When survey respondents' answers include pauses, stammers, and hedges, does this indicate that they are in danger of misinterpreting the survey question? Or are disfluencies so common in ordinary discourse that they are nondiagnostic? We analyzed respondents' first utterances after survey questions in a corpus of 42 laboratory-based telephone interviews (Schober & Conrad, 1997). Because respondents had answered questions on the basis of fictional scenarios, it was clear when their answers reflected a need for clarification. Half the respondents participated in strictly standardized interviews in which any clarification was prohibited, and half participated in more collaborative interviews in which clarification was encouraged. Results showed that respondents were more likely to pause longer, produce an *um* or *uh*, repair their utterance, and describe their circumstances in the first turn after a question was asked when answering about scenarios for which they needed clarification. Various combinations of discourse cues were also diagnostic of the need for clarification, although often only in the strictly standardized interviews.

Consider this invented interchange¹ from a telephone interview for the Current Population Survey, which is the primary data source for calculations of the U.S. unemployment rate:

¹In all excerpts, a period between two spaces (.) represents a pause. A hyphen at the end of a word ("that-") indicates that the word was cut off. Question marks indicate rising intonation, and utterance-final periods indicate falling or flat intonation regardless of whether the utterance is a question or an assertion.
Interviewer (I): Last week, did you do any work for pay?
Respondent (R): Uh, yes.

Does the respondent’s “uh” and pause indicate that she was having trouble with the question? Intuitively, the “yes” answer seems more tentative than it would have if the respondent hadn’t preceded it with the “uh” and pause. And the answer seems less tentative than it would have if the respondent had prefaced it with explicit disavowals and hedges, as in “Well, I get a tuition reduction for working at school, so that’s hard to answer. I guess the answer is yes.” A skilled and sensitive interviewer might pick up any or all such cues in deciding whether to probe further about why the question is hard to answer or whether to provide clarification about what the survey sponsors mean by “work for pay.”

A plausible alternative view is that perhaps pauses, hedges, and disfluencies such as “uh” are so common and nondiagnostic in ordinary discourse (or at least in surveys) that the interviewer ought not pay particular attention to them. After all, despite any initial fumbling, the respondent has presented an answer (“yes”) appropriate to the question’s requirements; perhaps the interviewer should not presume to question the respondent’s interpretation. In fact, allowing interviewers to question the respondent’s interpretation or probe further might lead to the sorts of bias and unreliability that plagued the survey industry at its inception and that led to the development of standardized interviewing methods designed to minimize the impact of interviewers on the survey process (see Beatty, 1995).

Which view is right? Do uh's and pauses and hedges reflect survey respondents’ need for clarification and likelihood of misinterpreting questions? Or are they simply slips reflecting speakers’ momentary planning and production difficulties but that do not reliably correlate with accuracy of the survey responses? In this study, we investigated the issue through coding of a corpus of 42 laboratory-based telephone survey interviews (Schober & Conrad, 1997) that allows us, as discourse researchers, special access to what interviewers meant by the questions and how respondents interpreted them. The corpus also allows us to compare the diagnosticity of respondents’ discourse cues in two quite different styles of survey interviewing: strictly standardized interviewing, in which the interpretation of survey terms is left up to the respondent, and more collaborative or “conversational” interviewing, in which interviewer and respondent work together to make sure survey terms are interpreted as intended.

CONCEPTUAL ALIGNMENT IN SURVEY INTERVIEWS

The extent to which survey respondents interpret words as intended is part of the broader question of the nature of conceptual alignment in conversation. Of course, survey interviews differ in important ways from other sorts of conversational inter-
actions in that they are scripted by outside parties, are usually not initiated by the question answerer, and often limit the ordinary mechanisms that conversationalists have for making sure they have understood each other, among other differences (see, e.g., Houtkoop-Steenstra, 2000; Schaeffer, 2002; Schaeffer & Maynard, 1996; Schober, 1999; Schober & Conrad, 2002; Suchman & Jordan, 1990). However, the processes of comprehending ordinary words in surveys is no doubt related to comprehension of those words in other circumstances.

A common view of conversational communication consistent with the assumptions underlying standardized survey interviewing is that people’s interpretations of what their partners mean are often accurate—or at least sufficiently accurate for their current purposes—from the moment they first hear their partners’ utterances. This view is implicit in theories that presume that people’s semantic systems are functionally identical. As one account (Pickering & Garrod, in press) puts it, interlocutors in dialogue often achieve conceptual alignment simply because addressees’ comprehension systems follow the same principles that speakers’ production systems follow; that is, once a speaker utters a word such as job, the addressee’s phonetic, morphological, syntactic, semantic, and pragmatic systems activate representations of job concordant with the speaker’s.

