APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **21**: 165–187 (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.1335

Bringing Features of Human Dialogue to Web Surveys

FREDERICK G. CONRAD[1]*, MICHAEL F. SCHOBER[2]
and TANIA COINER[2]

[1]*Institute for Social Research, University of Michigan, Ann Arbor, USA*
[2]*Department of Psychology, New School for Social Research, New York, USA*

SUMMARY

When web survey respondents self-administer a questionnaire, what they are doing is in many ways similar to what goes on in human–human interviews. The studies presented here demonstrate that enabling web survey respondents to engage in the equivalent of clarification dialogue can improve respondents' comprehension of questions and thus the accuracy of their answers, much as it can in human–human interviews. In two laboratory experiments, web survey respondents (1) answered more accurately when they could obtain clarification, that is, ground their understanding of survey questions, than when no clarification was available, and (2) answered particularly accurately with mixed-initiative clarification, where respondents could initiate clarification or the system could provide unsolicited clarification when respondents took too long to answer. Diagnosing the need for clarification based on respondent characteristics—in particular, age—proved more effective than relying on a generic model of all respondents' need for clarification. Although clarification dialogue increased response times, respondents preferred being able to request clarification than not. The current results suggest that bringing features of human dialogue to web surveys can exploit the advantages of both interviewer- and self-administration of questionnaires. Copyright © 2007 John Wiley & Sons, Ltd.

An increasingly important mode of collecting survey data is for respondents to self-administer questionnaires via a web browser. Much of the recent research on web surveys has been concerned with design and layout issues, for example, how multiple questions on a single page versus single pages per question affect answers (Dillman, 1999, pp. 395–396), or how arrows between questions and lines for entering text affect answers (Christian & Dillman, 2004). Research of this sort extends the design considerations that have long been discussed for paper self-administered questionnaires.

We propose that features of the web not available in paper questionnaires enable a new way of thinking about the design and self-administration of survey questions. In particular, we propose that web survey interaction can be conceived as a dialogue consisting of turns of interaction between a user (respondent) and the system (the interviewing agent). On this view, which is inspired by the collaborative view of comprehension in cognitive psychology and psycholinguistics (Clark, 1996), each move by the system—presenting a question, prompting for an answer—and each move by the respondent—clicking to proceed, reading a question, typing a number as an answer—corresponds to some physical

*Correspondence to: Frederick G. Conrad, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48104 USA. E-mail: fconrad@isr.umich.edu

or mental action in a spoken interview. Conceiving of the interaction this way not only highlights and clarifies the function of each move, but it also opens up the possibility that survey designers could implement web survey systems that include some of the advantages of human–human interview dialogue.

What exactly does such a view add? The picture from human–human dialogue is that saying something doesn't guarantee it will be understood. People engage in extra turns of dialogue to make sure that what the speaker intended has been understood—to ground their understanding (Clark & Brennan, 1991; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989). People ground their understanding to a criterion sufficient for their current purposes; in casual conversations (e.g. at a cocktail party), people may not need to understand precise details to satisfy their conversational goals, but in other settings (e.g. air traffic control conversations, calls to a technical help desk when your computer crashes, or conversations with your ex-spouse about child visitation) the stakes are higher. As people ground understanding, they also attempt to tailor their utterances and interpretations to their partner's knowledge states and discourse styles (Krauss & Fussell, 1996; Schober & Brennan, 2003).

This collaborative view of human conversation differs from traditional accounts of language use (what Akmajian, Demers, Farmer, & Harnish, 1990 called the 'message model' of communication), where listeners interpret utterances directly. The traditional view is that the meaning of an utterance is contained within the words themselves, and that the process of comprehension involves looking up those meanings in the mental dictionary and combining them appropriately (Conrad & Rips, 1986). A collaborative view argues that accurate comprehension also requires dialog so that people can clarify what is meant (Clark, 1996; Pickering & Garrod, 2004, and commentaries for discussion about the cognitive processes underlying dialogic interaction).

We argue that the standard way of designing web survey pages implicitly (perhaps unintentionally) embodies a message model of communication, in that it assumes that presenting a question is the same thing as its being understood appropriately. We propose that building into web surveys the potential for respondents to ground their understanding of the meaning of words in questions may allow for more accurate survey data collection that rests more firmly on principles of human cognition and interaction.

In the studies reported here, we thus investigate how a collaborative view of human conversation transfers to interaction with computers, in particular web-based questionnaires about non-sensitive facts and behaviours of the sort collected in censuses and many scientific surveys, and we examine whether a collaborative view can improve user interface design. In particular, we focus on the following intertwined features of collaboration in dialogue:

- Two-way initiation of clarification sequences: in spoken dialogue, both parties (the speaker and the addressee) can diagnose whether the addressee has understood appropriately and can thus initiate the speaking turns required to make clear what particular words are intended to mean.
- Variable criteria of understanding: in dialogue, people whose criteria of understanding are high (who believe understanding correctly really matters), and who know they lack needed information, are more likely to initiate a clarification sequence.
- Tailoring of clarification to one's partner's particular needs: people in dialogue don't treat the same action by every partner as meaning the same thing, but tailor their assessments and responses. A delay in response by a quick partner may mean the partner needs clarification, but a delay by a slow partner probably doesn't.

DOI: 10.1002/acp

A side benefit of this line of investigation is that it forces us to specify details of a collaborative view that can test its limits and refine our theories of human communication.

MIXED-INITIATIVE CLARIFICATION IN SURVEYS

In the two studies reported here, we contrast two approaches to bringing dialogue-like clarification to web survey questionnaires. Under one approach, clarification is user- (i.e. respondent-) initiated: if the user explicitly requests clarification, the system provides it. This requires users to recognise that they need clarification and to be willing to ask for it. Under the other approach, clarification is mixed-initiative: the system also provides (or offers to provide) clarification when it diagnoses misunderstanding, based on user behaviour. For example, in a desktop or speech interface a system could provide clarification when the user takes too long to act; in a speech interface a system could provide clarification when the user's speech is hesitant or disfluent (containing *um*s and *uh*s, restarts, etc.).

Current dialog systems for web surveys sometimes allow user-initiated clarification. For example, the US Census in 2010 may offer web-based data collection in which respondents can click for the definition of residence—what is meant by 'live or stay' in 'Do any of these people live or stay somewhere else?'. But to our knowledge no production survey systems volunteer clarification to respondents.

One reason that even user-initiated clarification is rare in current systems is that this violates the spirit of strict standardisation developed for human–human interviews, where the interpretation of questions should be left entirely up to respondents (Fowler & Mangione, 1990). The argument for standardisation is that if interviewers help respondents to interpret questions, they might influence responses, but if interviewers read scripted questions and provide only 'neutral' feedback, responses are less likely to be biased. We have demonstrated, in contrast, that in human–human (telephone) interviews even supposedly non-biasing feedback by interviewers can affect responses (Schober & Conrad, 1997, 2002). More importantly, strict standardisation can actually harm data quality in interviews because it prevents respondents from grounding their understanding of the questions. This is a problem because people's interpretations of seemingly straightforward questions like 'How many bedrooms are there in your house?' can vary enormously; without grounding their understanding of questions, respondents may conceive of questions in unintended ways, and the resulting data may not fulfil the survey designers' purposes (Clark & Schober, 1991). We have shown that respondents in strictly standardised telephone interviews may answer less accurately than respondents in more collaborative interviews where respondents can ground their understanding of questions with the interviewers (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad, & Fricker, 2004).

