CHAPTER 10

Cues of Communication Difficulty in Telephone Interviews

Frederick G. Conrad
University of Michigan[1], *USA*

Michael F. Schober
New School for Social Research, USA

Wil Dijkstra
Vrije Universiteit of Amsterdam, The Netherlands

When people converse, they do not just send and receive messages. They also give each other ongoing visual, auditory, and textual cues about the extent to which they believe they are understanding each other—about the extent to which their utterances are "grounded," to use Clark's (1996) terminology (see also Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Clark and Brennan, 1991, among others), as well as about their emotional reactions to what their partners are saying. Cues that addressees give to speakers—nods of the head, looks of confusion, back channels like "uh-huh" or "huh?", and requests for clarification like "what do you mean?" or reactions like "ouch!"—can alter what speakers say, such that both parties can be seen as molding each other's language use simultaneously.

Much the same sort of thing goes on during survey interviews. While respondents are answering questions, they simultaneously are giving cues about their comprehension and their reactions. Respondents who simply answer a question smoothly, without delays

[1] This work was supported in part by NSF Grant No. SES-0551294. The studies described here were supported by NSF Grant Nos. SBR-9730140 and ISS-0081550, with additional support from the Bureau of Labor Statistics and the Vrije Universiteit of Amsterdam.

Advances in Telephone Survey Methodology, Edited by James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster

or requests for clarification, are giving evidence that they believe they have understood the question well enough. Respondents who ask for clarification explicitly, as in

I: How many hours do you usually work?
R: Does "usually" mean on average?

are indicating explicitly that the communication is in danger of going wrong without clarification of a term in the survey question. Respondents who "report" rather than directly answering a question, as in

I: How many hours do you usually work?
R: Well, some weeks I work 60 hours and other weeks I work 30.

are indicating, less explicitly, that their circumstances do not map onto the questions in a straightforward way and that they would like the interviewer to make the decision for them based on their description of their circumstances. The respondent in this example is "reporting" potentially relevant facts rather than directly answering the question (see Drew, 1984; Schaeffer and Maynard, 1996).

Respondents can also implicitly signal (intentionally or not) their need for clarification or discomfort with a question by answering in ways that indicate trouble coming up with an answer:

I: How many hours do you usually work?
R: "Well... uh... usually fifty."

As we will discuss further, paralinguistic features of language like pauses, "well," and "uh" have been shown to be potential signals (intended or not) of need for clarification, plausibly just as informative as facial cues like looks of confusion. In the arena of telephone surveys, these features may constitute "paradata" (to use Couper's term [1998a, 2000b]) that telephone interviewers, and telephone interviewing systems of the future, can exploit.

In this chapter, we explore cues of comprehension difficulty displayed by respondents in telephone interviews and how these differ from cues displayed in face-to-face and other modes of interviewing. We discuss how interviewing techniques may moderate the informativeness of these cues, presenting evidence that whether and how interviewers present clarification in response to the cues can actually change their prevalence—how likely respondents are to emit them.

The approach taken in the studies described here is laboratory-based experimentation using questions about facts and behaviors. We use relatively small samples of respondents answering survey questions in situations where we have independent evidence about the facts about which they are answering—either from fictional scenarios from which respondents are to answer or from extensive postsurvey questioning to probe further into respondents' actual circumstances. The focus has been on cues of the extent to which respondents have understood questions, as opposed to on emotional reactions or rapport. The focus has also largely been on respondents' cues for questions about nonsensitive facts and behaviors, rather than on responses to sensitive questions or questions about attitudes and opinions; obviously, these would be important areas to study further.

Our laboratory approach clearly has its strengths and weaknesses. The power of the laboratory situation is that it allows us to manipulate features of interviews, either by training interviewers in particular techniques or by simulating telephone speech interfaces, so that causal inferences can be drawn. Thus far, the evidence from one larger scale replication in a U.S. national telephone sample (Conrad and Schober, 2000) of an earlier laboratory study (Schober and Conrad, 1997) suggests that the statistically reliable findings from these laboratory studies are likely to generalize to larger populations. But obviously, without full testing it is unknown whether all the results will so generalize, and we intend our discussion here to be suggestive of areas worth exploring rather than definitive or providing immediate prescriptions for practice in larger scale telephone surveys.

10.1 CUES IN TELEPHONE CONVERSATION

To a certain extent, the grounding cues in telephone surveys reflect the constraints of grounding understanding on telephones more generally, which differ from the constraints on grounding in other media (see Clark and Brennan, 1991, for extended discussion; see also Williams, 1977; Whittaker, 2003). Telephone interlocutors are audible to each other, but they do not have access to visual cues such as eye gaze, gestures, and frowns (although this is changing with the advent of video telephony). Unlike people conversing via handwritten letters, e-mail, or instant messaging, telephone interlocutors have access to audible paralinguistic cues that can be useful for understanding what their partner means: timing cues (long delays can be a sign of trouble—see Brennan and Williams, 1995), intonational cues (rising intonation, under the right circumstances, can be an indicator of doubt [Brennan and Williams, 1995]), and other discourse features like *ums* and *uhs*, sentence restarts, and other disfluencies that can give evidence about what speakers do and do not mean (e.g., Fox Tree, 1995; Clark, 1996; Brennan and Schober, 2001). Unless the phone connection is poor, telephone interlocutors perceive each other's signals more or less instantaneously, without the kinds of delays that can occur in handwritten letters, e-mail, or even (on a different scale) "instant" messaging. Because the channel of communication is speech, it leaves no reviewable trace that would allow further inspection, again unlike e-mail or chat room discourse, or messages left on telephone answering machines. And telephone conversationalists can produce and receive messages simultaneously, allowing overlapping speech (of course, within limits—complete overlaps can lead to communication breakdown); this differs from one-way forms of communication such as voice mail messages or walkie-talkie radio communication.