An important point about conversational comprehension that is not embodied in standardized survey interviewing practices is that interlocutors in ordinary conversation who recognize that they have not understood each other well enough for their current purposes can initiate repair sequences to clarify misinterpretations (Clark, 1996; Clark & Wilkes-Gibbs, 1986; Schegloff, 1972, 1997; Schegloff, Jefferson, & Sacks, 1977; among many others). This allows conversational participants ultimately to achieve conceptual alignment even if they initially misunderstood each other in ways that nonparticipants cannot (Goffman, 1976; McGregor, 1986; Schober & Clark, 1989). By leaving the interpretation of survey terms up to respondents, strictly standardized interviewing procedures remove important mechanisms for achieving conceptual alignment from the survey process (Houtkoop-Steenstra, 2000; Schober, 1999; Schober & Conrad, 2002).

In part because of this, the survey interview is an arena of interaction in which unnoticed misunderstandings of words can be surprisingly frequent. As demonstrated with a range of qualitative and quantitative laboratory and observational methods (e.g., Belson, 1981, 1986; Clark & Schober, 1991; Conrad & Schober, 2000; Houtkoop-Steenstra, 2000; Schaeffer & Maynard, 1996; Schober & Conrad, 1997, 2003; Suchman & Jordan, 1990), survey respondents can have quite different interpretations of seemingly banal terms in survey questions like you, job, household, or furniture than the survey designers. For example, Belson (1981) reported that when respondents were asked “For how many hours do you usually watch television on a weekday? This includes evening viewing” (p. 127), 15% of the respondents interpreted “you” as including other people too, 33% included times the television was on but they had not been paying attention to it, and 61% in-
cluded days other than the 5 weekdays intended or excluded some weekdays for other reasons.

The survey interview is thus a useful situation in which to examine conceptual misalignment and its resolution. First, unlike in large-scale corpus studies of impromptu conversations, in surveys about facts and behaviors (e.g., asking questions about a respondent’s job) it is possible to have independent evidence about what speakers (survey designers and interviewers) mean by what they say (e.g., published definitions for what should count and should not count as a job). It is also possible to have corroborative evidence about addressees’ (survey respondents’) conceptualizations: If a respondent’s actual life circumstances are known, can be inferred (either from direct observation, from checks of official records, or from subsequent reinterviewing), or can be simulated (from fictional scenarios in a laboratory), then a respondent’s answers to survey questions provide an unobtrusive window into whether their conceptualizations match the survey designer’s. Although these research advantages can be gained in more artificial laboratory tasks (e.g., referential communication tasks in which speakers engage in conversations about unfamiliar objects), the survey situation has the benefit of being a realistic situation that people encounter in day-to-day life and one in which conceptual alignment potentially has substantial real-world implications (for example, U.S. unemployment statistics are based on tens of thousands of telephone interviews per month in which questions about jobs are asked).

In this study, we capitalize on work manipulating features of survey interactions to examine the effects on survey data quality (e.g., Cannell, Miller, & Oksenberg, 1981; Conrad & Schober, 2000; Dijkstra, 1987; Schober & Conrad, 1997; Schober, Conrad, & Fricker, 2004). In particular, we examine the diagnosticity of discourse cues in a set of interviews that manipulated two factors that have been shown to affect respondents’ comprehension accuracy (Schober & Conrad, 1997).

The first factor is the prototypicality of the respondent’s circumstances with respect to concepts in the survey questions. Consider a respondent to the Current Point of Purchase Survey who is asked “Last year, did you purchase or have expenses for household furniture?” If the respondent did not purchase anything even remotely resembling household furniture, the answer is easy. If the respondent bought a chair or a dining room table, the answer is also easy, as those are prototypical pieces of household furniture; the mapping between the respondent’s circumstances and the survey designers’ intentions should be straightforward. However, if the respondent bought a floor lamp, the answer is less straightforward; for some purposes, one might consider a floor lamp a piece of household furniture, and for other purposes, one might not. The sponsoring agency for this survey (the Bureau of Labor Statistics) has created official definitions for what should be counted as a furniture purchase based on their own econometric purposes and on years of encountering idiosyncratic situations in survey pretests, and for their purposes, the floor lamp should not be counted as a furniture purchase. In such “com-
MISALIGNMENT CUES IN SURVEYS

At the heart of survey sampling are "misaligned mapping" situations, where respondents have been shown to answer questions in ways that demonstrate that their conceptualizations can often be misaligned with the survey designers'—for example, counting floor lamp purchases as household furniture purchases.

The second manipulated factor is the style of interviewing. In the corpus, half the interviews were carried out according to the strictest standardization: Interpretation of the terms was left entirely up to respondents, even if they requested clarification. This condition represents the most stringent test of how misaligned respondents' conceptions are with survey designers' because there is virtually no opportunity for interviewers to steer respondents' interpretations. This style of interviewing has been held up as an ideal for large-scale surveys (Fowler & Mangione, 1990) because it is argued to minimize the impact of individual interviewers on responses; if some respondents receive definitions and others do not, the interview cannot be considered strictly standardized. The other half of the interviews were collaborative: Interviewers and respondents could work together to assure that respondents had understood the questions as intended, and interviewers could rely on discourse evidence to decide whether to present clarification. This sort of interviewing has been promoted by critics of strict standardization in the survey world (Suchman & Jordan, 1990). Thus, the experimental design embodied in the corpus allows one to see whether respondents' cues are differently diagnostic when interviewers are and are not licensed to respond to them.