One important thing to note about web surveys is that respondents' interest in making sure they understand questions as the survey designers have intended is often low. We suspect this is because respondents usually do not initiate the survey dialogue; they provide information to the system rather than retrieving information from the system, as they do with a database query system or a web search interface. Users may thus care less about precisely understanding the words in survey questions (misunderstanding has few consequences for the user) than understanding the words in an on-line job application or an on-line health claims form (where misunderstandings can be costly). Initial empirical

DOI: 10.1002/acp

evidence for this account can be found in Schober, Conrad, Fricker and Ehlen's (2003) laboratory study in which users either answered survey questions (*provided* information) or conducted a comparable web search (*obtained* information); users who *obtained* information requested clarification by clicking on highlighted links almost twice as often as the users who *provided* information.

EXPERIMENTAL METHODS

In the studies reported here we assess whether text-based web interviewing systems that enable users to clarify survey concepts actually improve comprehension of questions (and thus improve response accuracy), as a collaborative approach would predict. We also examine the effects of clarification on task duration—clarification probably takes more time. In the context of web surveys, for which response rates are typically low, longer task durations are likely to discourage respondents who have begun the questionnaire from finishing, and are thus undesirable. Finally, we examine the effects of clarification on user satisfaction; even if clarification improves comprehension, it could be annoying.

Both experiments used a desktop interface, in which the computer displayed questions in a web browser[1] and the user entered responses with the keyboard and mouse (Figure 1). When users were allowed to request clarification, they did so by clicking on highlighted text (Figure 1a, "live" appears in blue and is underlined). In response to such requests, the system displayed a definition below the question (Figure 1b). When clarification was not available, the user interface was identical to what appears in Figure 1 except that no text was highlighted or linked to definitions. In both experiments the system could initiate clarification in some conditions. This was triggered by periods of respondent inactivity (no typing or clicking) that exceeded a specified interval (the inactivity threshold). Inactivity is one possible cue of need for clarification in interviews (Bassili & Scott, 1996; Conrad, Schober, & Dijkstra, in press; Schober & Bloom, 2004) and one that can be easily detected in web-based data collection (see Couper, 2000, on collecting 'paradata' in web surveys). Unsolicited definitions were displayed by the system exactly as they were when respondents requested them (Figure 1b).

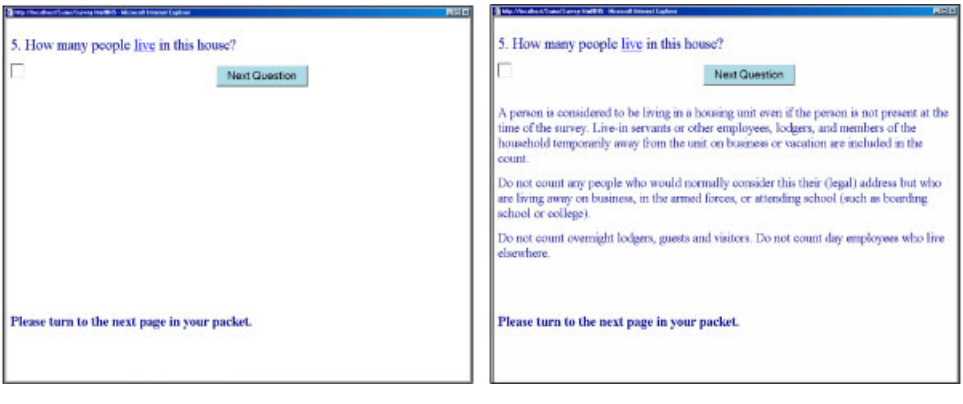

Figure 1. Survey question with hyperlink (a) before and (b) after definition is displayed

[1]In the first experiment, the interface was developed specifically for the study, so was not a commercial browser; in the second experiment, the interface was a commercial browser (Internet Explorer).

DOI: 10.1002/acp

In both studies, all respondents were asked non-sensitive factual/behavioural questions from on-going US government surveys. For each question, the survey designers had developed official definitions for the key concepts, which clarified whether, for example, a floor lamp should be considered a piece of household furniture, or whether a student away at college should be considered to be living at home.

Respondents answered these questions on the basis of fictional scenarios, so that we could measure response accuracy: there was a correct answer for each scenario based on the official definition for a concept in the question. For each question there were two alternate scenarios, one straightforward and one complicated. With the straightforward scenario, the survey question was designed to be easy for respondents to interpret—to map onto the respondent's (fictional) circumstances in a straightforward way. For example, the question 'How many people live in this house?' was presented with either the following straightforward or complicated scenario:

Straightforward

The Gutierrez family owns the 3-bedroom house at 4694 Marwood Drive. The family has four members: Maria and Pablo Gutierrez, and their two children Linda and Marta. There is one bedroom for Maria and Pablo, one for Marta, and one for Linda.

Complicated

The Gutierrez family owns the 3-bedroom house at 4694 Marwood Drive. The family has four members: Maria and Pablo Gutierrez, and their two children Linda and Marta. There is one bedroom for Maria and Pablo, one for Marta, and one for Linda. Linda is a college student. Although her legal address is still 4694 Marwood Drive, she stays at the college dorms all year, except for holidays and vacations.

The straightforward scenario describes a conventional nuclear family with two parents and two children. The complicated scenario concerns the same family except that one of the children is a college student living on campus. Without knowing the definition, it is ambiguous whether the child at college should be counted as living in the home; the definition (second paragraph, Figure 1b) clarifies that this child should not be included in the count.

For each question, the respondent answered on the basis of either the straightforward or complicated version of each scenario; half of the scenarios presented to each respondent were straightforward and half were complicated. Respondents were given as much time as they needed to study and become familiar with the scenarios before beginning the experimental session, and all of the scenarios were available to the respondents while they answered questions.

After respondents had answered the experimental survey questions, they were presented with 15 to 17 questions (depending on their experimental condition) about their experience with the particular user interface to which they had been assigned. Most of these questions concerned their satisfaction, including an assessment of the overall experience (rated from 'very bad' to 'very good') and their preference for interviewer- versus computerised self-administration or paper self-administration of similar questionnaires in the future (presented as 'yes'–'no' choices). Respondents used the mouse to click and enter their responses to the user satisfaction questions.

DOI: 10.1002/acp

EXPERIMENT 1: RESPONDENTS' AWARENESS OF NEEDING CLARIFICATION

In this experiment we examined how enabling web survey respondents to ground their understanding affects the accuracy of their answers and their subjective experience, and how respondents' awareness of their need for clarification affects their likelihood of clicking for definitions. Survey respondents are often not deeply vested in the accuracy of their answers; if they misinterpret a question and consequently misreport, there are no direct consequences for them. Their primary goal is likely to involve finishing the interview or data collection session. Given respondents' goals, requesting clarification may seem tangential. Yet, exposure to definitions can be essential for respondents to understand questions as they are intended, at least in telephone interviews (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober et al., 2004). For web survey respondents to initiate a clarification dialogue they must recognise the possibility that they interpret the question differently than intended, they must value interpreting it correctly and it must be easy for them to obtain clarification.

There is little reason, *a priori*, that a respondent would expect ordinary words in a survey question to be used with extra-ordinary meanings. After all, respondents are invited to participate in surveys, and so it would be reasonable for them to assume the questions are designed for them (Clark & Schober, 1991, refer to this as the *presumption of interpretability*). In this experiment, we informed half of those respondents who could obtain clarification that the ordinary words in the questions could be used in ways they might not expect and that obtaining clarification would be essential if they were to understand as intended (the 'clarification essential' group):

> It may be that if you don't obtain definitions in this way, you won't be able to get the right answer, because you may be thinking about the question differently than the people who wrote it.

