Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 555–582

Disfluencies and Gaze Aversion in Unreliable Responses to Survey Questions

Michael F. Schober[1]*, Frederick G. Conrad*[2]*, Wil Dijkstra*[3]*, and Yfke P. Ongena*[4]

When survey respondents answer survey questions, they can also produce "paradata" (Couper 2000, 2008): behavioral evidence about their response process. The study reported here demonstrates that two kinds of respondent paradata – fluency of speech and gaze direction during answers – identify answers that are likely to be problematic, as measured by changes in answers during the interview or afterward on a post-interview questionnaire. Answers with disfluencies were less reliable both face to face and on the telephone than fluent answers, and particularly diagnostic of unreliability face to face. Interviewers' *responsivity* can affect both the prevalence and potential diagnosticity of paradata: both disfluent speech and gaze aversion were more frequent and diagnostic in conversational interviews, where interviewers could provide clarification if respondents requested it or the interviewer judged it was needed, than in strictly standardized interviews where clarification was not provided even if the respondent asked for it.

Key words: Respondent paradata; respondent cues of processing difficulty; interviewing mode; face-to-face interviewing; telephone interviewing; conversational interviewing; standardized interviewing.

1. Introduction

When survey respondents answer survey questions, they can provide information beyond the content of their answers. As Couper (2000, 2008) termed it, respondents provide *paradata* along with their answers (the survey data): extra evidence about their response process, and thus perhaps about the quality of their answers. Depending on the mode of the survey, different kinds of cues potentially constitute useful paradata (Conrad et al. 2008). For example, in a textual web survey a respondent's delay before answering can give evidence about how much trouble she is having answering the question (e.g., Conrad et al. 2007; Yan and Tourangeau 2008); in a telephone interview a respondent's *um*s and *uh*s can be informative about the extent to which she needs clarification (e.g., Schober and Bloom 2004), and her delays can signal various problems with answers (Bassili and Scott 1996; Draisma and Dijkstra 2004; Ehlen et al. 2007; Schaeffer and Maynard 2002). Paradata are almost certainly exploited by interviewers who adjust the tone or style of an interview to

[1] Department of Psychology, New School for Social Research, 80 Fifth Avenue, New York, NY 10003, U.S.A. Email: schober@newschool.edu. (Corresponding Author)
[2] Institute for Social Research, University of Michigan, Ann Arbor, MI, U.S.A.
[3] Social Research Methods, Free University of Amsterdam, Amsterdam, Netherlands.
[4] University of Groningen, Groningen, Netherlands
Acknowledgments: The authors thank Daniel Nielsen, Jan Smit, Gerhard Van de Bunt and Brady West for their help. This project was supported by the U.S. National Science Foundation Grant Nos. IIS-0081550, SES-0551294, and SES-1025645; the New School for Social Research; the Survey Research Center at the University of Michigan; and the Free University.

match the respondent's needs; they are also potentially exploitable by automated interviewing systems to provide tailored clarification or otherwise adapt to respondents (see papers in Conrad and Schober 2008).

Despite the general recognition that respondent paradata can be informative, survey researchers do not yet have a comprehensive body of knowledge about which kinds of paradata are dependable indicators of respondents' cognitive or emotional states, and under which circumstances. We propose that a careful analysis of the paradata available in different survey modes and the paradata that are produced in different interviewing techniques is needed to build on what is known thus far about details of interviewer-respondent interaction (see, e.g., Cannell et al. 1981; Dykema et al. 1997; Houtkoop-Steenstra 2000; Maynard et al. 2002; Oksenberg et al. 1991; Schaeffer 1991) and inform interviewer hiring, training and practice.

Such an analysis can potentially build on the larger body of evidence about paradata from studies of discourse in noninterview situations (although the term "paradata" is not used in these studies). For example, laboratory studies of answering trivia questions (Brennan and Williams 1995; Smith and Clark 1993; Swerts and Krahmer 2005) have demonstrated that paralinguistic displays (*um*s and repairs) and visual displays (eyebrow movement, smiles, gaze aversion, and "funny face" – diversion from a neutral expression) not only correspond with speakers' lack of confidence ("feeling of knowing") in their answers but can be used by observers to judge that confidence ("feeling of another's knowing"). Studies of other kinds of discourse demonstrate that disfluencies and speech errors can be evidence of speakers' planning and production difficulties (e.g., Fromkin 1973, 1980; Goldman-Eisler 1958; Levelt 1989) and of the complexity, conceptual difficulty or novelty of what they are trying to say (e.g., Barr 2003; Bortfeld et al. 2001; Fox Tree and Clark 1997).

But there is no guarantee that results from studies in other domains will generalize to survey interviewing situations. Survey respondents answer about their own behaviors and opinions rather than retrieving nonautobiographical facts from memory (answers to trivia questions) or referring to objects in scenes they are viewing (as in various psychology experiments). In addition, the particular nature of probing and questioning in subsequent dialogue takes a very particular form in interviews quite unlike other dialogue situations (see, e.g., Houtkoop-Steenstra 2000, papers in Maynard et al. 2002, and Schober and Conrad 2002), and quite unlike laboratory experiments that involve no dialogue. In survey interviews, respondents can have trouble answering for any number of reasons: they can have trouble recalling relevant information or deciding what they think, they can have comprehension problems (trouble knowing what the questioner intends by a term, trouble mapping the question concepts onto their personal circumstances), and they can have trouble formulating or articulating an answer. Any of these kinds of trouble could plausibly result in a problematic (unreliable or inaccurate) answer, and the associated processing difficulties might be evidenced in audio or visual paradata that are produced along with the answer – whether those are intentional communicative *signals* or unintended *symptoms* of processing difficulty (Clark 1996). Of course, problematic answers in surveys can be uttered without any potential indicators of trouble, and answers with potential indicators can be accurate; this is why additional research is needed to establish the relationship between how a survey answer is produced and the quality of that answer.

In the study reported here, our main research question is to what extent two kinds of respondent paradata – fluency of speech and direction of gaze – can diagnose or predict data quality of answers in a corpus of face to face (FTF) and telephone interviews asking nonsensitive factual and opinion questions. In particular, we ask whether the diagnosticity of these paradata is affected (a) by the mode of interviewing (telephone vs. FTF) and (b) by interviewers' responsivity to these paradata, that is, by whether interviewers clarify questions after respondents produce potential indicators of trouble.

Why Focus on Disfluencies and Gaze Aversion?

Disfluencies. Audio paradata in surveys include both linguistic and paralinguistic paradata. Linguistic paradata include words that respondents utter to explicitly inform the interviewer about their processing difficulty, the state of their comprehension (e.g., Mathiowetz 1998, 1999; Oksenberg et al. 1991) or their emotional state. For example, respondents can say that they didn't hear the question ("Could you repeat that?"), that they need clarification ("What do you mean by 'work for pay'?"), or that they feel uncomfortable ("I don't think I want to answer that question"). They can explicitly indicate various other kinds of reactions to the interview ("I never thought about that before"; "I have no idea"; "That's an interesting question"; etc.). They can also "report" rather than selecting a response option from those provided (see Drew 1984; Schaeffer and Maynard 2002, 2008; Schober and Bloom 2004), indicating a mismatch between the question and their circumstances; for example they might answer "I bought tires for a truck" rather than "yes" or "no" in response to the question "Last year, did you have any purchases or expenses for car tires?".

Paralinguistic paradata are the parts of respondents' answers that are not words. These can include speech disfluencies: *um*s and *uh*s (*em*s and *er*s in British transcriptions), pauses and hesitations either before or during an answer, repairs ("three- I mean two") and restarts ("thr- three"). They also include intonational contours: rising intonation in an answer may signal a respondent's uncertainty or need for clarification ("Three?"). Word stress can act as an implicit signal for the interviewer to correct what the respondent recognizes is a potential misinterpretation ("I bought *truck* tires"). Other acoustic cues can indicate emotional distress or irritation (see, e.g., Scherer 2003), and laughter can sometimes indicate discomfort with an answer (e.g., during an answer to a question about sexual behaviors), although it can also sometimes reflect and promote bonding and rapport with the interviewer (see Lavin and Maynard 2002).