POTENTIAL DISCOURSE CUES OF CONCEPTUAL MISALIGNMENT

From examining the discourse literature, we identified several potential cues of respondents' conceptual misalignment with survey designers. Explicit requests for clarification such as "What do you mean by household furniture?" or "Is a floor lamp a piece of household furniture?" are obviously the best such cues. But explicit requests are far less frequent than one might expect, both in surveys (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober et al., 2004) and in other discourse settings. Sometimes people wait to hear more, as if all were well, in hopes that something their partner says later will clarify what was meant. People may want to avoid challenging their partners, to avoid looking stupid, or to maintain the flow of the conversation. And sometimes they honestly don't recognize that they haven't understood. For example, college students in one study (Graesser & McMahen, 1993) rarely asked questions about math problems even when the problems contained anomalous and irrelevant information. In this corpus, respondents explicitly requested clarification far less often than they needed it, as demonstrated by the extent to which their answers deviated from what the survey designers had
intended. It is exactly because of the rarity of explicit requests that we examine other possible discourse cues.

One possible set of cues are *speech disfluencies* such as *ums* and *uhhs*, too-long pauses in the wrong places, and self-corrections. In this study, we focused on the following disfluencies:

- **Midclause or preutterance pauses** lasting longer than 1 s (see Jefferson, 1989, for evidence that a 1-s minimum is the reasonable lower bound). An example of a midclause pause of this sort would be found in a response to the household furniture question such as “Yes I (1.2-s pause) did buy some furniture.” An example of a preutterance pause would be “(2.3-s pause) Yes I did buy some furniture.”
- **Fillers** (also called filled pauses) such as *um, uh,* and *mm,* as in a respondent’s answer “Yes I did buy some um furniture” or “Uh yes I did buy some furniture.”
- **Repairs,** in which speakers correct what they say (see Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Levelt, 1989). In the answer “Y- y- yes I did buy a fl- some furniture,” the respondent would have made several repairs.

Empirical evidence from ordinary (nonsurvey) discourse has long shown that these disfluencies are linked to speakers’ planning and production difficulties (Fromkin, 1973, 1980; Goldman-Eisler, 1958; Levelt, 1989), the complexity or conceptual difficulty of what they are trying to say (Barr, 2003; Bortfeld et al., 2001), the novelty of the information they are presenting (Fox Tree & Clark, 1997), and their uncertainty or lack of confidence in what they are saying (Brennan & Williams, 1995; Smith & Clark, 1993). Of course, none of these is a priori guaranteed to reflect complicated mappings in survey responding; speakers might well be fluent and confident in their inaccurate interpretations and uncertain about their accurate ones. And, although listeners in other discourse settings have been shown to be able to make use of disfluencies as they disambiguate speakers’ utterances (Arnold, Fagnano, & Tanenhaus, 2003; Brennan & Schober, 2001; Brennan & Williams, 1995, Fox Tree, 1995), this doesn’t guarantee that respondents’ disfluencies are necessarily useful to interviewers as evidence of respondents’ nonprototypical circumstances.

Another potential cue of respondents’ potential conceptual misalignment is **hedges:** approximators such as *about,* as in “We have about four bedrooms,” and shields such as *I think,* as in “I think we have four bedrooms” (see Prince, Frader, & Bosk, 1982, for a taxonomy of hedges). Speakers have been shown to hedge more when they have more potential ways to describe objects (Brennan & Clark, 1996); certainly the common interpretation of hedges is that they denote speaker uncertainty or reluctance.
A third set of potential cues of conceptual misalignment are discourse markers such as *well* (as in “Well, yes I did buy some furniture”) and *oh* (as in “Oh, I did buy some furniture”). As Schiffrin (1987) defines them, discourse markers, which are often found in the initial position of an utterance, operate both locally and globally in discourse. Some discourse markers 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).

The final potential cue of conceptual misalignment we examine in this study is what Drew (1984) called *reporting* (see also Schaeffer & Maynard, 2002), in which one answers a question in a way that leaves the responsibility of answering to the person who posed the question. For 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 household furniture question with “I bought a floor lamp.”

Are any of these cues used differently when their audience is likely to be able to recognize and respond to them? Evidence on this score is scant, although people have been demonstrated to be more disfluent when speaking to human partners than when they speak to computers (Oviatt, 1995). There are many potential reasons for this, but one possibility is that speakers are sensitive to the fact that machines may not be able to comprehend their disfluencies. If any or all of the cues investigated here are not only symptoms of speakers’ trouble but also full-blown communicative signals, as Clark (1994, 1996) has proposed, then speakers (here, the respondents) should produce them more in circumstances in which their addressees (the interviewers) are licensed to provide clarification in response to them: in the collaborative rather than strictly standardized interviews.