The other respondents who could obtain clarification were told simply that definitions were available (the 'clarification available' group), but the instructions did not stress the relation between definitions and response accuracy (see Appendix A for the full set of instructions). Our hypothesis was that respondents in the clarification essential group would be more likely than respondents in the clarification available group to suspect that they misunderstand a particular term or phrase. By recognising this 'conceptual misalignment', they should be more willing to ground their understanding by requesting a definition.

We also varied the way the survey system provided clarification. When clarification was user-initiated, respondents could request the official definition for a survey concept by clicking the mouse on highlighted text in the question. When clarification was mixed-initiative, the system could also offer a definition when respondents were 'slow' to respond. This was defined as taking longer than the median response time for complicated scenarios when no clarification was available. This offer ('Do you want help?') was presented in a dialog box, allowing users to reject it by clicking 'no' and to accept it by clicking 'yes'.

In addition to the four conditions created by crossing the instructions (clarification essential and clarification available) with the type of clarification (user-initiated and mixed-initiative), we included a fifth condition in which no clarification was available. This

DOI: 10.1002/acp

is much like a strictly standardised interview in which interviewers cannot engage in dialogue about the meaning of question wording.

Participants

Fifty-four (22 female and 32 male) respondents from the Washington, DC, area were recruited from an advertisement in the *Washington Post* and paid to participate. Thirteen participants were Black, 38 were White, and 3 were Asian. The educational backgrounds varied, with 24 having completed high school only, 21 with college degrees, and 9 with postgraduate education. Most (44) reported using a computer every day; 5 reported using a computer once a week, 2 once a month, and 3 once a year. These characteristics were roughly balanced across the experimental conditions. Eleven participants were recruited for each experimental condition except the user-initiated clarification, definitions available condition for which 10 participants were recruited.

Questions

Respondents answered 12 questions on the basis of fictional scenarios, half of which were straightforward and half complicated, presented in a counterbalanced order. The questions were taken from ongoing US government surveys, and had been used in earlier studies of human–human survey interviews (Schober & Conrad, 1997; Schober, Conrad, & Fricker, 2004). Four questions were about employment, from the Current Population Survey (e.g. 'Last week, did you do any work for pay?'); four questions were about housing, from the Consumer Price Index Housing survey (e.g. 'How many people live in this house?'); four questions were about purchases, from the Current Point of Purchase Survey (e.g. 'During the past year, have you purchased or had expenses for household furniture?'). The order of question domain (employment, housing and purchases) was counterbalanced across users. The order of questions within a domain was fixed and followed the order the questions appeared in the surveys from which they were taken. The questions appear in Appendix B.

Results

Clarification and response accuracy

Respondents who could obtain clarification (user-initiated and mixed initiative) provided substantially more accurate answers (based on how the official definition for the key concept in each question corresponded to the scenario) than did those unable to obtain clarification, $F(2,51) = 3.56$, $p < .05$. In particular, respondents provided reliably more accurate answers when they could click for clarification (user-initiated) than when no clarification was available, $F(1,51) = 5.76$, $p < .05$, and equally accurate answers when clarification was user-initiated and mixed initiative, $F(1,51) < 1$, n.s.

As one would expect, across all groups, respondents provided reliably less accurate answers for complicated than straightforward scenarios, $F(1,51) = 93.40$, $p < 0.001$. But their answers for complicated scenarios were reliably more accurate when they could receive clarification. In contrast, accuracy was uniformly high for straightforward scenarios whether clarification was available or not. Thus clarification (none versus any) interacted with scenario type (complication versus straightforward), $F(2,51) = 3.26$, $p < 0.05$ (Table 1). These results are consistent with those in Schober, Conrad and Fricker's (2004) telephone interviews using the same survey questions and scenarios, where respondents who could

DOI: 10.1002/acp

Table 1. Per cent correct responses, Experiment 1

	Scenarios	
Clarification	Straightforward	Complicated
None	98.5	40.9
User-initiated	98.4	67.5
Mixed initiative	100	66.4

Table 2. Per cent correct responses, complicated scenarios, Experiment 1

	Definitions	
	Available	Essential
User-initiated	55.0	78.8
Mixed-initiative	49.4	83.3

receive clarification answered more accurately for complicated scenarios. This suggests that data quality in textual web surveys can be improved by implementing strategies that work in human interviews (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober et al., 2004).

Looking just at complicated scenarios, response accuracy was enormously affected by the instructions to the respondent about the need for clarification. As Table 2 shows, when respondents had been told that definitions were essential, they answered more accurately (81.1%) than when they had been told that definitions were merely available (52.1%), $F(1,39) = 11.66$, $p < 0.005$. The way clarification was provided did not affect response accuracy, with 67.5% accuracy for respondents in the user-initiated clarification group and 66.4% for respondents in the mixed-initiative clarification group, $F(1,39) < 1$, *n.s.* The effect of instructions was not different when, in addition to honoring respondent requests for clarification the system could volunteer it (mixed initiative) than when respondents could only request clarification (user-initiated), so instructions (clarification essential versus available) did not interact with initiative (user–initiated versus mixed initiative), $F(1,39) = 0.36$, *n.s.* Note that the opportunity to receive clarification by itself does not necessarily lead to improved accuracy; respondents told that definitions were merely available did not answer reliably more accurately (52.1%) than respondents for whom clarification was unavailable (40.9%), $F(1,51) = 1.33$, *n.s.* As shown below this is because respondents in the clarification available condition did not request clarification very often.

Can we be sure that greater response accuracy results from getting clarification? The pattern is clear. When respondents did not get clarification, 34.8% of their answers for complicated scenarios were accurate. When respondents requested clarification, 90.1% of their answers for complicated scenarios were accurate. And when respondents received unsolicited clarification, 87.5% of their answers to complicated scenarios were accurate.[2]

[2]The per cents reported here were computed by first calculating an accuracy score (per cent of complicated scenario questions answered correctly) for each respondent for whom a score was possible, and then averaging across all the respondents. So the percentages are based on the 40 (of 54 possible) respondents who answered at least one question with a complicated scenario without getting clarification, the 31 (of 43 possible) respondents who requested clarification at least once, and the 8 (of 22 possible) respondents who received unsolicited clarification at least once. The ideal analysis is not possible because some respondents never received clarification and others always did, making the data too sparse for within-subjects analysis; there were only a small number of respondents who had data in all the clarification categories (no clarification, user-initiated and system-initiated).