We focus on disfluencies in particular for several reasons. First, as paralinguistic phenomena they are relatively frequent, unlike explicit linguistic paradata, which can be rarer; see, for example, Conrad and Schober (2000), in which respondents rarely explicitly requested clarification even when they needed it. Disfluencies are likely to be prevalent enough to allow statistical comparisons, and thus to be potentially practical on a large scale for interviewers or automated interviewing systems to exploit. Disfluencies have the advantage that there is relatively little ambiguity about their occurrence; rising intonation, in contrast, requires more complex measurement tools for researchers, and there may not be consistent agreement within linguistic subcultures about its meaning, as McLemore (1991) and Cameron (2001, pp. 112–114) have documented for speech styles with

frequent rising intonation ("uptalk" or "talking in questions"). Finally, speech disfluencies have been argued to occur in potentially problematic answers in telephone interviews (Draisma and Dijkstra 2004; Schaeffer et al. 2008; Schaeffer and Maynard 2002; Schober and Bloom 2004).

Gaze aversion. Less is known about visual than audio paradata in surveys. It is likely that global information about the respondent's appearance and demeanor can suggest whether the respondent is ready for and attending to the interview. The respondent's posture, for instance, leaning forward or leaning back, may give evidence of their attentiveness, nervousness, or engagement (Person, D'Mello and Olney 2008). As communication researchers have demonstrated in non-survey situations, respondents' facial expressions and head movements (furrowed brows, smiles, nods, head turns) potentially reflect (or explicitly signal) engagement, boredom, amusement, or confusion (see, e.g., Swerts and Krahmer (2005) on "audiovisual prosody" that reflects non-confidence in an answer to a trivia question). Eye movements have been shown to be particularly informative; direction of gaze can demonstrate what speakers are referring to (e.g., Hanna and Brennan 2007), when they are holding the floor or ready to pass the floor to another speaker (e.g., Goodwin 1991), or when they are searching for a word (Goodwin and Goodwin 1986). Gaze aversion – looking away from one's conversational partner – is another cue of potential utility in face to face interviews; several studies have demonstrated that people tend to avert their gaze while answering difficult questions (e.g., Doherty-Sneddon et al. 2002; Glenberg et al. 1998) or when they are not confident in their answers (Swerts and Krahmer 2005). The argument is that people avert the gaze of the questioner in order to temporarily eliminate visual (facial) information which might be distracting and hard to ignore.

In the current study we focus on gaze aversion in particular for two reasons. First, the empirical literature on gaze aversion in non-survey situations points in the same direction: listeners look away from the speaker when engaged in difficult cognitive tasks about whose outcome they lack confidence. It is plausible that this extends to survey settings and reflects respondents' processing difficulty. Second, unlike other visual paradata like facial expressions, gaze aversion is easy to observe without special training or aptitude, both for researchers and for interviewers; systematic coding of facial expressions, in contrast, can require extremely specialized knowledge, and interviewers may vary in face-reading skills. That is, interviewers might disagree on the meaning of a facial expression, but they are likely to agree, if they are paying attention, on whether a respondent has looked away during an answer.

Why Might Diagnosticity of Paradata Vary by Mode?

Audio paradata are transmitted in both FTF and telephone interviews, but the extent to which they are diagnostic of data quality may vary between modes. In telephone interviews, respondents only have the audio channel available for communication. They cannot assume that interviewers could possibly see their facial expressions or gaze direction; in fact, the notion of gaze aversion cannot even be defined when there is no interviewer from whom the respondent can avert their gaze. Thus it is only audio paradata that could be potentially diagnostic – at least for the interviewer – in telephone

interviews. In FTF interviews, respondents can display (intentionally or not) evidence of processing difficulties not only through the auditory channel but also visually, and so they *can* assume that attentive interviewers have access to additional (potentially diagnostic) visual paradata.

The availability of perceptible visual displays in FTF interviews could change the diagnosticity of audio paradata, because visual displays might replace some of the audio paradata in expressing or communicating processing difficulty. If so, then some moments of processing difficulty might be expressed only visually and not audibly, and so the audio displays would diagnose a smaller proportion of episodes of difficult processing in FTF than telephone interviews. Thus the audio paradata in the aggregate would end up being less informative because there are fewer observations. Alternatively, on those fewer occasions in FTF interviews when only audio displays are produced they might be particularly diagnostic because the respondent has not exploited alternative visual means of expressing or communicating processing difficulty, placing the communicative burden entirely on what is audible.

There is reason to hypothesize that audio and visual paradata complement each other. We know from other domains that visual signals – e.g., physically placing an object – can take the place of words (Brennan 1990, 2004; Clark and Krych 2004). Perhaps when interviewers and respondents cannot see each other, as on the telephone, respondents compensate for the lack of visual information by displaying more verbal cues of comprehension difficulty (cf. Whittaker 2003). Swerts and Krahmer (2005) found that visual and audio paradata together allow observers to make better judgments of question-answerers' confidence in their answers to trivia questions than either alone, but whether this generalizes to interviewing situations is unclear. As far as we know, there are no studies on whether visual paradata are always redundant with audio paradata in FTF survey interviews, or whether visual paradata replace (or further emphasize) audio paradata in FTF interviews.

Why Might Diagnosticity Vary by Interviewers' Responsivity?

Interlocutors in general – not just in interviews – can respond to each other's communicative displays quite subtly, picking up on and changing what they say based on their partner's gaze cues, fleeting facial expressions, vocal signs of uncertainty or approval, and so on (see, e.g., Clark 1994, 1996; Goodwin 1991; Schegloff 1984, 1998; Schober and Brennan 2003 among many others). This raises the possibility that how an interviewer reacts to a respondent's audio and visual display could affect the kinds of display that a respondent produces, and thus the extent to which the corresponding paradata are diagnostic of the respondent's processing difficulty. In fact, Schober and Bloom (2004) have demonstrated exactly this in analyses of audio paradata in a corpus of telephone interviews in which respondents answered about fictional scenarios. In the current study we therefore compare the diagnosticity of respondent paradata under two different interviewing techniques: (1) one which encourages attention to and substantive reaction to respondent behaviors that could suggest need for clarification (e.g., "It sounds like you're having some trouble. What can I help you with?"), and (2) one which allows only nonsubstantive reactions (e.g., "let me repeat the question") to respondents' explicit or implicit requests for clarification.

The interviewing technique that encourages substantive reaction to any evidence that a respondent may need clarification is what we have called "conversational" interviewing (Conrad and Schober 2000; Schober and Conrad 1997; Schober et al. 2004). As further detailed below, interviewers using this approach are trained to say what they believe is required to ensure the respondent understands what the survey designers mean by the terms in their questions; interviewers should provide definitions when explicitly asked for them, and they should offer definitions whenever they get the sense that clarification might be helpful. Although the training does not discuss respondent paradata, interviewers are instructed to attend to anything in what a respondent says or does that might suggest that clarification is needed, whether it has been requested or not.

We contrast this with an interviewing technique that requires nonsubstantive reactions: strictly standardized interviewing, following Fowler and Mangione's (1990) prescriptions. In this technique, interviewers are required to administer nondirective probes like "let me repeat the question" when respondents explicitly ask for clarification, and they are expressly forbidden from providing substantive definitions. (The logic is that clarifying words in a question for some respondents would mean that not all respondents would receive the same stimulus). Although Fowler and Mangione (1990) do not explicitly mention respondents' audio or visual displays, their technique would prohibit interviewers from providing a definition in response to spoken or visual potential indicators of trouble.

Respondents' paradata may be differently diagnostic of the respondent's processing difficulty in conversational than in standardized interviews. The potential for a conversational interviewer to help when respondents provide evidence of their processing difficulty may increase respondents' likelihood of displaying such evidence (intentionally or not). This could accurately inform conversational interviewers about respondents' needs more often than in standardized interviews. It is, of course, entirely possible that audio or visual displays are produced by respondents in the same ways no matter how interviewers react; it is also possible that the effects of interviewer reaction may differ FTF and on the telephone. The current study allows us to find out.

Measuring Quality of Answers

To assess the diagnostic value of paradata in the current study, we need to measure which answers are problematic. In this study, respondents answer questions about their own lives rather than fictional scenarios (as they did in Schober and Bloom 2004), and so we cannot measure response accuracy (validity) directly as we have in prior laboratory studies (Conrad et al. 2007; Ehlen et al. 2007; Schober and Conrad 1997; Schober et al. 2004). Instead, we measure two different kinds of (un)reliability: (1) response change (or its complement, consistency) during a question-answer (Q-A) sequence, that is, the respondent first answers the question and then changes the answer before the interviewer asks the next question, and (2) change or consistency between responses in the interview and responses to the same questions in a self-administered post-interview questionnaire where definitions accompany the questions.