**METHOD**

To determine whether these various discourse cues occur reliably more often in situations in which respondents have been shown to be likely to misunderstand survey concepts, we contrasted their use in (a) straightforward situations, such as when a respondent who purchased an end table or a sofa (objects that clearly count as furniture) is asked about household furniture purchases, with (b) situations such as our floor lamp purchasing respondent’s, in which her circumstances map onto the question concepts in a more complicated way. To examine whether the interviewer’s potential uptake of these cues alters the respondent’s production of them, we contrasted their use in strictly standardized and conversational interviews. We focused on potential discourse cues in the respondent’s first turn after the question was asked rather than those in subsequent turns. Although in longer question–answer sequences, interviewers might be able to use paralinguistic evidence from subsequent turns, a substantial number of sequences (70% in the standardized in-
terviews) consisted of simply the interviewer's question followed by the respondent's answer followed by the interviewer's acknowledgment.

Note that the design of the surveys in the corpus focuses on one particular kind of misunderstanding that can arise in a survey situation: potential misalignment with survey designers' definitions of survey concepts. There are, of course, many other possible sources of respondent inaccuracy in survey interviews (see Schaeffer, Maynard, & Cradock, 1993). Respondents might misunderstand what was said because, for example, they heard the interviewer say farm rather than firm. Respondents might misremember events, for example, recalling that they bought milk last week when they did not. Or they can misinterpret the nature of the response task, for example, assuming they should estimate rather than count all instances when answering a question about the frequency of a behavior (see Conrad, 1999). Each of these other kinds of misalignment may have its own sorts of cues.

The survey in the corpus included 12 questions from three different ongoing U.S. government surveys. Four questions adapted from the Consumer Price Index Housing survey were about housing (e.g., “How many people live in this house?”). Four questions adapted from the Current Point of Purchase Survey were about purchases (e.g., “Last year, did Carla purchase or have expenses for car tires?”). Four questions adapted from the Current Population Survey were about employment (e.g., “Last week did Pat have more than one job, including part-time, evening or weekend work?”).

In the interviews, respondents answered these questions based on fictional scenarios: purchase receipts, floor plans, or brief descriptions of situations. Respondents studied the scenarios before the interview, and they also had the scenarios in front of them during the telephone interview. For each question there were two alternate scenarios, one of which mapped onto the question in a straightforward way (e.g., a purchase receipt for an end table for the household furniture question) and the other of which mapped onto the question in a complicated way (e.g., a purchase receipt for a floor lamp). For each respondent half of the scenarios had straightforward mappings and the other half had complicated mappings. Respondents nearly always gave correct answers for straightforward mappings (97% and 98% accuracy in standardized and collaborative interviews, respectively), but they were far more likely to give inaccurate answers for the complicated mappings, particularly in the standardized interviews (28% accuracy compared to 87% accuracy for the collaborative interviews).

Half of the respondents in the transcribed interviews participated in strictly standardized interviews in which interviewers left the interpretation of questions entirely up to the respondents. The other half participated in "conversational" interviews in which interviewers tried to make sure that respondents had interpreted the questions according to the official definitions, saying whatever it took to make this happen. In these interviews, respondents were instructed to ask for clarification whenever they had the slightest doubt about how to interpret survey questions;
We coded for each potential misalignment cue—pauses, fillers, repairs, hedges, discourse markers, and reporting—in the respondent utterance that immediately followed the interviewer's question. To verify whether the coding was reliable, two judges coded 4 of the 42 transcripts (two standardized and two conversational interviews selected at random). Reliability of the coding ranged from excellent (Cohen's kappa of .92 for pauses) to satisfactory (.68 for repairs), with an acceptable overall kappa of .75 (see Everitt & Haye, 1992). Because so few discourse markers appeared in the respondents' first utterances after the interviewer's question, we dropped them from further consideration.

We then asked two main questions. First, do respondents use any of these potential cues more when mappings are complicated? If so, this would suggest that these linguistic cues provide reasonable evidence for interviewers (or even computerized interviewing systems) that respondents may need clarification. Second, do respondents produce these cues differently in conversational than in standardized interviews? If so, this would suggest that respondents modify their use of the cues depending on whether they will be interactionally useful.

RESULTS

Explicit Requests for Clarification

Obviously the very best predictors that respondents may misunderstand survey questions are respondents' explicit requests for clarification in the first turn immediately following the interviewer's question (see Oksenberg, Cannell, & Kalton, 1991). Respondents could do this with explicit questions such as "What do you mean by household furniture?" or more elliptically, as in this interchange:

I: How many people live in this house?
R: Full-time?

As Table 1 shows, respondents rarely explicitly requested clarification in standardized interviews (no doubt because such requests would not be honored) and

In actual survey settings, whether or not the interviewer would be licensed to probe or provide definitions depends on just how standardized the interview is supposed to be; different survey centers implement standardized interviewing procedures surprisingly differently (Viterna & Maynard, 2002). And even within survey centers that train interviewers uniformly, different interviewers seem to vary in how strictly standardized their interviews are (Schober et al., 2004).
TABLE 1
Percentage of Questions for Which Respondents Explicitly Requested Clarification in First Turn After Question

<table>
<thead>
<tr>
<th>Interview</th>
<th>Straightforward</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Collaborative</td>
<td>16</td>
<td>34</td>
</tr>
</tbody>
</table>

reliably more often in conversational interviews, \( F(1, 40) = 16.00, p < .001 \), as they had been instructed (and when such requests were more likely to be effective). As the table also shows, respondents explicitly requested clarification much more often for complicated cases than for straightforward cases, \( F(1, 40) = 21.78, p < .001 \). This demonstrates that explicit requests for clarification, when they occur, are indeed excellent markers of respondents’ likelihood of misunderstanding survey questions.