Other facts about grounding understanding in telephone survey interviews arise from what is unique about standardized survey interaction. As various reviews have noted (see, e.g., Suchman and Jordan, 1990; Schober, 1999; Schaeffer, 2002; Schober and Conrad, 2002, among others), respondents in standardized surveys are in the unusual position of trying to ground their understanding with a partner (the interviewer) who is not the originator of her utterances. Typically, interviewers read questions scripted and pretested by survey designers, and the agenda of the survey interview is predetermined and instantiated entirely by that script. The agent who is genuinely responsible for what a question means (and thus the person with whom grounding by rights should happen) is entirely absent, and the interviewer, as a mere purveyor of the questions, becomes a grounding partner of uncertain status.

Respondents are also in the unusual position that their grounding cues—their requests for clarification, their sounds of hesitation—are likely to be ignored or to be responded to in ways that are quite different from what would happen in less controlled and scripted telephone conversations. So, for example, interviewers trained in the strictest form of standardization (see, e.g., Fowler and Mangione, 1990) are trained *not* to respond substantively to explicit or implicit requests for clarification, because if only some respondents receive clarification, then all respondents are not receiving the same stimuli (words). Interviewers are, instead, trained to repeat the question (thus reframing any request for clarification as if it had been a simple mishearing) or to use a neutral "whatever-it-means-to-you" response—to explicitly leave the interpretation of question meaning up to respondents. (Note that what counts as training in standardized interviewing can vary substantially from survey center to survey center [Viterna and Maynard, 2002], and interviewers all trained in the same practices can nonetheless vary substantially in how they implement that standardization [e.g., Schober et al., 2004, Study 2]. In practice, in actual telephone interviews, then, some grounding cues may function more as they do in nonsurvey telephone calls.)

The result in current standardized interviewing practice is that only some proportion of respondents' ordinary communication cues are addressed in ways that respondents are familiar with in other telephone conversations (see Schober and Conrad, 2002, for more detailed discussion and examples). As in other interactions on the phone, survey respondents who simply answer a question without further ado will be taken as having provided evidence that their comprehension was successful, and the interaction will proceed. Also, as in other telephone interactions, respondents who say "huh?" or "what was that?" when they did not hear what their partners said will be given another chance when their partner (the interviewer) repeats the question. But unlike in other telephone interactions, respondents in the most strictly standardized surveys will find that both their explicit grounding cues (requests for clarification) and their implick grounding cues (reports, pauses, *ums* and *uhs*) are ignored or treated as if they were requests for something else.

We have demonstrated in a series of prior studies that how interviewers respond to telephone respondents' explicit and implicit communication cues can have substantial effects on the success of the communication and thus the quality of the resulting survey data. In our first laboratory study (Schober and Conrad, 1997) we contrasted two extreme ways that interviewers might handle respondents' grounding cues: strictly standardized interviewing, where requests for clarification or implicit signs of confusion are ignored or rejected, and a more collaborative form of interviewing in which participants talk about what has been said to be sure they understand each other sufficiently. (We dubbed this conversational interviewing because it relies on the conversational process of grounding [Clark and Wilkes-Gibbs, 1986; Schober and Clark, 1989; Clark and Schaefer, 1989; Clark and Brennan, 1991; Clark, 1996].)

In the study, we measured comprehension and response accuracy by having 42 telephone respondents answer 12 questions from ongoing U.S. government surveys on the basis of fictional scenarios. Because we had access to official definitions of survey terms, we knew what the survey designers hoped would be included and excluded in respondents' answers, and could judge respondents' interpretations of the questions based on their answers. For each question, we designed two alternative scenarios that a respondent might answer; in one case, there was no likely ambiguity in how the question mapped onto the respondents' (fictional) circumstances, and in the other case there was potential for a mapping ambiguity. For example, a respondent answering a question about how many bedrooms there are in his house would likely find this a straightforward question to answer when the fictional scenario is a floor plan of a house with three rooms labeled "bedroom." But that respondent is likely to have more trouble if one of those rooms is labeled "originally designed as a den, this room is being used as a bedroom." The hypothesis was that how interviewers handled grounding cues should be particularly important for these more complicated mappings.

Both in this laboratory study and in a subsequent field study in which interviewers telephoned a national sample of respondents (Conrad and Schober, 2000), the evidence showed that interviews in which interviewers were licensed to respond to explicit and implicit grounding cues substantively—conversational interviews—improved understanding and response accuracy, particularly when respondents' circumstances mapped onto questions in complicated ways. The evidence from the field study showed that these complicated mappings are frequent enough in a national sample to suggest that when grounding cues are ignored (in strictly standardized interviewing conditions) data quality may be compromised. There was, however, a cost to responding to grounding cues in both the laboratory and field study: providing clarification takes time, and so increased response accuracy is accompanied by increased interview duration.

In another laboratory study (Schober et al., 2004, Experiment 1), we made a first attempt at disentangling how interviewers' responses to explicit grounding cues (respondents' requests for clarification) and implicit grounding cues might affect data quality. In that study, we compared response accuracy under strictly standardized conditions and four versions of conversational interviewing. In all four versions, interviewers were able to clarify concepts after respondents provided explicit grounding cues (what we called respondent-initiated clarification). The grounding cues differed in whether (1) interviewers could also volunteer clarification in response to implicit cues (mixed-initiative clarification) and (2) they could use their own words to clarify the questions (paraphrased versus verbatim clarification).