The logic for (1) is that answers that change during a Q-A sequence (with or without interviewer-provided clarification) are clearly problematic, even if we don't know whether the original or changed answer (or neither) is correct. At the very least changed answers of

this sort reflect a lack of commitment to the original answer and potential uncertainty about which answer to provide. The logic for (2) follows that used in Conrad and Schober (2000): if answers change when definitions are provided (beyond the rate of answer change when no definitions are provided), this is evidence that initial (mis)interpretations have been corrected by the definitions. So response change (unreliability) when the respondent is presented with a definition is evidence that the earlier answer had been problematic. A consistent (reliable) response when the respondent is presented with a definition post-interview is evidence that the earlier answer was nonproblematic.

Note that the technique we use for assessing problematic answers intentionally supplements the wording of the re-asked questions in the post-interview questionnaire by adding definitions. This means that respondents who encountered a definition during a conversational interview will experience the same question and definition in the questionnaire; respondents who did not encounter a definition during the interview (either in standardized interviews or in conversational interviews in which they were not presented with a definition) will be encountering these post-interview definitions for the first time. It is these differences that allow us to assess whether respondents' interpretations of the questions in the original interview were consistent with the definitions presented in the post-interview questionnaire, and thus allow us to measure data quality of their original answers. In Conrad and Schober (2000) this interpretation of response change was supported by evidence that answers for which respondents elaborated their thinking (providing lists of purchased items) were more likely to fit what the survey definitions required when clarification had been provided.

2. Study

This study was carried out in a laboratory, as opposed to field, setting to guarantee suitable video views and audio quality for subsequent analysis, and to make sure that the physical setting was fully comparable in all conditions.

Interviewers were randomly assigned to conduct either strictly standardized or conversational interviews, either on the telephone or FTF; this led to four experimental groups. A total of eight experienced professional Dutch interviewers (all female) participated, with two interviewers assigned to each of the four experimental groups; interviewers had prior experience in both FTF and telephone interviewing. Each interviewer conducted five or six interviews for a total of 42. Respondents were Dutch university students (15 males, 27 females, mean age 22.3 ranging from 19 to 28 years) who were paid roughly the equivalent of US $25 to participate. There were eleven respondents in each of the two standardized groups (FTF and telephone) and ten respondents in each of the two conversational groups (FTF and telephone). The 42 respondents were randomly assigned to one of the four experimental treatments. All interviews were conducted in Dutch and carried out at the Free University of Amsterdam in June of 2000.

Interviewer Training

Interviewers were recruited to participate in a methodological study. They were told that they would be video-recorded for scientific purposes, to improve the quality of survey data collection.

Concepts. All interviewers were trained on the survey concepts being measured in each question (see Appendix A). This primarily involved a supervisor, who was blind to which interviewing technique the interviewer would be implementing, assessing interviewers' competence with the definition for each concept through mock interviews.

Interviewing Technique. Interviewers were then trained in one interviewing technique or the other. Standardized interviewers were trained to strictly follow the prescriptions of Fowler and Mangione (1990). Interviewers were required to read the question as worded; if the respondent did not provide an adequate answer, that is, did not select one of the response options presented with the question, the interviewer was instructed to administer a nondirective probe such as "Let me repeat the question" or "Is that a 'Yes' or a 'No?'" If the respondent requested clarification, the interviewers could only respond with nondirective probes such as "Whatever it means to you."

The instruction for conversational interviewers followed the approach of Schober and Conrad (1997) and Conrad and Schober (2000). In this technique interviewers also were to read the question as worded, but they could subsequently provide clarification if respondents explicitly asked for it or if in the interviewer's judgment the respondent seemed to need it. Interviewers were instructed to say whatever seemed necessary for the respondent to understand as intended, all or part of the definition, verbatim or in the interviewer's own words. Interviewers were not given any special instructions about attending or responding to visual or verbal evidence of difficulty answering.

Experimental Setting

In all of the interviews, the questionnaire was displayed on a laptop computer in front of the interviewer. She read aloud the questions from the computer and entered answers into the computer. The definitions of the survey concepts were printed on a sheet of paper available to the interviewer during the session. For the telephone interviews, the interviewer and respondent were situated in separate buildings. For the FTF interviews, the interviewer and respondent were seated at a table in the same room. All interview sessions, whether conducted on the telephone or in person, were video recorded with separate images of the interviewer's and the respondent's faces. In the FTF sessions, an additional video image was recorded of both parties together, so that we could determine where they were looking and when they were looking at each other.

Survey Questions

The questionnaire consisted of 18 questions, seven of which concerned nonsensitive facts or behaviors (e.g., student status, employment status, and membership in clubs) and eleven of which explored respondents' opinions (six questions about asylum seekers and five about illegal aliens). (See Appendix A for the full list of questions in English translation). In order to assess the impact of definitions on response change, we administered a paper questionnaire after the interview in which respondents were asked to answer the same questions they had just answered except the first five (for these five questions we believed we would have access to official records for students that would have allowed us to assess response validity by comparing access to those records, even if official records can themselves have errors in them; unfortunately, this access ultimately was denied

for reasons beyond the authors' control). The questions in the paper questionnaire were accompanied by the definition for the relevant concept (see Appendix A). Respondents completed the questionnaire alone in the same room in which they had been interviewed; an experimenter entered the room to provide the questionnaire, and returned to the room when the respondent had finished.

3. Results

Interviewing Techniques

Before getting to analyses of the paradata, we first verified that the two interviewing techniques had indeed been implemented as interviewers had been trained and that the corpus of interviews had the characteristics of conversational and standardized interviews seen in prior studies (Conrad and Schober 2000; Schober and Conrad 1997; Schober et al. 2004) that would make it suitable for answering our research questions. Interviews were first transcribed and checked by a second transcriber to make sure that all disfluencies were accurately represented in the transcript, including all *um*s and *uh*s (*em*s and *eh*s in Dutch), perceptible pauses (judged by the transcribers as perceptible), repairs (immediate replacements of sounds, words or phrases) and (immediate) restarts. Pauses were notated with periods enclosed within parentheses, and repairs and restarts were notated with double dashes (- -). Interviews were then segmented into Q-A sequences: all behavior from the point at which the interviewer began to ask a question until the interviewer began to ask the next question. Each transcript was coded by one of 3 coders for functional events in the interview (e.g., asking the question, requesting clarification, providing an answer, repeating the question, providing clarification) using a coding scheme (see Appendix B); coders recorded their decisions in Sequence Viewer 4 (Dijkstra 2006; http://www.sequenceviewer.nl/) to allow the interaction and paradata analyses reported below.

To additionally verify reliability of transcription of disfluencies, 150 Q-A sequences (20%) were randomly selected from the total 756 sequences, equally distributed across telephone versus FTF and conversational versus standardized interviews, for independent transcription by a different researcher. Comparisons of these verification transcripts with the original transcripts revealed high reliability of the count of number of functional events with speech disfluencies (Pearson's $r = .946$) and of the number of speech disfluencies per Q-A sequence (which takes into account that in some events there may be more than one speech disfluency) (Pearson's $r = .933$). Reliability of the coding for functional events was measured through extra coding of the same 150 randomly selected Q-A sequences by an independent coder, and it proved to be *substantial* by Landis and Koch's (1977) criteria (Cohen's kappa = 0.743).

As a first piece of evidence on the suitability of the corpus for testing our research questions, interviewers provided clarification more often in conversational interviews (for an average of 33.5% of the questions per interview) than in standardized interviews (for an average of 0.5% of the questions per interview), $F(1,38) = 224.27$, $P < .0001$ (see Table 1). Clarification was provided at the same rates in telephone (16.8%) and FTF (17.2%) interviews, $F(1,38) = 0.02$, ns, and the differences in clarification rates for conversational and standardized interviewing did not differ in the different modes, interaction $F(1,38) = 0.36$, ns. All four conversational interviewers provided clarification

Table 1. Percent of questions per interview for which interviewers provided clarification (SE in parentheses)

	Telephone	FTF	Overall
Standardized	1.0 (2.1)	0 (2.1)	0.5 (1.5)
Conversational	32.7 (2.3)	34.3 (2.3)	33.5 (1.6)
Overall	16.8 (1.6)	17.2 (1.6)	17.0 (1.1)

at least some of the time but not all of the time, ranging from 21% to 44% of the Q-A sequences in the interviews they administered; this is consistent with their training to provide clarification when, in their judgment, clarification was needed. Thus we could be confident that interviewers implemented the technique as they had been trained.