DOI: 10.1002/acp

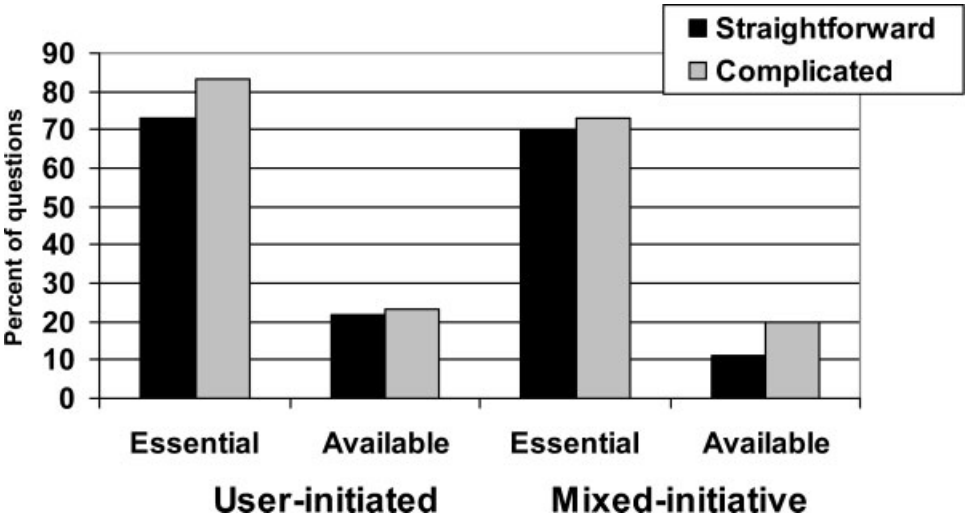

Figure 2. How often respondents requested (clicked for) clarification, Experiment 1

Note that unsolicited clarification was rare—only 8 (of 22 possible) respondents ever received it. This is because respondents who were told that clarification was essential tended to request clarification quickly, before the triggering threshold was exceeded (the system initiated 12% of all clarification for straightforward scenarios and 13% for complicated scenarios), and respondents who had been told that clarification was merely available tended to answer questions before the threshold had been exceeded (the system initiated 4% of all clarification for straightforward scenarios and 13% for complicated scenarios). So the pattern of data for the mixed initiative group does not differ much from the pattern for the user-initiated group; we follow up on this in Experiment 2.

Figure 2 shows how often respondents received clarification. As indicated above, this was mostly the result of respondent requests. The feature of the figure to note is that when respondents clicked, they did so indiscriminately. Respondents who were told that definitions were essential (the most accurate users overall) requested clarification not only for complicated scenarios, when clarification was helpful, but just as often for straightforward scenarios, when presumably it was not necessary. Similarly, respondents who were told definitions were available also did not distinguish between straightforward and complicated scenarios, even though with their infrequent requests for clarification one might expect them to have been more discriminating. But this doesn't mean that respondents were entirely indiscriminate; as discussed in the next section, respondents spent less time on definitions for straightforward than complicated scenarios.

Response time

Not surprisingly, obtaining and reading definitions took time. Respondents in the user-initiated clarification group averaged 43.7 s to answer each question, compared to 25.4 s per question for respondents in the no-clarification group, $F(1,51) = 4.37$, $p < 0.05$. How clarification was obtained did not reliably affect response times; respondents in the mixed initiative clarification group averaged 37.1 s per question, not reliably different from the 43.7 s per question in the user-initiated clarification group, $F(1,51) = 0.86$, *n.s.*

Respondents in all groups took longer to answer questions on the basis of complicated (39.1 s) than straightforward scenarios (31.7 s), $F(1,51) = 6.76$, $p = .012$, even though they

DOI: 10.1002/acp

clicked for definitions just as often in both cases. Thus, obtaining definitions is not always equally costly; respondents seemed to click and then quickly respond when they realised the definition wasn't needed, reducing the cost of unnecessarily clicking for definitions.

Focusing only on complicated scenarios, respondents took much longer to answer questions when they had been told that definitions were essential (54.4 s) than when they had been told that definitions were merely available (34.3 s), $F(1,39) = 4.63$, $p < 0.05$. This is consistent with the response accuracy results; respondents clicked for clarification and spent more time answering questions when they believed obtaining those definitions was essential. There was no difference in how long respondents took based on which clarification group (user-initiated or mixed initiative) they were in, $F(1,39) = 0.25$, *n.s.*, nor did instructions interact with clarification group, $F(1,39) = 0.06$, *n.s.*

These response time findings parallel those for human interviews. Clarification in survey dialogue takes time, but leads to greater response accuracy for complicated scenarios (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober et al., 2004).

Respondent satisfaction

The respondent satisfaction data suggest that, for most respondents, being able to interact with the system in order to get clarification improves the on-line survey experience. After answering the survey questions, most respondents reported that they would prefer ('yes' or 'no') participating in surveys in the future with a computer than with a human interviewer (between 82 and 90% across the five experimental groups) or with paper and pencil (between 80 and 91% across the groups). Respondents were asked to explain (by typing in open text) their preferences; for those who preferred a computer the explanations included the ability to get clarification, the ease and accuracy of data entry (writing is slower than clicking), speed of question presentation, anonymity, not feeling judged, lack of bias in how questions are presented, not wasting paper, and ease of comprehending questions. The respondents who preferred humans or paper to the computer reported that they missed the human interaction of an interview, that computers might be unfamiliar to some respondents or cause them discomfort, and that they felt an interviewer can provide superior clarification to that of a computer.

Most (8 of 11) respondents who could not get clarification reported that they would have asked for clarification if it had been available, and they collectively listed all but one question as requiring additional clarification. This suggests that interacting with dialog survey systems that don't allow clarification may be relatively frustrating.

Discussion

Our findings demonstrate that bringing features of dialogue—ability to clarify what questions mean—to a text-based web survey system can improve respondents' comprehension (and thus their response accuracy), without harming, and possibly increasing, user satisfaction. This comes at the cost of increased task duration, which could lower survey completion rates in actual web surveys, conducted outside the laboratory. But the increased task duration seems to be allocated in proportion to where it is needed; the greatest increase in response time comes for answers about complicated scenarios. So despite clicking indiscriminately for clarification, respondents seem to discriminate between definitions that are useful and those that are not.

Why did respondents click for clarification so indiscriminately? Most likely requesting clarification by clicking is easier or perhaps less inhibiting than asking an interviewer for

DOI: 10.1002/acp

clarification, as the satisfaction results suggest. Respondents who were told definitions were essential in the current experiment requested substantially more of them than respondents in conversational telephone interviews who received similar instructions and answered the same questions based on the same scenarios (83% for complicated scenarios in the current study versus 47% in Schober et al., 2004, Experiment 1). However, it seems that the ability to click for clarification is not, by itself, sufficient to promote frequent requests; respondents must also recognise the value of clarification. Respondents in the current study who were told merely that definitions were available requested substantially fewer of them (23%) than their counterparts in the interview study.

Overall, these results suggest that the success of dialogue-like clarification in textual web surveys depends both on how important respondents believe it is to understand accurately—their grounding criterion—and also on whether users recognise that system concepts may differ from their own. As is evident from the difference in clicking rates for respondents in the definitions-available and definitions-essential groups, respondents may not spontaneously recognise that their interpretations of ordinary terms like 'bedroom' and 'job' might differ from that of the system. This makes it harder for a system to diagnose respondents' need for clarification based on response times, because confident respondents may answer quickly yet inaccurately.

Experiment 1 implemented two features of human dialogue in a textual web survey: mixed-initiative clarification and variable grounding criteria. But other features of dialogue are present in interviews. Interviewers can use respondents' characteristics and speech cues to diagnose when a particular respondent needs clarification, reason about which parts of definitions would be appropriate to present at any given moment, what particular respondents are likely to misunderstand, etc. Interactive dialogue systems in other domains (tutoring, restaurant advice, travel planning) have attempted to implement these kinds of features of human dialogue (see, e.g. Moore, 1995, among many others). In Experiment 2, we focus on one of these features: diagnosis of need for clarification based on respondents' characteristics.