As in the Schober and Conrad (1997) study, respondents answered 12 questions on the basis of fictional scenarios designed to be unambiguous (straightforward) or to include a mapping ambiguity (complicated). For example, one straightforward scenario described a prototypical nuclear family with two parents and two children living in the home. Respondents used this to answer "How many people live in this house?" The complicated counterpart described a similar family in which one child was a college student living away from home. Should this child be counted as living in the house? The definition of "living in a home" created by the survey sponsors resolved the confusion (the child should be counted as living in the home) and so an interviewer able to provide this information and ground the respondent's understanding should collect more accurate survey data.

The evidence showed that, for complicated mappings, response accuracy was greatest when interviewers responded substantively to both explicit and implicit grounding cues—and, less importantly for the current discussion, when they were licensed to do this in their own words rather than following a script. Respondents answered questions with complicated-mapping scenarios most accurately in mixed-initiative, paraphrased clarification interviews (87 percent). Accuracy was at intermediate levels when interviewers could initiate clarification in response to implicit cues or paraphrase definitions but not both (mixed initiative, verbatim: 66 percent; respondent initiated, paraphrased: 55 percent) and just as high when the only clarification that interviewers could provide was verbatim definitions requested by respondents (respondent initiated, verbatim: 59 percent). In contrast, when no clarification was available, accuracy was disturbingly low (standardized interviews: 28 percent).

The pattern of data suggests, then, that explicit and implicit grounding cues by respondents may contribute independently to the success of communication in telephone interviews. Although accuracy was high under the mixed-initiative, paraphrased approach, respondents provided explicit grounding cues—initiated clarification—for only 47 percent of the cases where clarification was given. The rest of the time interviewers initiated clarification episodes because respondents had displayed uncertainty in their answers or failed to answer the questions definitively. The most common way in which respondents showed they were uncertain was to describe their situation (reports); they did this for 58 percent (37 of 64) of the complicated cases where interviewers intervened voluntarily. Interviewers also intervened voluntarily when respondents asked them to repeat the question (9 percent of cases, 6 of 64) and when respondents explicitly said they were not sure about the answer but did not ask for clarification (6 percent of cases, 4 of 64). (See Schober and Conrad, 1997 for further details.)

Why do respondents initiate clarification sequences less than, it would seem, they might benefit from them? There are at least two possibilities. First, respondents may not realize they are misinterpreting key concepts; respondents may presume that their initial interpretation is the right one—what Clark and Schober (1991) have called "the presumption of interpretability." 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 concept like *bedroom* or *living in a house*.

Either way, it is possible that respondents display implicit evidence of need for clarification that telephone interviewers could potentially exploit. Although a respondent may not be keenly aware of confusion, it could be that her processing difficulty or uncertainty (conscious or not) is reflected in her speech: in hedges like "about 50 hours a week" instead of just "50 hours a week," in paralinguistic behaviors like a doubtful-sounding tone of voice or (with video telephony) in visual indicators like a furrowed brow. A respondent who finds it too effortful or embarrassing to ask for help could nevertheless signal his need for help in similar ways. We now turn to an examination of the validity of such uncertainty cues. Do they reliably occur when respondents are in need of clarification about what a question means and how to answer?

10.2 SPOKEN CUES OF RESPONDENT NEED FOR CLARIFICATION

The studies just described show that explicit requests for clarification are reliable communication cues. They also show that telephone interviewers can successfully judge when respondents need clarification even when they have not explicitly requested it. Presumably, they do this based on respondents' implicit cues—although note that it is also possible for respondents to answer incorrectly without producing any overt indications of uncertainty. What particular cues allow interviewers to make this judgment?

A preliminary set of answers comes from Schober and Bloom's (2004) examination of respondents' paralinguistic behaviors in the first turn following a question's delivery in the mixed-initiative, paraphrased as well as standardized interviews from Schober, Conrad and Fricker (2004). The focus was on several classes of paralinguistic behavior that in ordinary (nonsurvey) discourse have been linked to speakers' planning and production difficulties (Goldman-Eisler, 1958; Fromkin, 1973, 1980; Levelt, 1989), the complexity or conceptual difficulty of what they are trying to say (Bortfeld et al., 2001; Barr, 2003), the novelty of the information they are presenting (Fox Tree and Clark, 1997), and their uncertainty or lack of confidence in what they are saying (Smith and Clark, 1993; Brennan and Williams, 1995). Among the verbal behaviors they examined were

- *Reporting.* A way of answering a question that leaves the responsibility of answering to the person who posed the question. To use the Schober and Bloom example, if person A asks person B "Do you like punk rock?" and B responds "I like The Clash," B has left it up to A to decide whether The Clash's music counts as punk rock. A survey analog of this would be answering the bedroom question with "I have two bedrooms, and we are using a room as a bedroom that was originally designed as a den."
- *Speech disfluencies.* Parts of utterances that are not words in the everyday sense and are often assumed to provide little semantic content to the utterance; these include fillers (like *ums* and *uhs*), prolonged pauses in the wrong places (like a 2 second pause at the start of a turn), and repairs (such as "Y- y- yes I did buy a fi- some furniture").
- *Discourse markers.* Words that can alert listeners that what comes next is unexpected or that the speaker is not entirely sure about the content of what is being uttered (Schiffrin, 1987). Examples include *well* (as in "Well, we have three bedrooms") and *oh* (as in "Oh, we have two bedrooms").
- *Hedges.* Words such as *about* (as in "We have about three bedrooms") and phrases like *I think* (as in "I think we have three bedrooms").