A second piece of evidence that the corpus was suitable is that clarification did indeed affect data quality as in our prior studies. As Table 2 shows, for the questions included in the post-interview questionnaire (Questions 6–18), 89.3% of final answers were the same in the interview and in the post-interview questionnaire when conversational interviewers had provided a definition during the interview (that is, an average of 10.7% of answers changed in the post-interview questionnaire which included definitions). As expected, these answers were significantly more reliable than final answers in those Q-A sequences in which conversational interviewers hadn't provided clarification (67.2%), within-subjects $F(1,38) = 17.80$, $P < .001$, and in standardized interviews (78.1%), in which interviewers almost never provided clarification, between-subjects $F(1,39) = 6.11$, $P < .02$. There were no differences in reliability between telephone and FTF interviews, nor was there any interaction with interviewing technique.

A third piece of evidence on the suitability of the corpus is that conversational interviews took longer than standardized interviews, as one would expect when clarification (which takes time) is given versus when it is not. As Table 3 shows, Q-A sequences lasted 28.2 seconds on average in conversational interviews, but 16.4 seconds in standardized interviews, $F(1,38) = 61.0$, $P < .001$. Interview duration was no different in FTF and telephone modes (unlike in Groves and Kahn 1979), nor did interviewing technique interact with mode, $Fs < 1$.

Thus we are confident that the interviewers had administered the two interviewing techniques as intended and that our analyses of audio and visual paradata in the two techniques would be based on interviews with the qualities we expected.

Table 2. Reliability of final answers, Qs 6–18 (SE in parentheses)

	Telephone	FTF	Overall
Standardized	77.6 (4.5)	78.6 (4.5)	78.1 (3.2)
Conversational, Q-A sequences without clarification*	62.0 (6.0)	72.4 (5.6)	67.2 (4.1)
Conversational, Q-A sequences with clarification*	89.4 (4.9)	89.2 (4.7)	89.3 (3.4)

* This is a within-subjects comparison

Table 3. Q-A sequences' duration in secs (SE in parentheses)

	Telephone	FTF	Overall
Standardized	17.2 (1.5)	15.7 (1.5)	16.4 (1.1)
Conversational	28.1 (1.5)	28.3 (1.5)	28.2 (1.1)
	22.6 (1.1)	22.0 (1.1)	22.3 (0.8)

Paradata

We first verify that the potential indicators of trouble we are measuring are indeed frequent enough in the sample to ask our research questions. Note that this also gives practical evidence on whether those indicators are frequent enough that interviewers or automated interviewing systems could in principle benefit from exploiting them.

We then examine diagnosticity of the paradata. Our presentation of the findings on diagnosticity reflects the diagnostic problem that interviewers face: given an answer that includes potential indicators of trouble, how likely is it to be a good answer? An alternative analytic approach is to ask whether problematic answers are more likely to include diagnostic cues of response difficulty than nonproblematic answers, as in Schober and Bloom (2004). We have analyzed this data set in both directions (with paradata as independent and dependent variables) and both sets of analyses show essentially the same pattern of results.

For ease of exposition, we first report results about disfluencies, and then about gaze aversion.

Disfluencies

Prevalence of speech disfluencies. We coded every *um* and *uh*, perceptible pause within and between conversational turns, and every repair and restart in each Q-A sequence, through a Sequence Viewer utility that automatically assigned a code based on the notations in the transcript. We treated *um* and *uh* as instances of the same thing, although, as Clark and Fox Tree (2002) note, they may indicate different kinds of trouble.

Our first question was whether respondents produced disfluencies at different rates in telephone and FTF interviews. As Table 4 shows, counting all disfluencies – *um*s and *uh*s, pauses, and repairs and restarts – respondents produced at least one disfluency in their answer (wherever it appeared in the Q-A sequence) in a greater percentage of the Q-A sequences in telephone interviews (57.0%) than they did FTF (42.1%), $F(1,38) = 10.56$, $P = .002$. (Throughout, the patterns of results are the same whether one counts *um*s and

Table 4. Percent of Q-A sequences that included at least one respondent disfluency, that is, ums and uhs, pauses, and repairs and restarts (SE in parentheses)

	Telephone	FTF	Overall
Standardized	53.7 (4.5)	33.8 (4.5)	43.8 (3.1)
Conversational	60.3 (4.7)	50.5 (4.7)	55.4 (3.3)
Overall	57.0 (3.2)	42.1 (3.2)	

*uh*s or all disfluencies; we will report on all disfluencies, but note that the great majority of disfluencies – 78.9% – were *um*s and *uh*s). The overall pattern of a higher rate of disfluencies over the telephone than FTF is consistent with what has been found in studies of telephone conversation of other kinds (Williams 1977).

Unexpectedly, the rate of disfluencies on the telephone was higher than has been observed in studies of other kinds of discourse in which speakers could only hear each other (e.g., Bortfeld et al. 2001, who observed a rate of about 6 disfluencies per 100 words in a laboratory card-matching task in which participants could not see each other). Here respondents' rate of *um*s and *uh*s during their answer on the telephone was 12.2 per 100 words, reliably higher than the FTF rate of 6.4 per 100 words, $F(1,38) = 12.22$, $P = .001$ (see Table 5); no other effects of interviewing technique or interactions were significant. Disfluency rates varied substantially between different questions; for example, respondents were particularly disfluent (19.1 *um*s and *uh*s per 100 words at some point during the Q-A sequence) while answering the question about how many methods courses they had taken (Q7), compared to a rate of 5.7 per 100 words for Q1-Q3. To the extent that disfluencies reflect processing difficulty, this makes sense; answering Q7 involves demanding mental operations: recalling many courses, determining whether each qualifies, and incrementing a running tally, while Q1-Q3 simply require choosing one of two response options (e.g., whether one is a "full time" or "part time" student).

Interviewing technique also affected the prevalence of disfluencies. Respondents produced at least one disfluency during their answer in a significantly greater percentage of Q-A sequences in conversational interviews (55.4%) than in standardized interviews (43.8%), $F(1,38) = 6.46$, $P = .015$ (see Table 4). This is consistent with the pattern for disfluencies in the (telephone) interviews in Schober and Bloom (2004) and supports the proposal that the interviewer's responsivity can actually change the prevalence of disfluencies. There was no reliable interaction between interviewing mode and interviewing technique.

Can these findings be explained by the influence of individual interviewers? It is, in principle, possible that different interviewers elicited different rates of respondent disfluency, although it is difficult to imagine what interviewer behavior might be involved in such an effect. Nonetheless, if interviewers differ in the respondent disfluency rates with which they are associated and if those with higher rates happened to have been assigned to the telephone or conversational interviewing conditions, this could explain the disfluency results which we are attributing to mode and interviewing technique. To examine this possibility, we computed ρ_{int} for respondent *um* and *uh* rate. This statistic (also labeled "rho-int") was developed by Kish (1962) to measure the degree to which variance (usually

Table 5. Rate of respondent ums and uhs per 100 words (SE in parentheses)

	Telephone	FTF	Overall
Standardized	14.3 (1.6)	5.8 (1.6)	10.0 (1.1)
Conversational	10.2 (1.7)	7.1 (1.7)	8.7 (1.2)
Overall	12.2 (1.2)	6.4 (1.2)	9.3 (0.8)

of responses but in our case disfluency rates) is correlated with individual interviewers (see Biemer and Lyberg (2003) for an introduction).

We calculated ρ_{int} from a mixed model ANOVA consisting of four independent variables: respondents, interviewers, mode and interviewing technique, in which respondents were nested within interviewers and interviewers were nested within mode and interviewing technique. At first blush, interviewer variance for this measure was large (.069), but this is almost entirely attributable to the experimental treatments (mode and interviewing technique) rather than individual interviewers. That is, when we re-run these analyses removing mode and interviewing technique from the model, that is, carrying out a more pure test of different effects of individual interviewers, interviewer-related variance becomes so small that ρ_{int} is effectively zero, despite the fact that small numbers of interviewers can inflate ρ_{int} values. The bottom line is that it seems to be the treatments and not individual interviewers that are driving disfluency rates.

Diagnosticity of disfluencies: Reliability during Q-A sequence. As the first row of Table 6 shows, respondents overall were more likely to change their first answer during the Q-A sequence when it included a disfluency (changing on average 9.8% of their answers) than when it did not (2.1%), $F(1,38) = 11.68$, $P = .002$. These findings are based on a threeway ANOVA with one within-subjects factor, disfluency (present or absent), and two between-subjects factors, mode (telephone or FTF) and interviewing technique (standardized or conversational); as all respondents produced at least one answer with a disfluency, all 42 respondents are included in this analysis.