EXPERIMENT 2: DIAGNOSING RESPONDENTS' NEED FOR CLARIFICATION

In conversation, speakers can often tell when their addressees do not understand them. Skilled interviewers bring this ability to bear in surveys, particularly when they are empowered to provide clarification. What cues do they use to diagnose the respondent's need for clarification? Interviewers seem to rely not only on respondents' explicit requests for clarification, but also on paralinguistic and visual cues provided (intentionally or not) by respondents (Conrad et al., in press; Schober & Bloom, 2004). They also can adjust their diagnosis based on what they know about the respondent's characteristics: their gender, ethnicity, employment status, education, age, etc.

In Experiment 2, we implement 'user models' (Kay, 1995) based on respondents' age. Survey methods research has shown that age affects responding, largely because working memory declines (Knäuper, 1999). More germane to our application, the cognitive aging literature documents a general slowing of behaviour with age (Salthouse, 1976), so that older web survey respondents will answer more slowly than younger ones. If this is the case, the same period of inactivity by old and young respondents may have different

DOI: 10.1002/acp

meanings: a short lag may indicate confusion or processing difficulty for a young respondent but simply ordinary thinking for an older respondent.

We contrasted five user interfaces. In the first there was no clarification available to respondents. The second was user-initiated, where clarification was available to the respondent by clicking. The third embodied a 'generic user model', where the respondent could request clarification but the system provided clarification if the respondent's inactivity exceeded a fixed threshold. The fourth was built around 'group-based user models', identical in approach to generic user models except that the inactivity threshold depended on the respondent's age group. Note that the third and fourth interfaces both embody mixed initiative clarification. In the fifth interface, the definition always appeared along with the survey question.

Respondents completed a brief on-line tutorial prior to the start of the experimental session to learn how to click for definitions. In all but the noclarification condition, the tutorial recognised them to obtain a definition but were not given any detail about how or when clarification might be helpful, just as in the definitions available instructions in Experiment 1. When the system volunteered clarification in the actual experiment, the definition was displayed and a tone sounded simultaneously to alert respondents to the arrival of the definition. This contrasts with Experiment 1 in which the system gave the respondents the opportunity to reject clarification offered by the system.

Questions

All respondents answered 10 questions (used by Conrad & Schober, 2000) on the basis of fictional scenarios (see Appendix C). Five purchase questions were drawn from the Current Point of Purchase Survey and five housing questions from the Consumer Price Index Housing survey. Half of the respondents answered the housing questions first, and the other half answered the purchase questions first. As in Experiment 1, half of the scenarios presented to each respondent were complicated and half were straightforward, in counterbalanced orders.

Thresholds

To establish the inactivity thresholds, we examined response times for the first 20 respondents in the no-clarification condition as well as the response time for the 12 respondents in the user-initiated condition who did not request clarification. Across the questions, response times for straightforward and complicated items were most different at the 40th percentile, so we used this time as the inactivity threshold in the generic user model. The group-based user models were also based on the 40th percentile response time for complicated mappings but computed separately for old (slower) and young (faster) respondents.

Participants

One hundred fourteen paid participants were recruited from the New York City area through an advertisement in the *Village Voice* and fliers posted at senior centres. There were 56 females and 58 males. Half of the participants were young (defined here as less than 35 years old) with a mean age of 26.8, and half were old (defined as over 65 years old) with a mean age of 72.4. Ethnicities, educational backgrounds and experience with computers

DOI: 10.1002/acp

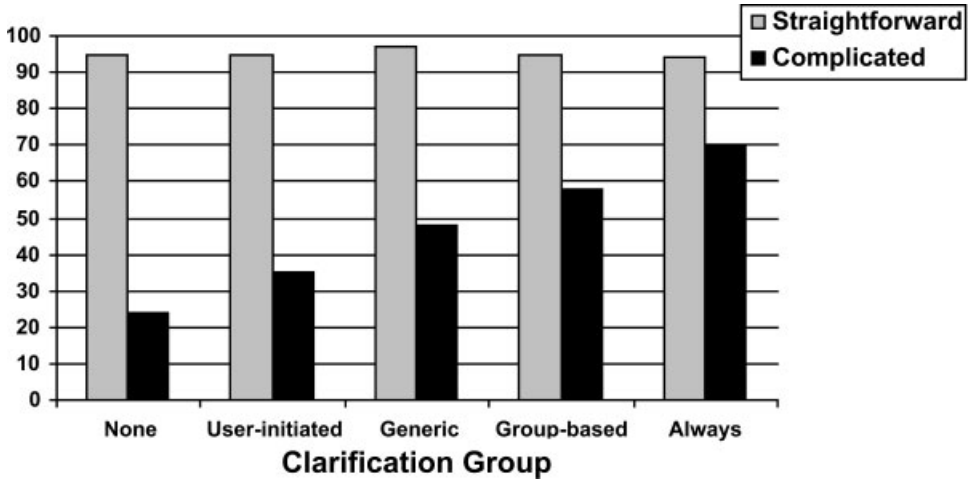

Figure 3. Response accuracy, Experiment 2

were balanced across age groups. There were 10 old and 10 young respondents in each group that received clarification, with 17 old and 17 young respondents in the no clarification group, for modelling purposes. Participants' ages ranged from 20 to 82 in the No Clarification condition, 18 to 86 in the User-initiated condition, 21 to 83 in the Generic Model condition, 21 to 83 in the Group-based Model condition, and 20 to 86 in the Definitions Always condition.

Results

Clarification and accuracy

As can be seen in Figure 3, all respondents were quite accurate when answering on the basis of straightforward scenarios (95% of questions answered correctly), but for complicated scenarios, accuracy varied depending on clarification group as revealed by the interaction of clarification group and scenario type $F(4,\ 104) = 16.58$, $p < 0.001$. Looking just at complicated scenarios, accuracy increased linearly across the five groups, $F(1,104) = 8.16$, $p < 0.001$. When respondents could not obtain clarification at all for complicated mappings, accuracy was quite poor (24% of questions answered correctly). When the system didn't provide clarification, but respondents could obtain definitions by clicking (user-initiated), accuracy was better (35% of questions answered correctly). Accuracy was better still when the system also volunteered clarification on the basis of a generic model (48%). Presumably this increment in accuracy reflects the benefits of additional system-initiated clarification on occasions on which respondents did not realise their interpretation differed from the designers' and did not request the clarification they needed. Accuracy was better still when the system took respondents' age into account (58% correct for group based participants) than when the thresholds were generic. Accuracy was best of all when respondents received definitions along with the questions (70% of questions answered correctly).

Although group-based user modelling boosted accuracy above generic user-modelling for complicated scenarios, these benefits interacted with age $F(4,104) = 3.22$, $p = 0.016$. Older respondents performed equally well with group-based (46% correct) and generic (50% correct) modelling, $F(1,52) < 1$, *n.s.* In contrast, the younger respondents performed

DOI: 10.1002/acp

Table 3. Per cent of questions for which clarification was requested by respondent, initiated by system and total clarification, complicated scenarios, Experiment 2

	Requested by R		System-initiated		Total	
Age group	Young	Old	Young	Old	Young	Old
Clarification						
None	0	0	0	0	0	0
User-initiated	34	6	0	0	34	6
Generic	24	16	48	62	72	78
Group-based	48	2	46	66	94	68
Always	0	0	100	100	100	100

better with group-based (70% correct) than with generic modelling (46% correct), $F(1,52) = 5.65$, $p < 0.05$. In fact, younger respondents performed as well with group-based user modelling as when they always received definitions (72% correct), $F(1,52) < 1$, *n.s.* This suggests that the group based model for young respondents accurately detected when they did and, perhaps more importantly, when they did not need clarification. Because accuracy was no better when clarification was always presented, the extra clarification in this condition must have been largely unnecessary. Older respondents, in contrast, showed the largest increase in accuracy between user-initiated (22% correct) and generic model (50% correct) clarification, $F(1,52) = 8.87$, $p = 0.005$ but no advantage from group-based over generic clarification. The fact that they derived no extra benefit from group-based clarification could suggest that the threshold was not set appropriately for these respondents, a possibility we return to later.

Accuracy was related to the frequency with which clarification was provided. The younger respondents, as just indicated, were more accurate in the group-based than generic clarification condition, and they received clarification for more complicated scenarios in the group-based (94%) than in the generic (72%) conditions (see Table 3). In contrast, the older respondents were slightly (but not significantly) less accurate in the group-based than generic conditions and, accordingly, received slightly less clarification for complicated scenarios in the grouped-based (68%) than generic (78%) conditions, interaction of age group $\times$ clarification group $F(1,36) = 4.82$, $p = 0.035$. If we compare all responses for complicated scenarios for which any clarification was provided (i.e. requested by the respondent, provided by the system, or always displayed) to all responses for which no clarification was provided, response accuracy was substantially higher with clarification (66%) than without (28%), F (1, 568) $= 132.29$, $p < 0.001$.

Although frequency of clarification is related to accuracy, above a certain amount more clarification was not particularly helpful. Increasing the frequency of clarification to 100% (the clarification always condition) for younger respondents did not increase their accuracy relative to the 94% frequency with which it was provided under group-based user modelling (accuracy was 72% and 70% in the clarification always and group-based user modelling conditions, respectively). This suggests that the threshold for system-initiated clarification in the group-based condition was probably diagnostic of these respondents' need for clarification and the extra clarification in the always condition was apparently provided when it was not needed.

The generally low rates of user-initiated clarification (Table 3) indicate that respondents were not good at recognising when they needed clarification. This is especially true for older respondents, who initiated clarification far less often (8% of the time on average

DOI: 10.1002/acp

across the three groups that allowed respondents to request clarification) than younger respondents did (35.3%), $F(1,54) = 14.45$, $p < 0.001$. Thus, system-initiated clarification seems to helpfully supplement the rare respondent requests for clarification. As Table 3 shows, in the generic and group-based clarification groups a large per cent of the clarification came at the system's initiative rather than the respondent's request. One possibility alluded to earlier is that for older respondents the group-based model, that is, the inactivity threshold, was not set optimally. Because the total number of questions for which older respondents were given clarification (68%) is smaller than the comparable figure for young respondents (94%) it could be that the system should have been programmed to offer clarification sooner. On the other hand, because older respondents requested clarification so infrequently (for only 2% of the questions in the group-based condition) it is possible that they didn't adequately recognise they might misunderstand and that misunderstanding might have led them to answer incorrectly. It could be that different instructions could help or earlier—and therefore more—system-initiated clarification might sensitise them to the possibility of interpreting ordinary words differently than intended. Another piece of the puzzle might be that older respondents find clicking to be more effortful than do younger respondents and so other interfaces might be helpful (see Conrad et al., 2006 for a discussion of the effort associated with clicking for clarification).

Response time

As in Experiment 1, when respondents were given clarification, they answered after a longer amount of time than when they were not given any clarification, suggesting that they read and considered the definitions. In particular, in the no clarification condition the average response time was 31.2 s, but in the other conditions (where clarification was available upon respondents' request, the system's initiative or always displayed) the average response time was 38.6 s ($F[1,110] = 5.14$, $p < 0.05$). If we directly compare response times when clarification was provided (irrespective of the exact method of presentation) to response times when no clarification was provided there is a substantial difference: mean response time with clarification was 46.3 s and mean response time without clarification was 27.7 s, $F\ (1,\ 1138) = 47.30$, $p < 0.001$.

This slowing of responses when clarification is provided is evident for both young and old respondents. Looking just at those conditions where clarification was available, respondents in both age groups were fastest (and least accurate) when the only clarification they received was user-initiated (18.6 s for younger respondents and 32.5 s for older respondents). In contrast, respondents in both age groups were slowest (and most accurate) when definitions were displayed all the time (37.2 s for younger respondents and 58.2 s for older respondents) as shown by a simple contrast between user-initiated and always conditions $F(1,\ 104) = 4.82$, $p < 0.001$.

Note that user modelling seems to have affected response times in a sensible way. For complicated scenarios, young respondents were faster (though not reliably) in the group-based (29.1 s) than generic (38.0 s) condition, presumably because the system-initiated clarification in the group-based condition was delivered earlier than in the generic condition; in contrast older respondents were slower in the group-based (61.4 s) than generic condition (41.1 s) ($F[1,52] = 6.36$, $p < 0.05$), presumably because the system-initiated clarification in the group-based condition was delivered later than in the generic condition. We also see that across all clarification groups, older respondents took substantially longer (48.7 s) than younger respondents did (30.7 s), $F(1,\ 104) = 18.51$,

DOI: 10.1002/acp

$p < 0.001$, which is consistent with much work in the cognitive aging literature (see, e.g. the classic work of Salthouse, 1982).

Respondent satisfaction

Respondents were moderately satisfied overall, with a mean evaluation of 2.76, where 4 is 'very good' and 0 is 'very bad'. They were particularly satisfied with respondent-initiated clarification (3.40), reliably more than with no clarification (2.39), $F(1,100) = 412.70$, $p < 0.001$. Their ratings for the other conditions (2.89 for the generic model, 2.50 for the group-based model, and 2.63 for clarification always) were not reliably higher than for the no-clarification control. So it appears that the respondents wanted the ability to obtain clarification by requesting it, but they did not welcome system-initiated clarification, presumably because it was unsolicited. The pattern of satisfaction was the same for young and old respondents, interaction of age and clarification condition $F(4,100) < 1$, *n.s.* Note that the preference for respondent-initiated clarification was not related to accuracy. As we already reported, respondents were more accurate under mixed initiative clarification (generic and group-based conditions) than when clarification was entirely user-initiated.

When asked whether they would prefer interviewer- or computer-administration of similar questionnaires in the future, respondents were about evenly split in the no clarification condition (47% preferring an interviewer, 53% preferring a computer overall), presumably because an interviewer might be able to provide the clarification they were unable to obtain in the current experiment. However, in the conditions where clarification was available, respondents indicated they would prefer a computer to an interviewer in future questionnaires (74% preferring a computer), presumably because they were satisfied with computer-based clarification in the current experiment. The difference in preferences across the experimental groups was marginally significant ($\chi^2[4] = 9.07$, $p = 0.06$). The pattern differed somewhat for young and old respondents (see Table 4). In the group-based clarification condition old respondents exhibited a marked preference for an interviewer (30% preferred a computer), while the opposite was true for young respondents in this condition (90% registered a preference for a computer). The pattern was quite different for the definitions always condition where 90% of the older respondent indicated a preference for a computer while only 44% of young respondents indicated a similar preference, $\chi^2(4) = 11.30$, $p < 0.05$ for the pattern of preferences across the 10 age × clarification groups. One explanation for the older, group-based respondents' preference for an interviewer is that the system provided definitions to older respondents in the group-based condition after too long a delay; because they frequently did not request clarification, they may have been in a confused state for an uncomfortably long amount of time before the system eventually offered them clarification. One reason that younger respondents in the

Table 4. Proportion of respondents preferring a computer to an interviewer, Experiment 2

	Age Group		
	Young	Old	Overall
Clarification			
None	0.82	0.20	0.53
User-initiated	1.00	0.80	0.90
Generic	0.78	0.78	0.78
Group-based	0.90	0.30	0.60
Always	0.44	0.90	0.68

DOI: 10.1002/acp

definitions always condition so strongly preferred an interviewer might be that often the definitions were not necessary or helpful, and younger respondents may have assumed that an interviewer would be less likely to provide clarification that is not needed.