Just as the explicit requests for clarification occurred more often for complicated than for straightforward mappings, so did reporting and some speech disfluencies. In addition to producing longer pauses, respondents produced fillers, pauses, and repairs reliably more frequently for complicated than for straightforward situations. This suggests that these communication cues—whether they indicate processing trouble, uncertainty, or even intentional grounding requests—count as valid markers of need for clarification. Combinations of some cues were even more diagnostic of need for clarification. For example, fillers and repairs, and fillers and reports, appeared together more often in complicated than straightforward situations, even more than one of these cues alone. Hedges and discourse markers, in contrast, appeared no differently in answers for complicated than straightforward scenarios, which suggests that, at least for these questions, they are not diagnostic of need for clarification.

Obviously, this set of results is based on a relatively small sample (41) of interviews using particular fact-and-behavior-based questions, and so we should be cautious about overstating their generality. But it is possible that if telephone interviewers can be trained to attend to and detect the cues that are particularly reliable indicators of need for clarification, particularly in combination with one another, they might be able to volunteer clarification in particularly judicious ways, explaining the question's meaning when it is actually needed and refraining from offering help when it is not needed.

An additional set of findings is that the use of various communication cues was affected by what interviewers were licensed to respond to: Respondents used some cues differently in conversational than in standardized interviews. Not surprisingly, explicit requests for clarification were far more likely in conversational (mixed-initiative, paraphrased) than standardized interviews; respondents no doubt recognized that explicit requests in strictly standardized interviews would be unlikely to elicit substantive grounding help. And along the same lines, perhaps it is not surprising that respondents produced reports (e.g., "She bought a floor lamp" when asked "Did Dana purchase or have expenses for household furniture?") more often in conversational than standardized interviews. As with explicit requests for clarification, reporting is an effective strategy only if interviewers are able to react substantively, perhaps determining and recording the answer on the basis of the report. In a standardized interview, reporting is less likely to help the respondent to ground his understanding; the interviewer is most likely to respond with a nondirective (and nongrounding) probe like "Would that be 'yes' or 'no'?"

But some of the disfluencies were also differentially present in standardized and conversational interviews—and not always in the direction one might expect. For example, fillers (*ums* and *uhs*) were actually *less* frequent in conversational interviews than in standardized interviews. Why might this be? Perhaps, when telephone respondents are deterred from requesting clarification (e.g., as seems to be the case in standardized interviews) their speech is more likely to reflect unresolved uncertainty. There may be a trade-off between having the ability to ask for clarification and the production of certain communication cues. In any case, these findings suggest a link between how an interviewer may respond to particular grounding cues and which cues the respondent produces.

10.3 SPOKEN VERSUS VISUAL CUES OF COMPREHENSION PROBLEMS

Are the cues we have been investigating specific to telephone surveys? *Do* respondents present different cues, as grounding theory proposes, in interviews that also allow the exchange of visual information? It is possible that respondents are exactly as likely to display confusion facially, in their voices, and via explicit requests for clarification whether or not the interviewer can see them. Or are respondents sensitive to the relative richness of cues afforded by the mode of data collection? Might 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?

At present, little is known about the particularity of communication cues in telephone surveys, despite various comparisons between telephone and face-to-face interaction in other settings (e.g., Williams, 1977; Whittaker, 2003). Large-scale comparisons between telephone and face-to-face interviewing, such as de Leeuw and van der Zouwen's (1988) meta-analysis, have found few, if any, differences in data quality that could be attributed to mode differences. Historically, the lack of difference between the modes has been good news: in the early days of telephone interviewing, this was celebrated as evidence that the new mode—and its concomitant benefits—had come of age.

We propose an alternative view of the situation: Respondents should produce mode-specific communication cues exactly to the extent that they are useful for grounding understanding or otherwise communicating useful information in the interview. In strictly standardized interviews, where interviewers are restricted from substantively responding to even explicit requests for clarification, the availability of a communication channel for grounding should make little difference. But when interviewers are licensed to use a particular communication cue—in some sort of conversational interview—the availability of a cue should be relevant because it can be exploited by the interviewer. More generally, if we open up the possibility that grounding cues are potentially useful to interviewers, then a new set of research questions emerges concerning when and how audio and visual cues might be redundant with each other and when they might complement each other in survey interviews.

As a first step toward exploring these kinds of questions, Conrad et al. (2004) carried out a laboratory study in which 42 Dutch respondents were asked about their own lives in either conversational or standardized interviews that were conducted over the telephone (four interviewers) or face-to-face (four other interviewers). 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 postinterview questionnaire, the change could be attributed to a change in their understanding brought about by reading the definition in the questionnaire, suggesting they had misinterpreted the question during the telephone interview.

The analysis was focused on one question about membership in the Dutch institution *verenigingen* 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. Indeed, answers changed more after standardized interviews than conversational ones for this question, suggesting that clarification during the (conversational) telephone interview had been beneficial to respondents' understanding and the accuracy of their answers.[2] However, there were no differences due to mode (telephone versus face-to-face). Why might this be 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 reliably more *ums* and *uhs* on the telephone (8.0 per 100 words) than face-to-face (6.1 per hundred words), as if they recognized that the interviewers could not see them on the telephone and so would need extra auditory evidence of their difficulty in answering. This was true both in standardized and conversational interviews, which suggests that interviewer responsiveness to cues is not the driving force behind the differential levels of disfluency. In the conversational interviews, telephone interviewers who provided clarification in response to disfluencies did so much sooner (after 4.2 "moves"—more or less equivalent to speaking turns) than the face-to-face interviewers (11.4 moves). Although the sample is too small for these results to be more than suggestive, it is possible that spoken cues of trouble can be taken as particularly revealing on the telephone.