The diagnosticity of disfluencies during the first answer in the Q-A sequence varied by mode of interviewing. In particular, disfluencies during this first answer were significantly more diagnostic in FTF interviews (14.5% rate of change for disfluent answers vs. 1.6% for fluent answers) than in telephone interviews (5.1% rate of change for disfluent answers vs. 2.6% for fluent answers), $F(1,38) = 5.30$, $P = .027$ for the interaction of disfluency and mode (see Table 6 for the full set of means and SEs from this analysis). The diagnosticity of these disfluencies also varied (marginally) by interviewing technique. If we compare diagnosticity of disfluencies between conversational and standardized interviews, collapsing across telephone and FTF interviews, disfluencies were marginally more diagnostic in conversational interviews (15.4% rate of change for disfluent answers vs. 3.6% for fluent answers) than in standardized interviews (4.1% rate of change for disfluent

Table 6. Unreliability of responses: percent of initial answers changed during Q-A sequence (SE in parentheses)

	Fluent	Disfluent
Overall	2.1 (0.7)	9.8 (2.2)
Telephone	2.6 (1.0)	5.1 (3.1)
Standardized	1.1 (1.4)	1.6 (4.2)
Conversational	4.0 (1.4)	8.6 (4.4)
FTF	1.6 (1.0)	14.5 (3.1)
Standardized	0.0 (1.4)	6.7 (4.2
Conversational	3.3 (1.4)	22.3 (4.4)

These analyses exclude the three listing questions (Q4, Q5 and Q6) for which response change during an answer cannot be unambiguously coded because it is unclear when an initial response is unreliable or simply partial.

answers vs. 0.6% for fluent answers), $F(1,38) = 3.39$, $P = .073$ for the interaction of disfluency and interviewing technique. No other interactions were statistically significant.

Diagnosticity of disfluencies: Reliability of answers as measured post-interview. Recall that in this corpus conversational interviewing led to more reliable answers (as measured post-interview) than standardized interviewing particularly in those cases where the conversational interviewers provided clarification; when they did not, answers were no more reliable. Thus if disfluencies are diagnostic of unreliable answers (as measured post-interview), they should predict response change in those cases where respondents' interpretations were *not* corrected during the interview, that is, in conversational interviews when clarification was not given and in standardized interviews. When clarification had been given, disfluencies in the original answer should not predict response change, because the problems diagnosed by the disfluency should have been resolved by the clarification.

This was exactly the pattern observed. In order to compare reliability for disfluent and fluent answers in conversational interviews where no clarification had been given and in standardized interviews, we carried out a threeway ANOVA with one within-subjects factor, disfluency (present or absent), and two between-subjects factors, mode (telephone or FTF) and interviewing technique (standardized or conversational without clarification). If we collapse the data for all respondents included in the analysis, the overall pattern is that in both cases (standardized interviews and conversational interviews where no clarification had been given) respondents' disfluent answers were more likely to be unreliable (32.1%) than their fluent answers (21.9%), $F(1,36) = 4.55$, $P < .05$. The means and SEs for all experimental conditions are presented in Table 7A. As expected, this did not vary by interviewing technique (conversational interviews without clarification are essentially standardized) or by mode, nor were there any interactions.

In contrast, disfluencies were no longer predictive of post-experiment response change when interviewers *had* provided clarification in conversational interviews. This can be seen when we compare reliability for disfluent and fluent answers in conversational interviews where clarification had been given and in standardized interviews, in a threeway ANOVA with one within-subjects factor, disfluency (present or absent), and two between-subjects factors, interviewing technique (standardized or conversational with clarification) and mode (telephone or FTF). As Table 7B shows, disfluent answers in the conversational interviews with clarification were 100% reliable (0% response change on

*Table 7A. Unreliability of final answers, Qs 6–18, compared to answers on post-interview questionnaire: percent of changed answers (SE in parentheses)**

	Fluent	Disfluent
Standardized (n = 21)	15.5 (5.3)	30.3 (4.8)
Telephone	16.3 (7.7)	26.3 (7.0)
FTF	14.8 (7.4)	34.4 (6.7)
Conversational interviews with Q-A sequences without clarification (n = 19)	28.2 (5.6)	34.0 (5.1)
Telephone	32.4 (8.1)	36.4 (7.4)
FTF	24.0 (7.7)	31.5 (7.0)

* These analyses include all respondents but two, who either were not disfluent or did not receive clarification.

*Table 7B. Unreliability of final answers, Qs 6–18, compared to answers on post-interview questionnaire: percent of changed answers (SE in parentheses)**

	Fluent	Disfluent
Standardized (n = 21)	15.5 (4.7)	30.3 (4.8)
Telephone	16.3 (6.9)	26.3 (6.9)
FTF	14.8 (6.5)	34.4 (6.6)
Conversational interviews with Q-A sequences with clarification (n = 8)	18.7 (7.7)	0.0 (7.7)
Telephone	37.5 (10.8)	0.0 (10.9)
FTF	0.0 (10.9)	0.0 (10.9)

* These analyses include those respondents in conversational interviews who had at least one Q-A sequence with a disfluent answer followed by clarification.

the post-experiment questionnaire, versus 18.7% response change for fluent answers), while disfluent answers in standardized interviews were unreliable 30.3% of the time (compared to 15.5% response change for fluent answers), interaction of disfluency and interviewing technique $F(1,25) = 6.95$, $P = .014$.

Altogether, these results show that speech disfluencies are indeed diagnostic of unreliable answers. They are frequent enough to be useful, and they are produced in predictably different ways in different modes (respondents were more likely to be disfluent during an answer on the telephone than FTF) and with differential interviewer responsivity (respondents were more likely to be disfluent in conversational than standardized interviews). And by two different measures of unreliability, answers with disfluencies were more likely to be unreliable. First, they were more likely to change within the Q-A sequence. Second, they were more likely to be corrected post-survey when respondents were provided with clarification – unless respondents had already been provided with clarification during the interview itself.

Gaze Aversion

Prevalence of gaze aversion. The video recordings of the FTF interviews allowed clear views of when respondents looked away from interviewers, turning their heads and averting their gaze (see Figure 1). (Of course we could not examine respondents' direction of gaze in the telephone interviews because the respondent was alone in the room without an interviewer so there was no stable fixation point from which to measure deviation). The start of gaze aversion was defined by eye movement away from the interviewer; the precise moment in time (to within one video frame) at which gaze aversion started could be unambiguously measured by moving the video one frame backwards or forwards. Based on double-coding of a sample of 79 randomly selected Q-A sequences (20% of all Q-A sequences in FTF interviews, with roughly half in conversational and half in standardized interviews), measurement was indeed unambiguous; the two coders' identification of the number of instances of gaze aversion correlated $r(79) = .990$, $P < .0001$, and measures of the duration of gaze aversion correlated $r(79) = .996$, $P < .0001$.

Based on this measurement, there were 65 identifiable Q-A sequences in the 21 FTF interviews in which there was at least one instance of gaze aversion. Almost all respondents (19 of 21) averted their gaze at least once during an answering phase, and many did so on

Fig. 1. Respondent (right) averting gaze from interviewer while answering question. (Fotographer: Wil Dijkstra, VU University, Amsterdam)

several questions, up to a maximum of eleven questions. Note that this creates a smaller sample than for the audio paradata, which were observable in both telephone and FTF interviews, but with enough statistical power to carry out a parallel set of analyses.

Respondents averted their gaze at least once during a greater percentage of their answers in conversational interviews (24.7%) than in standardized (11.4%) interviews, $F(1,19) = 5.16$, $P = .035$. Thus, as with the audio paradata, it seems that interviewing technique affects how often respondents produce this visual display. Certainly different interviewing techniques lead to different opportunities to produce visual indicators of trouble; conversational interviews are longer because they sometimes include the presentation of definitions, and so there is simply more time in which gaze aversion could occur. It is also possible that respondents in a FTF conversational interview use gaze aversion to display communication difficulty, much as in ordinary interaction – because interviewers, like ordinary conversational partners, can react substantively to evidence of need for clarification. As was the case with audio paradata, there is no evidence that different interviewers elicited different amounts of gaze aversion: ρ_{int} was effectively zero for FTF interviewers.