Discussion

Implementing another feature of dialogue—modelling one's partner based on their characteristics—in a text-based web survey improved overall data quality. More than in Experiment 1, system-initiated clarification improved response accuracy. These data demonstrate that two-way initiation of clarification in web surveys, as in interviews (Schober et al., 2004), can lead to greater accuracy than relying on respondents to request clarification only when they think they need it.

But the modelling worked differently for older and younger respondents. Group-based models improved response accuracy for younger respondents more than for older respondents, and older respondents found system-initiated clarification particularly irritating. This may be an artifact of the particular thresholds that we selected in this experiment; future studies could disentangle what range of thresholds, or what other characteristics of respondents, might be modelable. For example, in addition to overall speed of responding, one could imagine modelling other age-related characteristics of respondents like working memory capacity (Knäuper, 1999), or other characteristics that are not necessarily related to age, like computer experience or education. Finally, instead of group-based characteristics, it might be even more effective to construct individual user models. Individual respondents' uncertainty could be assessed with inactivity, just like the groups in the current study. However, individual thresholds would need to be set on the basis of earlier behaviour in a web session, for example, response times on a small number of questions requiring clarification.

GENERAL DISCUSSION AND CONCLUSIONS

Overall, these data suggest that bringing features of human dialogue to web surveys is a promising approach that opens up new possibilities for increasing survey data quality. In particular, presenting clarification for words in questions can improve the accuracy of answers when respondents are answering about complicated situations. Also, implementing two-way initiation of clarification can allow web survey respondents to answer more accurately than leaving the initiation of clarification only up to respondents. The need for designing interviewing systems that supplement respondents' initiation of clarification is evident from our Experiment 1 finding that respondents only clicked for clarification frequently when they were experimentally instructed to do so. This is particularly striking given that participants in our experiments had multiple incentives to accurately understand the survey questions; they were paid to participate and so may have felt compelled to perform well, and an experimenter was present whom respondents could have believed knew whether their answers were accurate.

The results also suggest that tailored diagnosis of the respondent's need for clarification via user-modelling (here at the group level, but also potentially at the individual level) is another promising area to explore for web surveys. As with mixed-initiative clarification more generally, adapting this aspect of human–human collaboration to the web survey setting may well allow self-administration to embody some of what is most successful in

DOI: 10.1002/acp

interviewer-administered data collection. But note that this requires rethinking some of the basic assumptions about meaning from the strictest views of standardisation theory (Fowler & Mangione, 1990): survey designers will need to be willing to allow only some respondents to be given definitions for survey terms. Particularly in web surveys this can be done in a precisely determined way that reduces fears about the interviewer bias that can be introduced from allowing interviewers to intervene as they see fit, and ultimately it may allow for the standardisation of meaning (rather than wording) that Suchman and Jordan (1990) have promoted.

Obviously, these studies are laboratory-based and exploratory, rather than large-scale field tests of refined techniques and technologies. Note also that the generalisability of these findings depends on how frequently respondents in real-world web surveys have complicated situations that are akin to our complicated scenarios. If such situations are as frequent as our telephone interview data suggest (Conrad & Schober, 2000)—or if their frequency is unknown—it may well be worth designing web surveys with clarification capability. While we strongly believe that bringing features of dialogue to self-administered surveys is likely to be fruitful, we see these findings as raising questions for further exploration rather than providing immediate prescriptions for web survey designers. A few immediate practical questions come to mind:

Will real-world respondents read and use definitions in a web survey?
Little is yet known about whether in an actual web survey the extra time required for clarification would decrease respondent satisfaction and increase break-offs. Web respondents 'in the wild' may be even less concerned than laboratory participants with grasping what they consider to be the nuance of a survey question and so may be reluctant to request clarification. Conrad, Couper, Tourangeau and Peytchev (2006) observed that only between 14% and 22% of actual web respondents (recruited from an opt-in panel) clicked for definitions, but that features of the interface and the definitions could dramatically increase rates of clicking.

Why not provide clarification all the time?
In Experiment 2, response accuracy was greatest when definitions were displayed along with each survey question. Yet respondents were not happy about it; satisfaction was generally low for the always condition, and young respondents showed a preference for a human interviewer to a computer that presented definitions all the time. In addition, response times were greatest for this group, presumably because respondents read definitions which they did not need. These serious drawbacks to providing clarification all the time suggest that more tailored clarification is worth exploring further; clarification tailored to individual respondents could help them interpret questions as intended when they really need clarification but not when they don't. Alternatively, one might investigate other features of dialogue to import into web surveys or other ways of heightening respondents' awareness that they need clarification, like including parts of definitions along with original questions to sensitise respondents to definitional complexities (Lind, Schober, & Conrad, 2001).

What about respondents who are certain that their erroneous interpretation is correct?
In Experiment 1, respondents who were told only that clarification was available often entered incorrect answers before the system offered clarification. They seemed to lack

DOI: 10.1002/acp

scepticism about what they had taken the question to mean, thus displaying little evidence of uncertainty. But very quick answers may indicate that these respondents had not recognised the potential ambiguity in the question. One approach is to provide feedback to users about the brevity of their response latency and offer them a definition and chance to answer again. We have explored this approach in the context of speech interfaces to automated questionnaires (Ehlen, Schober, & Conrad, 2005) and found it to improve accuracy beyond the sort of mixed initiative used in the current study. It would be a straightforward matter, conceptually, to add a lower bound to the inactivity threshold for a mixed-initiative, desktop interface to a web questionnaire.

Beyond the practical (though preliminary) value of these findings for web survey design, we believe that these findings require us to refine a theory of human–human collaboration by explicitly introducing the notion of initiative for grounding understanding (following Walker and Whittaker's (1990) proposals about conversational initiative more generally). Our findings that comprehension success can vary depending on whether the user or system takes the initiative should be extended to the human realm; a collaborative theory should include who takes the responsibility for clarifying meaning. In many cases, speakers are responsible for what they mean, and listeners assume that what speakers say is readily interpretable to them in the current context (Clark and Schober's (1991) interpretability presumption). But in situations where the speaker is less competent than the addressee, the addressee may take responsibility for the meaning, and may initiate clarification (Schober, 1998, 2005; Schober & Brennan, 2003). Who should be responsible under what circumstances, and what determines how speakers decide whose effort should be minimised, are important questions for a theory of collaboration.

Because the web is such a rich and flexible medium, web surveys of the future will not be monolithic. Here we have enhanced textual web surveys in a few preliminary ways that suggest some of the possibilities. But a survey administered via web could include speech, video, embodied conversational agents, or other as-yet-unknown features; web surveys might be self-administered, human-administered (e.g via live chat) or a hybrid; web surveys could have complex artificial intelligence governing dialogue moves. Surveys via the web may not only provide data comparable to human interviews but even lead to improved data; a growing body of evidence suggests that in some circumstances—answering questions about sensitive behaviours (Tourangeau & Smith, 1996) or getting correction from a tutor (Schofield, 1995)—interacting with a machine might be preferable to interacting with a person. As the current study suggests, features of human dialogue capability can enhance text-based web surveys, but what works for some respondents may not work for others. We will need to develop clearer distinctions about when particular aspects of human interviews are desirable to implement in web surveys and when they are not.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants No. IIS-0081550 and SES-0551294. In addition, the Bureau of Labor Statistics provided support. We thank Susan Brennan, Cathy Dippo, Patrick Ehlen, Scott Fricker, Susan Schnipke and Clyde Tucker. The opinions expressed here are those of the authors and not those of the Bureau of Labor Statistics.

DOI: 10.1002/acp

REFERENCES

Akmajian, A., Demers, R. A., Farmer, A. K., & Harnish, R. M. (1990). *Linguistics: An introduction to language and communication* (3rd ed.). Cambridge, MA: MIT.

Bassili, J., & Scott, S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*, 390–399.

Christian, L., & Dillman, D. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly, 68*, 57–80.

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

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: APA.

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

Clark, H. H., & Schober, M. F. (1991). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 15–48). New York: Russell Sage Foundation.

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

Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2006). Use and non-use of clarification features in web surveys. *Journal of Official Statistics, 22*, 245–270.

Conrad, F. G., & Rips, L. J. (1986). Conceptual combination and the given/new distinction. *Journal of Memory and Language, 25*, 255–278.

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

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

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review, 18*, 384–396.

Dillman, D. A. (1999). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: Wiley.

Ehlen, P., Schober, M. F., & Conrad, F. G. (2005). Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Proceedings of the Symposium on Dialogue Modeling and Generation, 15th Annual meeting of the Society for Text & Discourse,* Vrije Universiteit, Amsterdam, 2005.

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

Kay, J. (1995). Vive la difference! Individualized interaction with users. In C. S. Mellish (Ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 978–984). San Mateo, CA: Morgan Kaufmann.

Knäuper, B. (1999). Age differences in question and response order effects. In N. Schwarz, D. Park, B. Knäuper, & S. Sudman (Eds.), *Cognition, aging, and self-reports*. Philadelphia: Taylor & Francis.

Krauss, R. M., & Fussell, S. R. (1996). Social psychological models of interpersonal communication. In E. T. Higgins, & A. Kruglanski (Eds.), *Social psychology: A handbook of basic principles* (pp. 655–701). New York: Guilford.

Moore, J. D. (1995). *Participating in explanatory dialogues: Interpreting and responding to questions in context*. Cambridge, MA: MIT.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169–190.

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

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

Schober, M. F. (1998). Different kinds of conversational perspective-taking. In S. R. Fussell, & R. J. Kreuz (Eds.), *Social and cognitive psychological approaches to interpersonal communication* (pp. 145–174). Mahwah, NJ: Lawrence Erlbaum.

DOI: 10.1002/acp

Schober, M. F. (2005). Spatial dialogue between partners with mismatched abilities. *Proceedings of Workshop on Spatial Language and Dialogue*, Hanse Wissenschaftskolleg, Delmenhorst, Germany.
Schober, M. F., & Bloom, J. E. (2004). Discourse cues that respondents have misunderstood survey questions. *Discourse Processes, 38*, 287–308.
Schober, M. F., & Brennan, S. E. (2003). Processes of interactive spoken discourse: The role of the partner. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 123–164). Mahwah, NJ: Lawrence Erlbaum Associates.
Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*, 211–232.
Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61*, 576–602.
Schober, M. F., & Conrad, F. G. (2002). A collaborative view of standardized survey interviews. In D. Maynard, H. Houtkoop, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 67–94). New York: Wiley.
Schober, M. F., Conrad, F. G., & Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology, 18*, 169–188.
Schober, M., Conrad, F., Ehlen, P., & Fricker, S. (2003). How web surveys differ from other kinds of user interfaces. *Proceedings of the American Statistical Association, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.
Schofield, J. W. (1995). *Computer and classroom culture*. Cambridge, UK: Cambridge University Press.
Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85, 232–253.*
Tourangeau, R., & Smith, T. (1996). Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*, 275–304.
Walker, M. A., & Whittaker, S. (1990). Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings 28th Annual Meeting of the Assocation for Computational Linguistics*, 70–78.

APPENDIX A: INSTRUCTIONS TO RESPONDENTS IN EXPERIMENT 1 CONDITIONS WHERE IT WAS POSSIBLE TO OBTAIN CLARIFICATION

Clarification essential

Sometimes these survey questions use ordinary words with slightly different meanings than you may be used to. This is because surveys sometimes need to have technical definitions different from ordinary definitions. This questionnaire program has been built so that at least some of these words or phrases with special meanings appear in blue. This means that if you position the mouse cursor over the blue word or phrase and click the mouse the computer will display a definition for that word or phrase. Please take advantage of this by clicking on the blue words if you have even the slightest doubt about what they mean.

It may be that if you don't obtain definitions in this way, you won't be able to get the right answer, because you may be thinking about the question differently than the people who wrote it. For example, imagine that your packet contains a shopping receipt that shows that Gina bought butter. If the corresponding question asked you 'Did Gina buy any fats or oils?', you might want to say 'yes' because butter seems to be a fat. But the official definition of 'fats or oils' excludes butter, and so the correct answer would be 'no'. If you didn't obtain the definition of butter, you probably would get the wrong answer.

DOI: 10.1002/acp

Clarification available

You may notice that some of the words or phrases in the questions appear in blue. This means that optional information about these words or phrases is available if you position the mouse cursor over the blue text and click the mouse. You are not required to do this; it is entirely up to you.

APPENDIX B: QUESTIONS IN EXPERIMENT 1 (KEY CONCEPTS ARE ITALICISED)

Housing questions (from CPI Housing survey):

1. How many *bedrooms* are there in this house?
2. This question has two parts. How many *full bathrooms* are there in this house? How many *half bathrooms* are there?
3. How many *other rooms* are there, other than bedrooms and bathrooms?
4. How many people *live in this house*?

Work questions (from CPS survey):

1. Does anyone in this household have a *business* or a farm?
2. Last week, did Chris do any work for *pay*?
3. Last week, did Pat have *more than one job*, including part-time, evening or weekend work?
4. How many hours per week does Mindy *usually* work at her job?

Purchase questions (from CPOPS survey):

1. Has Carla purchased or had expenses for *car tyres*?
2. Has Alexander purchased or had expenses for *college tuition or fixed fees*?
3. Has Kelly purchased or had expenses for *household furniture*?
4. Has Dana purchased or had expenses for *meats and poultry*?

APPENDIX C. QUESTIONS IN EXPERIMENT 2 (KEY CONCEPTS ARE ITALICISED)

Housing questions (from CPI Housing survey):

1. How many *bedrooms* are there in this house?
2. How many *full bathrooms* are there in this house?
3. How many *half bathrooms* are there?
4. How many *other rooms* are there, other than bedrooms and bathrooms?

DOI: 10.1002/acp

5. How many people *live* in this house?

Purchase questions (from CPOPS survey):

6. Has Chris purchased or had expenses for *moving*?
7. Has Alexander purchased or had expenses for *telephones or telephone accessories?*
8. Has Pat purchased or had expenses for *inside home maintenance or repair services?*
9. Has Kelly *had expenses for household furniture?*
10. Has Carla *purchased or had expenses for whiskey or other liquors for home use*?

DOI: 10.1002/acp