What then are the visual cues available only face-to-face and 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, 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. (For further discussion of the communicative implications of gaze see, e.g., Doherty-Sneddon et al., 2002 and Goodwin, 2000.) The critical issue in the Conrad et al. (2004) study was whether respondents looked away more in conversational than standardized interviews when interviewers might possibly provide clarification based on these cues.

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 the 10 conversational, face-to-face interviews, as compared with 4.3 percent of the time in the 11 standardized, face-to-face 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

[2] See Conrad and Schober (2000) for another example of how questions requiring lists as answers produce more accurate data with conversational interviewing than standardized ones.

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. Curiously, conversational interviewers did not provide more clarification in response to this behavior, despite glancing at respondents at least once during 80 percent of their looking-away episodes. 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.

In addition to verbally signaling the need for clarification, speakers may supplement these cues visually (e.g., direction of gaze). If so, understanding might suffer on current telephones because, without visual cues, interviewers may miss opportunities to provide needed clarification. Alternatively, respondents may compensate for the limits of auditory-only communication by verbalizing their comprehension problems paralinguistically. Clearly, this warrants further investigation, particularly as video telephony becomes more practical (see Anderson, 2008 and Fuchs, 2008).

10.4 INTERACTING WITH AUTOMATED TELEPHONE INTERVIEWING SYSTEMS

It is currently unknown whether all interviewers, or only the most socially sensitive interviewers, can use verbal and visual cues of respondent uncertainty as a trigger for providing clarification. The division of attention that may have limited interviewers' use of gaze aversion in the Conrad et al. (2004) study could be a serious impediment. We propose that technology may be able to help. In particular, diagnostic software could be created that could take some of the attentional burden off interviewers by monitoring for spoken or even visual cues of respondent difficulty. One could even imagine deploying such technology as part of fully automated interviews in the not-so-distant future.

We have begun studying the effectiveness of this kind of diagnosis by simulating such technology with a "Wizard-of-Oz" (WOZ) technique (e.g., Dahlbäck et al., 1993). In this approach respondents believe they are interacting with an automated system via telephone but are actually interacting with a human (wizard) who presents preexisting speech files created to sound like synthesized speech. Unlike conventional speech recognition technology (as in some of today's interactive voice response [IVR] systems), the simulated dialogue technology is not limited to utterance recognition but can take into account discourse criteria like whether a concept has already been discussed and whether respondents' speech contains the kind of markers observed by Schober and Bloom (2004).

Here we describe two experiments using the WOZ technique to simulate automated interviewing technology. Respondents answer questions asked by the simulated interviewing system on the basis of fictional scenarios, just as in our studies of human telephone surveys, so that we have independent evidence about when they have interpreted questions as the survey designers intended. Note one advantage of this sort of study: The behavior of the "interviewer" can be manipulated with algorithmic precision in a way that is far less certain in training human interviewers.

10.5 DIAGNOSING RESPONDENT'S NEED FOR CLARIFICATION FROM COMMUNICATION CUES

In the first study (Bloom, 1999; Schober et al., 2000), a Wizard-of-Oz technique was used to simulate a speech interface. Users believed they were interacting with a computer, when actually a hidden experimenter presented the questions and scripted clarification. To enhance believability, we used an artificial-sounding computer voice (Apple's "Agnes" voice); virtually all respondents were convinced they were interacting with a computerized interviewing system, and the data from the few who doubted this were removed from the study.

In the first condition, the system could not provide clarification. This was similar to one of our strictly standardized interviews in that a confused respondent could not obtain a definition; if a respondent requested clarification, the system would repeat the question. In the second condition, clarification was based on explicit respondent-initiated grounding cues—the system would provide clarification if the respondent asked for it explicitly. In the third condition, clarification was based both on explicit and implicit respondent grounding cues (the initiative was mixed)—the system would also "automatically" provide full definitions when users displayed the cues of need for clarification cataloged in Schober and Bloom (2004). These included *ums*, *uhs*, pauses, repairs, and talks other than an answer. In the fourth condition, the system always provided clarification; no matter what the user did, the system would present the full official definition for every question.

The results with this simulated system in some respects parallel those for our studies of human interaction. As in Schober et al. (2004, Study 1), respondents were almost perfectly accurate when they answered about straightforward scenarios. For complicated scenarios, respondents were substantially more accurate when they were always given clarification (80 percent) than when they were never given clarification (33 percent).

But the pattern for grounding cues was somewhat different. Unlike in Schober et al. (2004), requiring explicit grounding cues (requests) in order to provide clarification was entirely ineffective, because respondents almost never asked for clarification despite being instructed to do so if they were "at all uncertain about the meaning of a word in a question." In the respondent-initiated clarification condition, the accuracy of respondents' answers was no better (29 percent) than when they were never given clarification. Most likely it did not occur to respondents that clarification was necessary; the presumption of interpretability (Clark and Schober, 1991) probably holds in computer-administered interviews. What *was* effective was relying on respondents' implicit grounding cues; response accuracy was reliably better when the system provided clarification in response to users' disfluencies and pauses (the mixed-initiative clarification condition) (59 percent), although not as good as when clarification was given always.