Diagnosticity of Gaze Aversion: Reliability During Q-A Sequence

The evidence is that gaze aversion did indeed predict unreliability of answers within a Q-A sequence. Among the 21 FTF interviews, there were 17 respondents (9 conversational and 8 standardized) who produced at least one answer with gaze aversion, which allowed us to compare reliability of answers with and without gaze aversion within-subjects. To do this,

Table 8. Unreliability of responses: percent of initial answers changed during Q-A sequence, FTF interviews (SE in parentheses)

	No gaze aversion	Gaze aversion
Overall	4.3 (1.8)	24.7 (8.5)
Standardized	1.0 (2.7)	18.8 (12.3)
Conversational	7.6 (2.5)	30.6 (11.6)

Analysis based on the 17 FTF respondents who produced at least one answer with gaze aversion.

we carried out a two-way ANOVA with one within-subjects factor, gaze aversion (present or absent), and one between-subjects factor, interviewing technique (standardized or conversational). As Table 8 shows, answers with gaze aversion were more likely to be unreliable within the Q-A sequence (24.7%) than answers without gaze aversion (4.3%), $F(1,15) = 4.94$, $P < .05$. The pattern was the same in both interviewing techniques, interaction $F(1,15) = 0.08$, *n.s.*, although perhaps we would see an interaction with a larger sample.

Diagnosticity of Gaze Aversion: Reliability of Answers As Measured Post-interview

Unlike disfluencies, gaze aversion did not predict unreliable answers between the interview and the post-experiment questionnaire. Answers with gaze aversion were no more likely to be unreliable (20.6%) than answers without gaze aversion (24.2%), $F(1,15) = 0.29$, n.s. Following our earlier logic, gaze aversion should predict response change only in the cases where interviewers had not provided clarification: in standardized interviews and in conversational interviews without clarification. Unfortunately we have too few cases for the full within-subjects comparisons we were able to do for disfluencies, but we can compare the cases where interviewers did not provide clarification in both kinds of interviewing. In this comparison, answers with gaze aversion were no more likely to be unreliable (27.3%) than answers without gaze aversion (27.3%), $F(1,15) = 0.0$, ns. And there was no evidence for an effect of interviewing technique on diagnosticity: in conversational interviews 26.3% of answers were unreliable with gaze aversion versus 32.3% without, and in standardized interviews 28.3% of answers were unreliable with gaze aversion versus 22.3% without, interaction $F(1,15) = 0.40$, n.s.

On the other hand, there were five respondents in conversational interviews for whom we could compare (within-subjects) the rate of unreliability for answers in which they averted their gaze and received clarification versus the rate for answers where they averted their gaze and did not receive clarification; the other respondents did not avert their gaze and both receive and not receive clarification. When these five respondents exhibited gaze aversion and received clarification, the rate of unreliable answers (0%) was significantly lower than the rate (32.8%) when they exhibited gaze aversion and did not receive clarification ($F(1,4) = 7.98$, $P < .05$). This is consistent with the notion that gaze aversion followed by clarification leads to more reliable answers than gaze aversion not followed by clarification. So at least part of the logic about unreliability of answers with gaze aversion as measured post-interview holds for a very small sample of respondents, but with only five respondents we see this result as more suggestive than conclusive.

Altogether, these results show that for one kind of visual paradata – gaze direction – respondents in FTF interviews were more likely to avert their gaze during an answer in an interview where the interviewer could provide clarification than in one where the interviewer couldn't. Answers with gaze aversion were more likely to be unreliable within the Q-A sequence than answers without gaze aversion. Answers with gaze aversion were not more likely to be unreliable as measured post-interview, in contrast to disfluencies for which there was such an effect.

4. Discussion

The findings in this study demonstrate that two kinds of respondent paradata – fluency of speech and the direction of gaze during answers to survey questions – can provide evidence about data quality in face to face interviews, and that speech disfluencies can provide evidence about data quality in both face to face and telephone interviews. For both interview modes, answers with these behaviors were more likely to be of poorer quality. The findings extend evidence from other domains of interaction that utterances with these behaviors are more likely to be problematic (unreliable, unconfident, wrong) than utterances without them. They also extend the related Schober and Bloom (2004) finding on speech disfluencies into interviews about autobiographical information and into FTF interviews.

Regarding our first research question, whether the diagnosticity of speech disfluencies is affected by the mode of interviewing (FTF vs. telephone), the evidence is clear. Although answers with disfluencies were less reliable in both modes, disfluencies were particularly diagnostic of unreliability FTF. Disfluencies were also less frequent in FTF interviews than on the phone, possibly because respondents have visual channels for displaying response difficulty beyond audio.

Regarding our second research question, the current findings demonstrate that in both FTF and telephone interviews the interviewer's ability to respond when the paradata indicate trouble affects the respondent's likelihood of indicating that trouble. That is, respondents produced more disfluencies and averted interviewers' gazes more often during answers in conversational interviews, a technique in which interviewers were trained to provide clarification if they got the sense that respondents needed it. And the evidence was that this was not an effect of individual interviewers' somehow eliciting more disfluencies, but rather the result of the experimental treatment – a more collaborative interviewing style that promotes clarification. To our knowledge this provides the only evidence thus far that an interlocutor's potential uptake increases a speaker's likelihood of producing a disfluency or averting gaze. (Oviatt (1995) found that speakers were more likely to be disfluent when speaking to another human than to a computer, but this could be the case for many reasons besides the interlocutor's potential uptake.)

How might these findings be usefully applied to reduce measurement error in survey interviews? We propose several different possibilities, each of which would require additional research in order to be effectively implemented. First, one could imagine implementing new selection criteria for interviewers to hire those who are intuitively able to recognize and make use of visual and auditory evidence of response difficulty. It is possible that current hiring practices already favor interviewers who are interpersonally

sensitive on multiple fronts, including the ability to attend to a respondent's audio and visual displays; but it is an empirical question whether this is in fact the case. If so one could imagine making the practice more deliberate.

Second, one could imagine explicitly training already-hired interviewers to detect and make use of the presence of these behaviors, assuming that attentiveness to them can be trained (an open question). Interviewers could be trained to use whatever interviewing techniques are available to them when they encounter evidence of a problematic answer, from additional neutral probing to engaging in clarification dialogue to resolve the trouble. Training materials could be created from existing audio and video recordings of interviews, demonstrating which kinds of verbal and visual behaviors are informative about problematic answers and what the possible subsequent interviewer actions might be.

If such attentiveness turns out not to be easily trainable (interpersonal skill does seem to vary across interviewers), one could imagine designing automated real-time support for helping less sensitive interviewers to recognize potential need for clarification, either for training or production purposes. For example, one could design automated speech recognition systems to monitor and provide evidence to interviewers about delays in the respondent's speech or *um*s and *uh*s (see Ehlen et al. 2007, for a preliminary system of this sort); one could design automated vision tools that could inform an inattentive interviewer about a respondent's gaze aversion, for example processing the video feed in a videomediated interview, or even from an interviewer's laptop in a FTF interview. With such tools, one could even imagine fully automated detection of gaze direction or speech disfluencies in an automated interviewing system. This, of course, would require additional knowledge about whether respondents avert gaze or produce disfluencies in the same way with an automated partner as with a human interviewer.

The findings in this study open the door to additional research on the uses of paradata in interviews and interviewing systems. First, beyond speech fluency and gaze direction it is plausible that other paradata – for example, response latency, vocal stress and tone, facial expressions, gestures, and posture, among others – are systematically related to the quality of responses. Which of these occur frequently enough to be useful, and how universally they are diagnostic across different respondent cultures, dialects, and individual expressive styles, is unknown. We assume that the base rates of potentially diagnostic behaviors – either within an interview mode or technique, across a culture, in an individual, or across different topics (see, e.g., Schachter et al. 1991) – are likely to be important factors in judging the utility of any particular instance of paradata. That is, an *um* produced by a respondent who never *um*s, or averted gaze by a respondent who mostly stares right at the interviewer, should be far more informative about the respondent's cognitive or interactive processes than an *um* produced by a respondent who is chronically disfluent or averted gaze by a respondent who barely maintains eye contact with the interviewer.