However, note that many complicated mappings associated with conceptual misalignment did not lead to explicit requests for clarification, which means that relying on respondents to provide them is insufficient. The following analyses of respondents’ first turns at talk after questions were asked exclude all question-answer sequences in which respondents explicitly requested clarification, because it is exactly in those cases that interviewers lack explicit evidence about need for clarification.

Pauses

The mere presence of a pause longer than 1 s in the first turn following the interviewer’s question was not a good predictor that respondents were likely to need clarification. As Table 2 shows, although respondents paused reliably more often for complicated mappings than for straightforward mappings in both standardized and conversational interviews, \( F(1, 40) = 6.40, p < .05 \), respondents paused so of-

TABLE 2
Percentage of Questions for Which Respondents Paused Longer Than 1 Sec in First Turn After Question, Excluding Explicit Requests for Clarification

<table>
<thead>
<tr>
<th>Interview</th>
<th>Straightforward</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized</td>
<td>50.8</td>
<td>66.2</td>
</tr>
<tr>
<td>Collaborative</td>
<td>56.6</td>
<td>68.1</td>
</tr>
</tbody>
</table>
ten for straightforward cases—more than half the time—that this cue was not particularly informative.

Other aspects of pausing, however, were reliable indicators of need for clarification—but only in standardized interviews. In the standardized interviews, the average length of respondents' pauses in their first turn after hearing the question was greater for complicated mappings (2.64 s) than for straightforward mappings (2.43 s), whereas in the collaborative interviews, there was no difference (2.89 s for both kinds of scenario), interaction $F(1, 38) = 4.81, p = .034$. In the standardized interviews, the average number of pauses in a respondent's first turn following the question was greater for complicated mappings (1.42) than for straightforward mappings (1.04), but there was no difference in the collaborative interviews (1.33 vs. 1.25, respectively), interaction $F(1, 38) = 5.18, p < .03$.

The fact that pauses are informative about need for clarification in standardized interviews but not in collaborative interviews shows that respondents are sensitive to the interviewing context—but not in the way one might predict if respondents present more signals of need for clarification when they are likely to be responded to. That is, if respondents at some level use pauses to signal their need for clarification when it will do them some good, they should pause longer and more often in the collaborative interviews, and the data are not consistent with this view.

Fillers

As with the pauses, fillers were good predictors of respondents' need for clarification in standardized interviews, but not in collaborative interviews. As Table 3 shows, in the standardized interviews, respondents were more than twice as likely to produce a filler in their first turn following the question for complicated mappings than for straightforward mappings. However, there was no difference (if anything, the reverse) in the collaborative interviews, interaction $F(1, 40) = 9.11, p < .005$.

Although respondents were substantially more likely to produce fillers in the first turns after the question in collaborative interviews (40%) than in the standardized interviews (22.5%), $F(1, 40) = 8.31, p = .006$, this is not necessarily evidence

<table>
<thead>
<tr>
<th>Scenario Type</th>
<th>Straightforward</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interview</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized</td>
<td>12.7</td>
<td>32.4</td>
</tr>
<tr>
<td>Collaborative</td>
<td>43.8</td>
<td>36.3</td>
</tr>
</tbody>
</table>
that they were using the fillers to display need for clarification for collaborative interviewers; note that in the collaborative interviews, they presented fillers just as often for straightforward scenarios for which they did not need clarification as for complicated scenarios for which they did.

Repairs

As Table 4 shows, in both types of interviews respondents were more than twice as likely to utter at least one repair in their first turn following the question for complicated mappings than for straightforward mappings, $F(1, 40) = 9.36, p < .005$. In the standardized interviews, the difference was even greater, with respondents producing repairs more than three times as often for complicated mappings than for straightforward mappings. Even though repairs were relatively rare in the standardized interviews, they were still very good predictors of when respondents might be likely to misinterpret questions.

Respondents were marginally more likely to produce repairs in collaborative interviews than in standardized interviews, $F(1, 40) = 3.40, p = .073$. This is consistent with the possibility that speakers utter repairs as signals of need for clarification that their listeners can pick up on. But, again, the fact that repairs also increased for straightforward scenarios in the conversational interviews suggests that their increased use does not simply reflect the need for clarification but perhaps other sorts of interactional demands.

Hedges

Hedging in the first turn immediately after the question was rare in this corpus, but the pattern was similar to that for repairs. Respondents in the standardized interviews hedged more often for complicated mappings (5.6% of question-answer sequences) than for straightforward mappings (1.5%), but the pattern was the reverse for collaborative interviews (4.7% and 10.3%, respectively), interaction $F(1, 40) =$

---

**TABLE 4**

<table>
<thead>
<tr>
<th>Scenario Type</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized</td>
<td>8.1</td>
</tr>
<tr>
<td>Collaborative</td>
<td>17.9</td>
</tr>
</tbody>
</table>
8.32, \( p = .006 \). Once again, hedging was marginally more frequent in collaborative interviews than in standardized interviews, \( F(1, 40) = 3.09, p = .087 \).