When the system provided clarification in response to implicit grounding cues, respondents were actually more likely to ask explicitly for clarification: Respondents asked questions more often in the mixed-initiative condition, presumably because they were more likely to recognize that clarification might be useful. These users also spoke less fluently, producing more *ums* and *uhs*—and there is some evidence that this tendency increased over the course of the interview. We speculate that this was because these users at some level recognized that the system was sensitive to their cues of uncertainty.

Why did respondents with the computer speech interface give explicit grounding cues (ask for clarification) so rarely? Perhaps even more than the respondents in the telephone interviews in Schober, et al. (2004) they found it relatively uncomfortable to articulate their confusion or uncertainty to a computer agent. But we cannot conclude this with certainty, as there are other differences that we suspect may have been even more important: Obtaining a definition with this particular speech interface was a more daunting prospect than getting a definition from a human interviewer, because the entire definition—not just the relevant parts—would be spoken, and this was time consuming (up to 108 seconds) and impossible to shut off. In contrast, human interviewers can potentially provide just the relevant part of the definition (as in the paraphrased clarification interviews in Schober et al., 2004) and respondents can interrupt the interviewer if necessary to circumvent the full delivery of the definition. Finally, respondents in the current study could not reject a system-initiated offer to provide a definition because the system did not offer—it simply provided—the definition. In the Schober et al. (2004) interviews, it was often the case that interviewers asked respondents if they wanted clarification.

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

Respondents rated the system more positively when it was responsive (respondent or mixed-initiative conditions). When the system was not responsive (no clarification or clarification always), users wanted more control and felt that interacting with the system was unnatural. Respondents did not report finding system-initiated clarification particularly more annoying than respondent-initiated clarification—which they almost never used.

Overall, these results suggests that enhancing the collaborative repertoire and diagnostic capability of a speech-interviewing system can improve comprehension accuracy without harming user satisfaction, as long as the system provides help only when it is necessary. But these improvements come at the cost of increased task duration, which raises questions about the practicality of a system with only these characteristics in real-world survey situations.

10.6 MODELING RESPONDENTS' SPEECH TO PROVIDE MORE TAILORED CLARIFICATION

We propose that systems may be able to provide more precisely tailored clarification to respondents by attending to their grounding cues in a more nuanced way. We demonstrated this in an experiment (Ehlen, 2005; Ehlen et al., 2007) in which we modeled different groups of respondents' relevant paralinguistic behaviors. In particular, the respondent modeling techniques allowed us to distinguish behaviors more likely to signal uncertainty from those that are less likely to do so. For example, someone who regularly says *well* and *uh* as part of their daily repertoire is less likely to be signaling comprehension difficulty with well or uh than someone who rarely uses them. A listener who makes this distinction is modeling the individual speaker. The same logic can apply to groups of speakers. Older speakers have been shown to be less fluent than younger speakers (e.g., Bortfeld et al., 2001), and so the same disfluency rate for a young and old speaker may indicate different states of understanding. In other words, the same level of *umm*ing might indicate problematic understanding for a younger speaker but ordinary speech for an older speaker. We applied this idea to automated interviewing by allowing the system to offer clarification on the basis of a generic model (same criteria for all respondents) and a stereotyped model (different criteria for old and young respondents).

One hundred respondents (50 older than 65 years of age and 50 under 40 years of age), answering 10 questions on the basis of fictional scenarios, participated in one of the five kinds of interviews: No Clarification, Respondent-Initiated Clarification, Required Clarification, Generic Respondent Model, and Stereotyped Respondent Model. The first two kinds of interviews, similar to their namesakes in the Bloom et al. study, generated the respondent speech that was used for the respondent models. In the Required Clarification interviews, respondents first answered each question; after this they were presented the full definition and could change their response if they chose to. These interviews served two functions. First, they provided a test bed for the models. In particular, they allowed us to ask how precisely the models predicted comprehension accuracy prior to the definition being presented. Second, they served as a benchmark of comprehension accuracy. Because definitions were presented for all questions, response accuracy under these conditions provided an upper bound on the benefits of clarification. In the Generic Respondent Model interviews, the system initiated clarification after certain speech conditions (discussed next) were met, regardless of respondent characteristics. In the Stereotyped Respondent model interviews, the conditions that triggered the system to provide clarification were different for older and younger respondents.

The respondent models were calculated with ordinary least-squares regression techniques in which the predicted behavior was response accuracy and the predictors were fillers, hedges, restarts, repeats, repairs, reports, mutters, confirmation pickups,[3] and pauses. While older respondents produced more spoken cues and longer

[3] An example of a confirmation pickup is "Usually, fifty" in response to "How many hours per week does Mindy usually work at her job?" because it picks up the term "usually" as a way of keeping it in play so that it can be confirmed or negotiated.

pauses than younger respondents, none of the cues improved the model beyond the predictive ability of pause length. If respondents answered too quickly or too slowly, they were more likely to be incorrect than if they answered within the intermediate (not too slow, not too fast) range. (We called this the "Goldilocks" range, in honor of the "just right" range of porridge temperatures and chair sizes in the "Goldilocks and the Three Bears" tale.) The Generic Goldilocks range was 2–7.2 seconds. The Goldilocks range for younger respondents ran from 4.6 to 10.2 seconds and for older respondents it ran from 2.6 to 4.35 seconds. Surprisingly, older people did not take longer to answer, in general, than did younger people. Rather, the range in which older respondents were likely to be accurate was smaller and faster than for younger respondents.