Another important arena for additional research is the extent to which different paradata co-occur or supplement one another, and the extent to which they replace each other. In our data set there is a hint that the co-occurrence of audio and visual paradata in FTF interviews is particularly diagnostic: Among the 36 sequences (of 315 FTF sequences) that involved both gaze aversion and disfluency, 9 (25%) resulted in answers that were unreliable between the interview and post-experiment questionnaire. The percentage of unreliable answers was notably lower among the 75 sequences that involved disfluencies

alone, where seven answers (9%) were unreliable; among the 13 sequences that involved gaze aversion alone, where one answer was unreliable (a rate of 8%); and among the 191 sequences involving neither disfluency nor gaze aversion, where only two of the answers (1%) were unreliable. Although these are so few cases that we would not want to conclude too much from them, they nonetheless are consistent with the possibility that answers that include displays in more than one channel may be particularly problematic.

Further research is also needed on whether the diagnosticity of different paradata varies for different kinds of questions than those examined here: open-ended questions that require more speech planning, sensitive or personal distress questions for which respondents may feel a greater need to present themselves in a positive light, or particularly complex and difficult questions that require deeper thought. We hypothesize that, in general, the prevalence and diagnosticity of behaviors that provide evidence of trouble answering will be greater for questions for which respondents must construct answers on the fly. And based on our findings, we assume that the diagnosticity of particular paradata is likely to vary in different modes. Given the proliferation of new modes and platforms of interviewing beyond FTF and telephone, it will be important to understand the availability and diagnosticity of different paradata in modes that implement survey dialogue differently, from videomediated interviews to web surveys to speech-IVR interviews, on desktop or mobile multimodal devices, and more.

Presumably not every piece of paradata is revealing about the accuracy or reliability of the speaker's utterance, nor about the speaker's affect or motivation or confidence. The practical challenge for survey researchers will be to understand when interventions that make use of respondent paradata – either by interviewers or automated interviewing systems during the interview itself, or in subsequent data analysis – lead to improved data quality. The theoretical challenge will be to map out, in different domains and styles of discourse, when which paradata are informative of which cognitive and affective states.

Appendix A: Questions and Definitions

Question 1
The first questions in this interview are about your education.
Are you a full-time or part-time student?

1. full-time
2. part-time

definition:
Whether a student is called a part-time of full-time student depends on the official registration form. This seems logical, but many part-time students (officially) participate for whatever reason in the full-time program, and consequently consider themselves (incorrectly) full-time students.

Question 2
Is this a full or shortened course of study?

1: normal
2: reduced

definition:
no definition

Question 3
What is your year of study?
definition:
The registration date determines which year of study a student is in. For instance, if a student was registered as a student by September 1999, he/she is a first year student. This also holds for students who participate in the shortened program (2 instead of 4 years), because the exemptions are based upon prior education (completed outside the Faculty of Social-Cultural Sciences). If a regular student takes up a second course of study, exemptions count. For instance, if a student decides to take up a second course of study and he/she is exempted from the first year, he/she is called a second year student.

Question 4
(not posed to freshmen: 17 respondents)
What is your field of study? [more than one answer is possible]
definition:
no definition

Question 5
Which methodological/statistical courses have you completed during your course of study?
definition:
English:
A course is considered an M&T (methodological/statistical) course when an employee of the Research Methodology Department teaches it and this department is responsible for the course.
In order to complete a course a student must sit for and pass an exam. The course is also considered as completed when a student is exempted from the course due to previous education at another institution.

Question 6
Now I will ask some questions about your membership in clubs.
Can you name all the clubs in which you are a member?
definition:

- An 'association' is a legal entity (local authorities and natural persons are legal entities as well).
- An association has members and aims for a certain goal which need not be idealistic.
- A person cannot be the owner of the association; there is no owner.
- An association has a non-profit seeking goal.
- Any profit may not be divided among its members but should be spent on the goal of the association.
- An association is normally established by a notarial deed, containing the articles of vereniging.

- Members of the board as well as of the association according to certain provisions bear personal responsibility for debts and the like.
- Membership is personal (unless it is stated otherwise in the articles of association).
- The members of the board are normally nominated from the members by the general meeting. Each member has a right to vote.
- Within six months (11 at most) the board should publish an annual report, including a financial report.

To mention:

- personal membership
- non-profit seeking goal, no division of profit among members, profit should be spend on the goal
- general meeting of members, board, annual report
- no owner

Question 7
How many paid jobs on the side have you had since July 1, 1999?
definition:
A respondent can have a job on the side only if the job is not his/her main activity. The number of jobs on the side depends on the number of employment contracts. If multiple duties are mentioned in one contract only one job is counted. In the case of multiple employers but the same kind of job multiple jobs are counted. In the case of moonlighting there is no legal contract and therefore no job. An employment contract is simply nothing more than a written or oral agreement between employer and employee.

Question 8
I would like to present some statements about asylum seekers and illegal aliens in the Netherlands. First I will present some questions about asylum seekers. We would like to know to what extent you agree or disagree with these statements. You have the following alternatives to choose from: fully agree; agree; neither agree nor disagree; disagree; fully disagree.
Asylum seekers come to Europe because they are in danger in their own country.

1: fully agree
2: agree
3: neither agree nor disagree
4: disagree
5: fully disagree

definition:
An asylum seeker is a person who irrespective of the reason, which can vary a great deal, seeks asylum in The Netherlands. Reasons may be:

- political and religious reasons,
- social-economical reasons,
- ethnic reasons and/or
- social reasons.

An illegal alien is

- a person who is refused asylum, has no status as a recognized fugitive and doesn't have permission to stay in The Netherlands
- a person who never applied as an asylum seeker and stays in the Netherlands without permission (except for holidays), or
- a "white illegal" person

A white illegal person

- is an undocumented alien who has worked for six continuous years in the Netherlands (and is able to show and prove this), and who has a social security number and valid passport.
- Until 1 January 1998 they were qualified for a residence permit.
- Each case is treated separately.
- A "white illegal" person is an illegal alien until he/she obtains the status of recognized fugitive.

For all remaining questions interviewers presented the same response alternatives (1–5) as those for Question 8, and the same definitions were used.

Question 9
Asylum seekers come to Europe to profit from welfare.

Question 10
The Netherlands should close its borders to all asylum seekers.

Question 11
Asylum seekers should make more efforts to adjust to Dutch norms.

Question 12
The areas surrounding asylum seekers' centers are unsafe.

Question 13
The Netherlands should receive asylum seekers with political grounds with open arms.

Question 14
The following statements are about illegal aliens and not about asylum seekers any more. We would like to know to what extent you agree or disagree with these statements. You have the following alternatives to choose from: fully agree; agree; neither agree nor disagree; disagree; fully disagree.
There is enough room in our country for everyone.

Question 15
Illegal aliens should not receive food stamps.

Question 16
Illegal aliens have rights, too.

Question 17
Illegal aliens should not be discriminated against.

Question 18
All illegal aliens deserve the same rights as Dutch citizens.

Appendix B: Coding Scheme for Functional Events

Respondent:

(1) Answers a question from the questionnaire (e.g.,: "I'm a member of a tennis club," "I'm not a member of any club"
(2) Answers question (or gives information) relevant to the definition (e.g., "I make a contribution," "There is an annual meeting")
(3) Any don't know answer (e.g., "Don't know if there is an annual meeting")
(4) Request clarification
(5) Standalone filler (*em* or *eh*)
(6) Answer other question, e.g., about other characteristics (e.g., "It's in Amsterdam", "The name is X")
(7) Report (describe circumstances) (e.g., "I play tennis")
(8) Request repeat of survey question/present survey question for confirmation
(9) Repeat previous answer at request of interviewer
(10) No more information (e.g., "That's all")
(11) Other, including confirmation of other's utterances and own repetitions

Interviewer:

(1) Read question exactly as worded (include corrected disfluencies)
(2) Read question with change in wording
(3) Repeat question or part of question
(4) Paraphrase question (re-present question or parts of question, deviating from original wording)
(5) State response alternatives
(6) Neutral probe (e.g., "whatever it means to you," "we need your interpretation," "let me repeat the question," "anything else?", "take your time to think"
(7) Read definition verbatim
(8) Paraphrase parts of definition (includes answering respondent's question about definition) (e.g., "A club has an annual meeting," "A sports club is also a club"
(9) request information from respondent pertaining to definition ("Do you make a contribution?" "Is there an annual meeting?", "is that a real club?")
(10) request description of other characteristics ("What is the name of the club?")
(11) Repeat/restate/elaborate respondent's answer
(12) Request repetition of answer from questionnaire
(13) Back channel (e.g., "uh-huh," "okay")
(14) Other

9. References

Barr, D.J. (2003). Paralinguistic Correlates of Conceptual Structure. Psychonomic Bulletin & Review, 10, 462–467.

Bassili, J.N. and Scott, B.S. (1996). Response Latency as a Signal to Question Problems in Survey Research. Public Opinion Quarterly, 60, 390–399.

Biemer, P. and Lyberg, L. (2003). An Introduction to Survey Quality. Hoboken, NJ: Wiley.

Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.R., and Brennan, S.E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. Language and Speech, 44, 123–149.

Brennan, S.E. (1990). Seeking and Providing Evidence for Mutual Understanding. Unpublished doctoral dissertation, Stanford University.

Brennan, S.E. (2004). How Conversation is Shaped by Visual and Spoken Evidence. In Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions, J.C. Trueswell and M.K. Tanenhaus (eds). Cambridge, MA: MIT Press, 95–130.

Brennan, S.E. and Williams, M. (1995). The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to Listeners About the Metacognitive States of Speakers. Journal of Memory and Language, 34, 383–398.

Cameron, D. (2001). Working with Spoken Discourse. Thousand Oaks, CA: SAGE Publications, Inc.

Cannell, C.F., Miller, P.V., and Oksenberg, L. (1981). Research on Interviewing Techniques. In Sociological Methodology, S. Leinhardt (ed.). San Francisco: Jossey-Bass, 389–437.

Clark, H.H. (1994). Managing Problems in Speaking. Speech Communication, 15, 243–250.

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

Clark, H.H. and Fox Tree, J.E. (2002). Using Uh and Um in Spontaneous Speaking. Cognition, 84, 73–111.

Clark, H.H. and Krych, M. (2004). Speaking While Monitoring Addressees for Understanding. Journal of Memory and Language, 50, 62–81.

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

Conrad, F.G. and Schober, M.F. (2008). Envisioning the Survey Interview of the Future. Hoboken, NJ: Wiley.

Conrad, F.G., Schober, M.F., and Coiner, T. (2007). Bringing Features of Dialogue to Web Surveys. Applied Cognitive Psychology, 21, 165–187.

Conrad, F.G., Schober, M.F., and Dijkstra, W. (2008). Cues of Communication Difficulty in Telephone Interviews. In Advances in Telephone Survey Methodology, J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster (eds). New York: Wiley, 212–230.

Couper, M.P. (2000). Usability Evaluation of Computer Assisted Survey Instruments. Social Science Computer Review, 18, 384–396.

Couper, M.P. (2008). Designing Effective Web Surveys. New York: Cambridge University Press.

Dijkstra, W. (2006). Sequence Viewer, version 4. Available at: http://www.sequenceviewer.nl.

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., and Doyle, C. (2002). Development of Gaze Aversion as Disengagement from Visual Information. Developmental Psychology, 38, 438–445.

Draisma, S. and Dijkstra, W. (2004). Response Latency and (para)Linguistic Expressions as Indicators of Response Error. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: Wiley.

Drew, Paul (1984). Speakers' Reportings in Invitation Sequences. In Structures of Social Action: Studies in Conversation Analysis, J.M. Atkinson and J. Heritage (eds). New York: Cambridge University Press, 129–151.

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

Ehlen, P., Schober, M.F., and Conrad, F.G. (2007). Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces. Discourse Processes, 44, 245–265.

Fox Tree, J.E. and Clark, H.H. (1997). Pronouncing "The" as "Thee" to Signal Problems in Speaking. Cognition, 62, 151–167.

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

Fromkin, V.A. (1973). Speech Errors as Linguistic Evidence. The Hague, Netherlands: Mouton.

Fromkin, V.A. (1980). Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand. New York: Academic Press.

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

Goldman-Eisler, R. (1958). Speech Production and the Predictability of Words in Context. Quarterly Journal of Experimental Psychology, 10, 96–106.

Goodwin, C. (1991). Conversational Organization: Interaction Between Speakers and Hearers. New York: Academic Press.

Goodwin, M.H. and Goodwin, C. (1986). Gesture and Coparticipation in the Activity of Searching for a Word. Semiotica, 62(1/2), 51–75.

Groves, R.M. and Kahn, R.L. (1979). Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic Press.

Hanna, J.E. and Brennan, S.E. (2007). Speakers' Eye Gaze Disambiguates Referring Expressions Early During Face-to-Face Conversation. Journal of Memory and Language, 57, 596–615.

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

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. Journal of the American Statistical Association, 57, 92–115.

Landis, J.R. and Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33, 159–174.

Lavin, D. and Maynard, D.W. (2002). Standardization vs. Rapport: How Interviewers Handle the Laughter of Respondents During Telephone Surveys. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley, 335–364.

Levelt, W.J.M. (1989). Speaking: From Intention to Articulation. Cambridge, MA: MIT Press.

Mathiowetz, N.A. (1998). Respondent Expressions of Uncertainty: Data Source for Imputation. Public Opinion Quarterly, 62, 47–56.

Mathiowetz, N.A. (1999). Respondent Uncertainty as Indicator of Response Quality. International Journal of Public Opinion Research, 11, 289–296.

Maynard, D.W., Houtkoop-Steenstra, H., Schaeffer, N.C., and van der Zouwen, J. (2002). Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview. New York: Wiley.

McLemore, C.A. (1991). The Pragmatic Interpretation of English Intonation: Sorority Speech. Unpublished doctoral dissertation, University of Texas, Austin.

Oksenberg, L., Cannell, C., and Kalton, G. (1991). New Strategies for Pretesting Survey Questions. Journal of Official Statistics, 7, 349–365.

Oviatt, S. (1995). Predicting Spoken Disfluencies During Human-Computer Interaction. Computer Speech and Language, 9, 19–35.

Person, N.K., D'Mello, S., and Olney, A. (2008). Toward Socially Intelligent Interviewing Systems. In Envisioning the Survey Interview of the Future, F.G. Conrad and M.F. Schober (eds). New York: Wiley, 195–214.

Schachter, S., Christenfeld, N., Ravina, B., and Bilous, F. (1991). Speech Disfluency and the Structure of Knowledge. Journal of Personality and Social Psychology, 60, 362–367.

Schaeffer, N.C. (1991). Conversation with a Purpose – or Conversation? Interaction in the Standardized Interview. In Survey Measurement and Process Quality, P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: John Wiley, 367–391.

Schaeffer, N.C. and Maynard, D.W. (2002). Occasions for Intervention: Interactional Resources for Comprehension in Standardized Survey Interviews. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley, 261–280.

Schaeffer, N.C. and Maynard, D.W. (2008). The Contemporary Standardized Survey Interview for Social Research. In Envisioning the Survey Interview of the Future, F.G. Conrad and M.F. Schober (eds). New York: Wiley, 31–57.

Schaeffer, N.C., Dykema, J., Garbarski, D., and Maynard, D.W. (2008). Verbal and Paralinguistic Behaviors in Cognitive Assessments in a Survey Interview. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Schegloff, E.A. (1984). On Some Gestures' Relation to Talk. In Structures of Social Action: Studies in Conversation Analysis, J.M. Atkinson and J. Heritage (eds). Cambridge: Cambridge University Press, 266–298.

Schegloff, E.A. (1998). Body Torque. Social Research, 65, 535–596.

Scherer, K.R. (2003). Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication, 40, 227–256.

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

Schober, M.F. and Brennan, S.E. (2003). Processes of Interactive Spoken Discourse: The Role of the Partner. Handbook of Discourse Processes, A.C. Graesser, M.A. Gernsbacher, and S.R. Goldman (eds). Mahwah, NJ: Lawrence Erlbaum Associates, 123–164.

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

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

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

Smith, V.L. and Clark, H.H. (1993). On the Course of Answering Questions. Journal of Memory and Language, 32, 25–38.

Swerts, M. and Krahmer, E. (2005). Audiovisual Prosody and Feeling of Knowing. Journal of Memory and Language, 53, 81–94.

Whittaker, S. (2003). Mediated Communication. In Handbook of Discourse Processes, A.C. Graesser, M.A. Gernsbacher, and S.R. Goldman (eds). Mahwah, NJ: Erlbaum, 243–286.

Williams, E. (1977). Experimental Comparisons of Face-to-Face and Mediated Communication: A Review. Psychological Bulletin, 84, 963–976.

Yan, T. and Tourangeau, R. (2008). Fast Times and Easy Questions: The Effects of Age. Experience and Question Complexity on Web Survey Response Times. Applied Cognitive Psychology, 22, 51–68.

Received October 2011
Revised July 2012