So hedges seem to be reasonable predictors of need for clarification in standardized interviews but not in collaborative interviews, in which the pattern of hedging is reversed. This is hard to explain; perhaps respondents in conversational interviews who were not sure about answers that should have been obvious were more likely to mark those answers with hedges as a way of reducing embarrassment.

### Reporting

Reporting was rare in the first turn after a question was asked even though reporting at some point was fairly common (see Schober & Conrad, 1997). As Table 5 shows, respondents reported reliably more often in conversational interviews than in standardized interviews, \( F(1, 40) = 4.52, p < .05 \). Although there was only a marginal effect of mapping on how often respondents reported, \( F(1, 40) = 3.38, p = .073 \), in standardized interviews, reporting appeared to be a perfect predictor of complicated mappings; respondents only reported when faced with a complicated mapping, \( F(1, 20) = 8.00, p < .01 \).

In this sample, then, reporting in the first turn after a question is asked is an excellent indicator of the need for clarification.

### Multiple Cues

Respondents' cues could co-occur in the same utterances, as in this explicit request for clarification from the corpus (cues in italics):

\[
\text{I: Has Kelly purchased or had expenses for household furniture.} \\
\text{R: (3.5 s) Um h- household furniture or are we (1.7 s) uh referrin' to uh fix- fixtures and accessories?}
\]

To examine whether the cues considered in combination were effective predictors of respondents' need for clarification, we carried out an exhaustive series of addi-

<table>
<thead>
<tr>
<th>Interview</th>
<th>Straightforward</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized</td>
<td>0</td>
<td>4.8</td>
</tr>
<tr>
<td>Collaborative</td>
<td>4.8</td>
<td>8.7</td>
</tr>
</tbody>
</table>

---

**TABLE 5**

Percentage of Questions for Which Respondents Reported Their Circumstances Rather than Answering in First Turn After Question
tional analyses. In one set, we examined the diagnosticity of the combination of any two, three, and four cues (combinations of all five cues were too rare to evaluate). In another set, we examined the diagnosticity of the presence of either of two, three, four, or five cues.

Several of the cues in combination were indeed reliable predictors of complicated mappings, and several combinations differed reliably in standardized and collaborative interviewing. As Table 6 shows, when the first turn after a question contained both a filler and a pause longer than 1 s, it was reliably more likely to reflect a complicated mapping than a straightforward mapping in both interview types, $F(1, 40) = 4.72, p = .036$; the difference between interviewing types was not reliable. Similarly, as Table 7 shows, when the first turn after a question contained both a pause and a repair, it was reliably more likely to reflect a complicated mapping than a straightforward mapping in both interview types, $F(1, 40) = 4.67, p = .037$; again, the difference between interviewing types was not reliable. The same pattern held for first turns that contained both a report and a pause, first turns that contained both a filler and a hedge, and marginally for first turns that contained a report and a hedge.

Although turns containing three or four cues were relatively rare, some combinations were highly reliable predictors of complicated mappings. Turns that contained a report with at least one filler and at least one pause happened rarely (1.6% of question–answer sequences in standardized interviews and 2.4% in collabora-

<table>
<thead>
<tr>
<th>Scenario Type</th>
<th>Straightforward</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized</td>
<td>7.9</td>
<td>18.3</td>
</tr>
<tr>
<td>Collaborative</td>
<td>15.4</td>
<td>23.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario Type</th>
<th>Straightforward</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized</td>
<td>1.5</td>
<td>5.6</td>
</tr>
<tr>
<td>Collaborative</td>
<td>3.5</td>
<td>8.3</td>
</tr>
</tbody>
</table>
MISALIGNMENT CUES IN SURVEYS

tive interviews), but they only happened for complicated mappings, $F(1, 40) = 5.44, p = .025$, making them perfect predictors of need for clarification. Turns that contained reports, pauses, and fillers were perfect predictors of need for clarification in standardized interviews (3.2% for complicated scenarios, 0% for straightforward) and good predictors in conversational interviews (7.9% for complicated scenarios, 3.2% for straightforward), $F(1, 40) = 7.04, p = .011$. And turns that contained repairs, pauses, and fillers were good predictors of need for clarification both in standardized interviews (2.4% for complicated scenarios, 0.8% for straightforward) and in conversational interviews (11.1% for complicated, 4.9% for straightforward), $F(1, 40) = 4.83, p = .034$.

Some combinations of cues occurred more frequently in conversational interviews than in standardized ones: fillers and repairs; fillers and hedges; hedges and repairs; repairs and reports; reports, pauses, and fillers; fillers, pauses, and repairs; and, marginally, hedges and reports; and fillers, repairs, and reports. But in many cases this happened without differentiating need for clarification. For example, turns that included reports, hedges, and repairs happened only in collaborative interviews (1.6% of turns in complicated scenarios and 1.6% of turns in straightforward scenarios) and never in standardized interviews, $F(1,40) = 4.11, p = .036$.

There were also many combinations that did not reflect complicated mappings or interviewing type at all. In most cases it was because they did not occur frequently enough to provide any evidence. For those that included hedges, in several cases (hedges, repairs, and reports; hedges, fillers, and pauses; fillers, hedges, reports; fillers, pauses, and hedges), there were enough cases to see a pattern, but the pattern led to no reliable differences. So although it may seem that multiple cues in a single turn are likely to reflect need for clarification, this isn’t necessarily always the case, particularly when hedges are involved.

When we examined combinations of cues in the second sense—whether the presence of either of two, three, four, or five cues is diagnostic of need for clarification—we saw that every possible combination led to reliable differences, although often only for standardized interviews. For example, as Table 8 shows, respondents’ first turns after questions were far more likely to contain either one or more fillers or one

<table>
<thead>
<tr>
<th>Scenario Type</th>
<th>Straightforward</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interview</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized</td>
<td>14.3</td>
<td>34.8</td>
</tr>
<tr>
<td>Collaborative</td>
<td>45.8</td>
<td>50.2</td>
</tr>
</tbody>
</table>

TABLE 8
Percentage of Questions for Which Respondents Produced Either One or More Fillers or One or More Repairs in First Turn After Question, Excluding Explicit Requests for Clarification
TABLE 9
Percentage of Questions for Which Respondents Produced Either One or More Fillers, One or More Repairs, One or More Hedges, or a Report in First Turn After Question, Excluding Explicit Requests for Clarification

<table>
<thead>
<tr>
<th>Scenario Type</th>
<th>Straightforward</th>
<th>Complicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interview</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized</td>
<td>17.5</td>
<td>41.9</td>
</tr>
<tr>
<td>Collaborative</td>
<td>48.3</td>
<td>50.2</td>
</tr>
</tbody>
</table>

or more repairs for complicated mappings than for straightforward mappings, $F(1, 40) = 7.16, p = .011$, and far more likely to contain either one of these cues in conversational interviews than in standardized interviews, $F(1, 40) = 11.65, p = .001$. But the presence of either cue was only diagnostic of need for clarification in the standardized interviews, interaction $F(1, 40) = 5.18, p = .028$. That is, roughly half the time in conversational interviews the first turn contained either a filler or a repair whether the respondent's scenario was straightforward or complicated.

Although the statistical patterns of reliability were the same when we included pauses in the analysis—that is, higher rates predicted need for clarification—the rates were so high (from 50% to 80% in each case), because pauses were so frequent, that we discounted these combinations as good predictors. But the presence of at least one of the other four cues was highly predictive of need for clarification in standardized interviews. As Table 9 shows, respondents' first turns after questions were much more likely to contain either a filler, hedge, repair, or report for complicated mappings than for straightforward mappings, $F(1, 40) = 7.60, p = .009$, and more likely to contain them in conversational than standardized interviews, $F(1, 40) = 5.56, p = .023$. Yet the presence of either one of these cues was only predictive in standardized interviews, interaction $F(1, 40) = 9.28, p = .004$.

In practical terms, the predictive value of considering either of two cues seems just about as good whether the two being considered are fillers and repairs, fillers and reports, hedges and reports, fillers and hedges, and so forth; the mean percentages were not very different in each case. In fact, as a comparison of Tables 8 and 9 shows, adding in additional cues beyond these two did not seem to add much predictive value.

DISCUSSION

Even when survey respondents do not explicitly ask for clarification, their linguistic and paralinguistic behaviors in the very first turn after a question is asked can reliably indicate that they have circumstances that do not map onto questions in a
straightforward way and thus that they are likely to produce inaccurate answers. In particular, respondent repairs and reports are good indicators of the need for clarification, and in standardized interviews, so are pause length, hedges, and ums and uhs. Other potential indicators—discourse markers—were not particularly good indicators, at least in this corpus. Turns with more than one cue or turns with either of two cues (a filler or a repair, a filler or a report, a hedge or a report, etc.) were especially diagnostic of respondents’ likelihood of needing clarification.

Respondents’ discourse cues varied depending on the kind of survey interview they participated in: Respondents ummed, repaired, and reported more often in collaborative interviews than in strictly standardized interviews. Quite often, the discourse cues were only diagnostic of need for clarification in the standardized interviews. What accounts for this? On the one hand, the fact that the cues were more frequent in collaborative interviews suggests that respondents at some level were tailoring their speech, producing more cues particularly when those cues were likely to be usable by the interviewers. This is consistent with Oviatt’s (1995) finding that people are likely to be less fluent when speaking with another human than when speaking with a computer, which presumably cannot make use of the evidence the disfluencies present. On the other hand, the fact that the cues were often diagnostic of need for clarification only in the standardized interviews and not in the collaborative interviews argues against the idea that survey respondents always signal their need for clarification. Perhaps the greater disfluency and reporting in collaborative interviews reflects greater cognitive effort or greater accountability for answers rather than diagnostic signaling of need for clarification.

So are uncertainty cues useful for survey interviewers—and survey designers—to pay attention to? We argue that they are, for several reasons. As Mathiowetz (1998) proposed, uncertainty expressions in respondents’ speech may be a good data source for statistical imputation models in large survey data sets. In our study, several discourse cues predicted complicated mappings, which in general are more likely to be inaccurate. Given how rarely survey respondents ask for clarification when they really need it, these sorts of cues might allow interviewers to provide clarification more appropriately—that is, when respondents need it. Although strictly standardized interviewing, in which interviewers are prevented from providing clarification, is often held up as best practice (Fowler & Mangione, 1990), practices vary from survey center to survey center (Viterna & Maynard, 2002); presumably, it is just the sorts of cues examined in this study that interviewers rely on when they provide clarification because they intuitively sensed that respondents needed it. (Of course, if the cues become nondiagnostic once interviewers begin responding to them, this is problematic.)

These data provide an empirical basis to support interviewers’ intuitions about when respondents need clarification. Obviously, further analyses of other sorts of corpora would be needed to ensure that the findings from these interviews based on fictional scenarios generalize to interviews about respondents’ own lives or to
other discourse settings. More extensive studies of individual respondents’ patterns of using various discourse cues are also warranted; an um presented by someone who never ums should be far more diagnostic than an um presented by a constant ummer. Nonetheless, the data we presented here suggest that even when respondents do not explicitly ask for clarification, their discourse cues can provide real insight for interviewers into the respondents’ need for clarification. And the data should provide a basis for giving clarification in speech-based computerized interviewing systems of the future. Such systems might be able to use the information that these cues convey to choose when to provide users with unsolicited clarification, which has been shown to improve response accuracy in both computer-administered textual (Conrad & Schober, 1999) and speech surveys (Bloom, 1999; Schober, Conrad, & Bloom, 1999, 2000).

Beyond their utility for the realm of surveys, these findings raise more general questions about the communicative functions of disfluencies and about the nature of conceptual alignment. First, the fact that the same disfluencies are differentially diagnostic in standardized and collaborative interviews suggests that disfluencies function differently in different discourse contexts, depending on their audience and their likelihood of uptake. Audience and uptake have not been central to most views of disfluency; disfluencies have most often been seen as products of individuals in isolation, rather than parts of discourse that are under collaborative influence. The findings here are at least partially consistent with the idea that disfluencies can function as communicative signals (Clark, 1994).

Second, although the discourse cues examined here are sometimes diagnostic of need for clarification, our findings demonstrate that need for clarification is not necessarily accompanied by discourse cues. That is, when our survey respondents uttered completely fluent answers, this was no guarantee that they had understood the questions exactly as intended. Unfortunately, the data do not allow us to distinguish when conceptual misalignments will be accompanied by discourse cues and when they will not. One could hypothesize that more severe misalignments should be more likely to be marked, but a theory of degrees of misalignment remains to be worked out (see Tourangeau et al., 2004, for a proposal that the number of overlapping features in definitions marks degree of alignment). Certainly the fact that severe misalignments were not always marked in this corpus suggests that survey respondents may not take the time or have the metacognitive resources to notice that their own conceptualizations for ordinary words such as furniture or work could differ from survey designers’ conceptions.

This leads to a conundrum: How can interviewers use respondents’ discourse cues to improve data accuracy when respondents are confident in their incorrect interpretations of survey questions? Without evidence to go on, interviewers would have to present clarification along with every question. This would be costly and annoying, and would probably lead to respondents’ dropping out of surveys at rates that would be unacceptable for important official statistics.
(Conrad & Schober, 2000). The usual solution is for questions to be reworded based on extensive pretesting (e.g., Fowler & Mangione, 1990), which does indeed lead to fewer misinterpretations; but it is unlikely that all possible conceptual misalignments within and across populations of respondents can ever be uncovered through pretesting—note that the questions used in the corpus in this study had been extensively pretested. And, ignoring the problem will not make it go away; if we simply follow the usual procedure of allowing misinterpretation whenever respondents are confident, we run the risk of basing societal judgments on inadequate survey data.

We suspect that what we observed in these surveys holds true for conceptual alignment in conversation more generally. Although sometimes interlocutors present each other with evidence of need for clarification—explicit requests for clarification, discourse cues reflecting processing difficulties—it is entirely possible that interlocutors’ conceptualizations are misaligned far more frequently than they suspect, without any cues of the misalignment. Of course, many of these misalignments may not matter for the interlocutors’ purposes; people’s goals for understanding each other precisely are often far less stringent than those in a government survey about facts and behaviors. Our data allow us to be clearer about at least one situation when these misalignments matter: when the need for precision is high enough and when the circumstances to which the interlocutors are referring are nonprototypical with respect to the categories under discussion.

ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under Grants SBR-9730140 and ISS-0081550 and a dissertation fellowship from the New School for Social Research. We thank Susan Brennan, Fred Conrad, and Bill Hirst for their comments and Cezar Giosan and Eduardo Vega for help with coding and analyses.

REFERENCES