Response accuracy (again focusing on complicated scenarios) increased across the different kinds of interviews much as in the previous study (Bloom, 1999; Schober et al., 2000): poorest when no clarification was available, better when respondents could request clarification but the system could not provide it and better still when the system could also provide clarification (on the basis of models or after each question). When response accuracy prior to the required definition was used to test the models, 53 percent of the answers outside the Generic Goldilocks range were inaccurate and 83 percent of the answers outside the Stereotyped Goldilocks ranges were inaccurate.

When the system actually provided clarification, response accuracy improved reliably (linear trend) from Generic to Stereotyped Respondent Modeling to Required Clarification (after the definition had been delivered). In fact, the accuracy with Stereotyped Respondent models was as high as with Required Clarification, yet the interviews were reliably faster. It seems that by tailoring clarification to the respondent's age group, the clarification was often provided when it was needed and rarely when it was not needed, thus minimizing the temporal costs of improving clarification.

10.7 CONCLUSIONS

The data described here suggest in a preliminary way that respondents' explicit and implicit cues of their states of comprehension provide exploitable evidence that telephone interviewers and future telephone interviewing systems could use to improve survey data quality. The cues we have investigated are primarily conveyed through language (explicit requests for clarification, saying something other than a direct answer) and paralanguage (too-long and too-short delays before answers, *ums* and *uhs* in answers, etc.). But visual cues (gaze aversion, looks of confusion) are potentially exploitable in telephone interfaces that include visual information, to the extent that such cues prove nonredundant with textual and paralinguistic cues.

Of course, much more would need to be known before the findings described here are translatable into practical prescriptions for telephone survey centers; the studies described here only begin to address the larger set of theoretical and practical questions that survey researchers of the future will need answers to. And even for interpreting and applying these studies, we should be very clear about several caveats.

Most of our studies are laboratory-based, relying on small samples of respondents answering questions about nonsensitive facts and behaviors, and professional interviewers given brief training in alternate interviewing techniques. How one generalizes from experimental manipulations to actual, large-scale surveys is not at all straightforward. Experiments demonstrate that certain phenomena *can* happen but not that they necessarily do happen under all circumstances. To clearly map experimental results to large-scale surveys, one must know how often the circumstances created in the laboratory actually occur "in the wild."

We have focused on accuracy of respondents' interpretation of questions (the extent to which their answers reflect the same interpretations as the survey designers'), rather than on other important indicators of data quality in surveys, such as response rates, completion, and break-off rates. Whether the findings will extend beyond the lab to larger samples of respondents, different kinds of questions, different interviewer populations, and additional measures of survey quality remains to be seen.

Also, the effects of attending to grounding cues are apparent particularly for situations where the respondent's circumstances are ambiguous with respect to (well-pretested) questions, and so the frequency with which this occurs in real-world settings places a limit on the utility of our findings. Thus far, the evidence suggests that these sorts of "complicated-mapping" circumstances are frequent enough to worry about in broader samples and under more natural conditions—see Conrad and Schober, 2000; Suessbrick et al., 2005—and that they apply to attitude and opinion questions as well as questions about facts and behaviors, but again we recommend caution in assuming this about every sample of respondents and questions. We should also note that there are other potential sources of trouble answering questions that we have not been investigating: trouble understanding technical terms, trouble deciding on the response task (e.g., estimate or count?), trouble retrieving answers from memory, and troubles resulting from ambiguous (polysemous) terms in questions. Whether the communication cues surrounding these other kinds of trouble are the same as those investigated here is an open question.

In general, we see the findings described here as raising a set of issues that need to be explored in much greater detail, and we hope that this discussion helps to prompt further research along these lines. In particular, we think it would be important to know the following:

(1) *How diagnostic of need for clarification are communication cues across different respondents and circumstances?* While we have seen good evidence across our samples that, for example, respondents use *um* more often in their first turn after a question is asked when the answer is likely to need clarification, it is also clear that there is substantial individual variability in discourse styles, dialects, and propensity to *um*. The *um* from a respondent who never *ums* is presumably informative in a way that an *um* from a respondent who regularly *ums* is not. Gaze aversion from a steadily gazing respondent is different in meaning than gaze aversion from a respondent who never looks at the interviewer. To what extent do sensitive interviewers already attend to baseline rates of any potential communicative cue—delay in responding, reporting, gazing, *umming*—as they decide whether to probe or clarify?

To what extent should interviewers be trained to attend to the individual variability of such cues, and to what extent should interviewing systems of the future be able to diagnose the individual variability of such cues?

It is entirely possible that what is a cue of the respondent's comprehension difficulty in one situation reflects a quite different internal state in another situation. Or, worse, the same cue might indicate different states in the same situation on different occasions. Consider the "looking away" cue. Because respondents look away longer in conversational interviews (when interviewers might react to the cue) than in standardized interviews (when they cannot), looking away would seem to be under respondents' control. But if respondents look away irrespective of interviewers' ability to react, this would more likely mean that looking away is an involuntary reflection of comprehension difficulty.

Alternatively, looking away could reflect something other than comprehension difficulty. It could indicate that the respondent is planning what to say next and does not want to be distracted by looking at the interviewer (cf. Glenberg et al., 1998). Or it could reflect a state almost diametrically opposed to needing help: Looking away could reflect respondents' desire to maintain the floor and not surrender it to the interviewer, a concern more in conversational than standardized interviews (see the discussion by Clark, 1996, of turn allocation rules, pp. 321–324). Finally, looking away (or any of the cues we have suggested reflect comprehension difficulty) could indicate ambivalence about answering truthfully. If a respondent is concerned that providing a truthful answer might somehow cause her harm, for example, by confessing to illegal conduct, she might look away or answer less fluently or pause longer before speaking than if she has no reservations about answering.

(2) *How does an interviewer's responsiveness to any communication cue affect the respondent's likelihood of using it?* Presumably, when strictly standardized interviewers ignore explicit request for clarification, they reduce the likelihood that any but the most perverse or conversationally insensitive of respondents will continue asking for clarification. Does the same go for more implicit communication cues? The preliminary evidence reported here hints that rates of disfluency may well be sensitive to interviewers' responsiveness to them. Surely this is not the sort of thing that is under respondents' conscious control, and it suggests a kind of dyadic regulation of individual process that is not part of many views of communication.

(3) *Are all interviewers equally sensitive to grounding cues?* Although reliable empirical evidence on this is rare, ordinary intuitions and clinical observation of the general population suggest that people can vary substantially in their interpersonal sensitivity: their empathy, perspective-taking ability, and ability to attend to subtle linguistic cues (see, e.g., Davis, 2005, and other chapters in Malle and Hodges, 2005; Schober and Brennan, 2003). Presumably, interviewers who are socially tone deaf do not survive long in the job; overly sensitive interviewers, for whom denying requests for clarification may be interpersonally aversive, may also not survive in a telephone survey center that requires the strictest of standardized practice. What is unknown is the extent to which sensitivity to such cues is trainable, or whether adult language users already have an ingrained repertoire of cues to which they attend that

is resistant to change. Presumably, there are individual differences in interviewers' sensitivity to grounding cues (the approach of Hall and Bernieri, 2001 might allow assessment of this kind of skill), which will constrain the effectiveness of training. In particular, for interviewers low in sensitivity to such cues, the additional task of monitoring for them may be unrealistically burdensome (see Japek, 2005, for a discussion of interviewer burden). Also, external constraints like time pressure to finish interviews may cause even the most interpersonally attuned interviewers to ignore potentially useful cues.

(4) *How nonredundant are communication cues?* Thus far, little is known—in general and in survey interviews—about the extent to which visual cues provide information distinct from that provided by tone of voice, delay, or the content of what is said. While one can find clear examples where a particular cue seems to be the only indicator of trouble, we just do not know whether a single cue is always sufficiently diagnostic to warrant intervention by an interviewer or an interviewing system. To complicate matters, it is possible that different respondents may have different discourse styles: One respondent's gaze aversion may always be accompanied by a pause and an *um*, while another's gaze aversion may provide unique nonredundant information. To the extent that cues are redundant, interviewers who already have a lot to attend to might be able to rely on the cues in the most easily available or measurable channel.

(5) *How multifunctional are communication cues?* Our approach thus far has been focused on the cognitive aspects of comprehending questions and how respondents' disfluencies and other cues provide relevant evidence. But every cue we have discussed—explicit requests for clarification, reports, hedges, and so on—is also a potential indicator of the respondent's emotional state, level of irritation, and likelihood of continuing the interview. Respondents could delay or rush responses not only because they have trouble understanding the question or have not thought hard enough, but also because they find a question intrusive, because they feel the interview has gone on too long, or because the interviewers' nonresponsiveness to a request for clarification is becoming trying. To what extent do grounding cues also provide evidence about the rapport and emotional alliance between interviewers and respondents? We suspect that although grounding cues and rapport cues are conceptually distinct, in practice they can be quite intertwined. For example, an interviewer's apology for the stiltedness of an interview ("I'm sorry, I can only repeat the question") can be a response to cues that the interview is going offtrack both on affective dimensions (the respondent's frustrated tone of voice) as well as on grounding dimensions (explicit and implicit indicators of need for clarification).

As the space of new technologies available for telephony expands, telephone interviews are beginning to share more features with face-to-face interviews (see Schober and Conrad, 2008 as well as other chapters in Conrad and Schober, 2008). Will additional visual information help improve comprehension and thus data quality? Whittaker (2003) has observed that across various (nonsurvey) domains there is little evidence in support of the *bandwidth hypothesis*: the idea that adding visual information to speech will improve the efficiency of communication. It may be that the total amount of usable information about a communicative partner's need for clarification is the same with or without video. It remains to be seen what the facts are for surveys with different populations of respondents, with individually and culturally variable communication styles, with different domains of questioning (sensitive and nonsensitive questions), and with different interviewing agents (human versus computer) with different capabilities (diagnosing and responding to requests for clarification versus leaving interpretation up to respondents). How these questions are answered will help shape future debates about how telephone interviews should be conducted.

Advances in Telephone Survey Methodology

JAMES M. LEPKOWSKI

Institute for Social Research
University of Michigan
Ann Arbor, MI

CLYDE TUCKER

Bureau of Labor Statistics
U.S. Department of Labor
Washington, DC

J. MICHAEL BRICK

Westat
Rockville, MD

EDITH DE LEEUW

Department of Methodology and Statistics
Utrecht University
The Netherlands

LILLI JAPEC

Department of Research and Development
University of Stockholm
Stockholm, Sweden

PAUL J. LAVRAKAS

Nielsen Media Research
New York, NY

MICHAEL W. LINK

Centers for Disease Control and Prevention
Atlanta, GA

ROBERTA L. SANGSTER

Bureau of Labor Statistics
U.S. Department of Labor
Washington, DC

1807 WILEY 2007
WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION