APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **18**: 169–188 (2004)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.955

Misunderstanding Standardized Language in Research Interviews

MICHAEL F. SCHOBER[1]*, FREDERICK G. CONRAD[2] and SCOTT S. FRICKER[3]

[1]*New School for Social Research, USA*
[2]*University of Michigan, USA*
[3]*Bureau of Labor Statistics, USA*

SUMMARY

Leaving the interpretation of words up to participants in standardized survey interviews, aptitude tests, and experiment instructions can lead to unintended interpretation; more collaborative interviewing methods can promote uniform understanding. In two laboratory studies (a factorial experiment and a more naturalistic investigation), respondents interpreted ordinary survey concepts like 'household furniture' and 'living in a house' quite differently than intended in strictly standardized interviews, when the interpretation was left entirely up to them. Comprehension was more accurate when interviewers responded to requests for clarification with non-standardized paraphrased definitions, and most accurate when interviewers also provided clarification whenever they suspected respondents needed it. Copyright © 2004 John Wiley & Sons, Ltd.

Most people probably assume that competent listeners or readers exposed to the same words or phrases will all end up with more or less the same interpretations. How else would people reach common understandings of, for example, newspaper articles, advertisements, or the items in a drop-down menu on a computer? Psychologists implicitly subscribe to this view when they assess attitudes and opinions in questionnaires, when they administer aptitude tests and clinical assessments, and when they instruct experimental participants. In order to reduce bias and promote comparability of responses, they standardize the materials: they expose all respondents and subjects to exactly the same words. They also require participants to interpret those words for themselves, so that researchers will not influence responses, as in these instructions for administering the Information subtest of the WAIS-III intelligence test:

> Read each question aloud exactly as it is written. If the response to a question is incomplete or unclear . . . it is permissible to say: 'Explain what you mean' or 'Tell me more about it'. However, do not ask leading questions or spell the words. Do not alter the wording of any question. (Wechsler, 1997, p. 136).

*Correspondence to: Michael F. Schober, Department of Psychology F330, New School University, 65 Fifth Avenue, New York, NY 10003, USA. E-mail: schober@newschool.edu

Contract/grant sponsor: National Science Foundation; contract/grant numbers: SBR-9730140 and ISS-0081550.
Contract/grant sponsor: Bureau of Labor Statistics.

If an examinee's response 'suggests he or she has misheard or misunderstood the exact meaning of the questions' the examiner should simply 'repeat the question' (p. 136), still leaving the interpretation up to the examinee.

Other domains in which practitioners advocate some version of standardized wording without clarification include survey research (e.g. Fowler & Mangione, 1990), personality testing (e.g. Butcher, 1972; Meehl, 1947), personnel selection (e.g. Huffcutt & Woehr, 1999; Kahn & Cannell, 1957), educational testing, physician training (e.g. Barrow, 1993; Pangaro et al., 1997), police lineups (e.g. Gwyer & Clifford, 1997; Smith, Kassin, & Ellsworth, 1989), police interviews (e.g. Fisher, Geiselman, & Raymond, 1987; Memon, Holley, Milne, Koehnken, & Bull, 1994), and user interface design (e.g. Pangalo, 1993; Tanaka, Eberts, & Salvendy, 1991), among others. Each of these fields has a history that has led it to adopt this approach (e.g. see Beatty, 1995; Wright, 1969); in each case standardization is seen as an important improvement over earlier less consistent methods.

But standardizing wording and leaving the interpretation up to the audience have not gone unchallenged. In survey research (e.g. Houtkoop-Steenstra, 2000; Suchman & Jordan, 1990), personnel interviews (Speak, 1967), educational testing (Marlaire & Maynard, 1990), medical assessment (e.g. Mishler, 1986), and user interface design (e.g. Grudin, 1989; Gulliksen & Sandblad, 1995), dissenters have argued that overreliance on standardized wording is problematic. The argument is that different listeners, readers or users can interpret the same words quite differently, and so standardized wording may actually decrease uniformity of interpretation. An alternative proposal has been for speakers (interviewers, testers, computer systems) to collaborate with the addressee (respondents, test-takers, users), saying whatever it takes to make sure that the addressee has understood the words as intended. Paradoxically, in order to standardize interpretation speakers may need to follow non-standardized 'conversational' practice, clarifying what their words mean when addressees ask for help or seem to need it, checking whether addressees have understood, helping addressees respond appropriately, etc. (see e.g. Schober & Conrad, 2002).

Here we examine when standardized language succeeds and when it does not, and what features of standardization matter the most. Standardized interactions actually differ from spontaneous ones in several fundamental ways; particular combinations of features lead to different forms of standardization. In the strictest standardization (see Fowler & Mangione, 1990), the wording is held constant for all addressees, rather than being tailored to a particular addressee. The wording is scripted rather than spontaneously produced, and the speaker is usually not the author of the words. In the strictest standardization, the speaker or author cannot (or is forbidden to) clarify what words mean, because the stimulus would not be standardized if only some addressees received clarification, or if clarification involved different words for different addressees. For example, if survey respondents ask what a question means, strictly standardized interviewers should not answer directly; rather, they should repeat the question, repeat the response alternatives, or tell respondents 'whatever it means to you' (Fowler & Mangione, 1990).

In less strict forms of standardization, wording *can* be clarified: a census taker might define 'home' for a respondent who needs clarification (and not for another respondent whose understanding seems adequate), a computer user can click on a help menu to find out what 'delete' will do (while another user might not), etc.[1] One possibility is for

[1]Note that under the strictest criteria for standardization, this would not qualify as standardized practice. Because not all participants will request clarification, not all participants receive exactly the same stimulus (see Fowler & Mangione, 1990).

standardized clarification to be presented only when the addressee requests it ('Um . . . how often do I need to stay in a place for it to count as my home?'). Another possibility is for the speaker to also provide clarification if the addressee seems to need it, even without being explicitly requested ('You seem confused—let me tell you what we mean by ''home'' '). Another variation is in how clarification is presented: in a scripted fashion (e.g. an interviewer reads an official definition for 'home' verbatim) or unscripted (the interviewer paraphrases the definition).

In the studies reported here we present evidence that under some circumstances strictly standardized interviewing leads to poor comprehension, and that less strictly standardized methods can lead to improved comprehension. We demonstrate this in the context of US federal surveys about facts and behaviours on which influential official statistics like the unemployment rate and the Consumer Price Index are based. Survey interviews like these are a useful domain in which to study language comprehension. The concepts in the questions are everyday notions like 'job', 'work for pay', and 'household furniture'. When the question concepts are well defined by survey designers, the extent to which respondents' answers are accurate—fit the official definitions—thus measures the degree to which respondents have understood the words in the questions as the survey designers intended.

In these sorts of interviews, misunderstandings can have real-world consequences for official statistics, leading to inaccurate estimates for politically sensitive figures like the unemployment and inflation rates. Despite the high stakes, there is little consensus about what counts as the best standardized procedure. Different survey organizations that consider themselves to implement standardization go about it in different ways. Some train their interviewers to adhere to the strictest standardization of wording, while others train interviewers to provide scripted clarification when respondents explicitly request it (Viterna & Maynard, 2002). As far as we know, no survey organization advocates allowing interviews to present *unscripted* clarification, as this might allow interviews to paraphrase inaccurately. But as we show in Study 2, even within a single organization interviewer behaviour can vary substantially, up to and including the presentation of unscripted clarification.

As we see it, strictly standardized wording, with no clarification, should lead to adequate comprehension only sometimes. For example, if a survey respondent who has purchased tyres for a Honda Accord (an ordinary sedan) is asked whether he purchased car tyres, both he and the survey designers are likely to conceive of his purchase as car tyres, and so he should not need extra clarification to answer correctly. But for a respondent who owns a pick-up truck it may be a different story. Both he and the survey designers may consider truck tyres to be examples of car tyres—or they may not. If they both happen to conceive of car tyres the same way, then clarification is unnecessary for accurate comprehension; but if one person's concept is different from the other's, then additional discussion (scripted or unscripted) may be necessary to converge on an interpretation. In general, when the respondent's circumstances do not map onto the survey designers' concepts in a straightforward way, respondents may need clarification on what should count, for current purposes, as an instance of the survey concept, if they are to comprehend questions as the survey designers intend (see Schober, 1998).

STUDY 1

In the experiment reported here, we systematically varied when and how respondents in telephone surveys could receive clarification from interviewers presenting standardized

wording. Interviewers either provided *no clarification*, clarification only when respondents requested it (*respondent-initiated clarification*) or clarification both when respondents requested it and whenever else they believed respondents needed it (*mixed-initiative clarification*). Some interviewers were trained to read scripted definitions *verbatim*, and others were allowed to *paraphrase* the definitions. Thus there were five experimental conditions, creating a 2×2 factorial design that crosses when respondents were given clarification (respondent-initiated or mixed-initiative) with how it was given (scripted or paraphrased), along with a control condition in which no clarification was given. Respondents answered questions about behaviours from ongoing large-scale US government surveys; they answered on the basis of fictional scenarios, so that we could directly assess response accuracy—the extent to which responses matched what was dictated by the official government definitions for key survey concepts, which indicated accurate comprehension. The scenarios were designed to map onto the survey questions in either a straightforward or complicated way, as in our earlier car tyre/truck tyre example.

We predict that comprehension should benefit when there is more opportunity for clarification—but only to the extent that respondents and survey designers interpret concepts in the survey questions differently, which should be more likely for scenarios with complicated mappings. When respondents' circumstances are straightforward, extra clarification should not be necessary.

Procedure

Some of the data that we report here—certain analyses in the no clarification and mixed-initiative paraphrased clarification conditions—have been reported in Schober and Conrad (1997). Here we treat the entire data set as one because the data across the five conditions are comparable on virtually every dimension. The interviewers are demographically comparable and have equivalent amounts of training and interviewing experience at the same interviewing facility; the respondents were recruited from the same diverse subject population using the same recruitment methods and are demographically comparable; and the experimental setting, survey questions, and fictional scenarios that respondents answered the questions about were identical. Also, it was not possible to replicate the Schober and Conrad (1997) conditions here because there were not enough interviewers at the telephone interviewing facility to which we had access. The Schober and Conrad (1997) data were collected in 1996, and the data for the remaining three conditions were collected in 1998.

Questions

All respondents were asked the same 12 questions from ongoing US government surveys. Four questions were about employment, adapted from the Current Population Survey (e.g. 'Last week, did Chris do any work for pay?'); four questions were about housing, adapted from the Consumer Price Index Housing Survey (e.g. 'How many people live in this house?'); four questions were about purchases, adapted from the Current Point of Purchase Survey (e.g. 'Has Alexander purchased or had expenses for college tuition or fixed fees?'). The questions had all been pretested for intelligibility, in some cases extensively. For each question, the sponsoring organization had developed official definitions for key concepts. For example, for 'How many people live in this house?' the official definition is this:

> A person is considered to be living in a housing unit even if the person is not present at the time of the survey. Live-in servants or other employees, lodgers, and members of the

> household temporarily away from the unit on business or vacation are included in the count. Do NOT count any people who would normally consider this their (legal) address but who are LIVING away on business, in the armed forces, or attending school (such as boarding school or college). Do NOT count overnight lodgers, guests and visitors. Do NOT count day employees who live elsewhere. (US Department of Commerce, 1994).

The complete set of questions and definitions can be found in Schober and Conrad (1997).

Within each domain (employment, housing, or purchases), respondents answered the four questions in the order in which they appeared in the survey from which they were drawn. The order of domains was counterbalanced across respondents.

Scenarios

Respondents answered on the basis of fictional scenarios. These consisted of floor plans, work descriptions, and purchase receipts; the scenarios were never seen by the interviewers. For each respondent, six scenarios described situations that mapped onto questions in a straightforward way, and the other six described situations that mapped onto questions in a complicated way; different respondents saw different scenarios for different questions. The ordering of the mappings was counterbalanced across respondents; for example, if one respondent answered questions 1 and 2 on the basis of scenarios that mapped onto the questions in a complicated way, a corresponding respondent would answer the same two questions on the basis of straightforward scenarios.

Complicated scenarios were created on the basis of the official concept definitions, which included statements of what circumstances or behaviours qualify for inclusion or exclusion. We selected particular components of definitions that allowed us to instantiate the potential ambiguities in comprehensible scenarios. For every complicated mapping, we created a corresponding straightforward scenario that avoided the ambiguity. As the comprehension accuracy results will confirm, the scenarios we created did indeed have the qualities we believed they would.

Thus, for the question 'Last week, did Pat have more than one job, including part-time, evening, or weekend work?', the corresponding definition for 'more than one job' states that:

> If an individual does the same type of work for more than one employer in an occupation where it is common to have more than one employer, do not consider the individual a multiple jobholder. Examples include private household or domestic workers including babysitters, chauffeurs, gardeners, handypersons, cooks, and maids.

The complicated scenario for this question described Pat as babysitting for different families on different days of the week. The straightforward scenario for this question described Pat as babysitting for one family all week, which is clearly one job. The official concept definition thus clarified what the correct answer should be; Pat should be considered to have one job even if she has multiple employers.

Participants

The 55 interviewers were professional Census Bureau interviewers (43 female, 12 male) calling from the Hagerstown, MD telephone facility. They averaged 59.5 months of interviewing experience, ranging from 2 to 165 months. There were no reliable differences in interviewing experience between the different interviewing groups, $F(4, 50) = 0.69$, *ns*. Each interviewer telephoned two respondents in the Bureau of Labor Statistics laboratory

in Washington, DC (except for one who only telephoned one respondent). The interviewers never saw the scenarios and so did not know what the correct answers were. Because the mappings differed in the two interviews conducted by each interviewer, they also could not predict the correct answers based on earlier ones.

The 109 paid respondents (two per interviewer, with one exception) were recruited from the Bureau of Labor Statistics subject pool; they had responded to an ad in the *Washington Post*. They represented a range of demographic characteristics: 60 were women and 49 were men; 30 were black, 73 were white, five were Asian, and one was Hispanic, and there were comparable numbers of respondents from each category in each group. Respondents averaged 15.8 years of education; the groups did not differ reliably in education.

Interviewer training

All interviewers were first trained together in groups on the key survey concepts for about an hour. This training included a quiz and group discussion; as the definitions could be quite long and complicated, we wanted to make sure that all interviewers understood them thoroughly.

Then interviewers received additional training in one of five interviewing techniques. In all cases interviewers were first to read the survey questions as worded. The strictly standardized group was trained to always leave the interpretation up to the respondent, and only to probe 'neutrally' if a respondent asked for clarification (following Fowler and Mangione's 1990 guidelines). The two respondent-initiated clarification groups were trained to clarify the meaning of questions only if respondents explicitly requested clarification. We defined explicit requests fairly rigorously; interviewers were only to provide clarification when respondents asked explicit questions like 'Does babysitting for two families count as one job or two?' or expressed their uncertainty directly, as in 'I'm not sure what you mean by that question.' Interviewers were not to provide clarification if respondents described their fictional circumstances rather than answering the question, as in 'Well, I babysit for two families', nor were they to provide clarification if respondents' answers merely sounded uncertain.

The two mixed-initiative clarification groups were trained to provide clarification whenever they felt respondents needed it, whether or not respondents had explicitly asked for clarification. This meant that they were licensed to provide unsolicited clarification when they deemed it necessary.

Among the four groups who provided clarification, two were trained to read scripted definitions; this consisted in reading at least one full sentence of the definition at a time, up to reading the entire definition. (Sentences could be read in any order the interviewers chose; the point was that interviewers should not have the leeway to select their own wording.) The other two groups were trained to explain the concepts in their own words, although they were allowed to read parts of definitions if they preferred. Thus interviewers in the mixed-initiative paraphrased clarification group essentially used the resources of ordinary conversation, while the others were restricted in one way or another.

Results

Before turning to the findings on comprehension accuracy and interview length, we first needed to verify that interviewers had implemented the different interviewing techniques correctly. We did this by closely examining transcripts of the interviews. From the 1296

question-answer sequences recorded, the authors coded each conversational move by interviewer and respondent according to a coding scheme based on the Schober and Conrad (1997) scheme (please contact the authors for details). Approximately 60% of the question-answer sequences were independently coded by two of the authors; coding agreement was 97.7% (agreement on 4263 of the 4371 moves coded), and discrepancies were resolved by discussion.

Interviewers implemented the different interviewing techniques as they had been trained. Interviewers in the no clarification group never provided definitions, while those in the other groups provided definitions for an average of 42.9% of the question-answer sequences, $F(1, 106) = 48.67$, $p < 0.0001$. Interviewers defined concepts reliably more often in the mixed-initiative groups (for 57.2% of the question-answer sequences) than in the respondent-initiated clarification groups (for 27.3% of the question-answer sequences), $F(1, 74) = 34.90$, $p < 0.0001$.

For further evidence that interviewers implemented the techniques as they were trained, we examined just those question-answer sequences where interviewers provided clarification. In the respondent-initiated clarification groups, interviewers almost always provided the clarification only after the respondent explicitly requested it (for 89.4% of these question-answer sequences), as in this example:[2]

I: Has Kelly purchased or had expenses for household furniture?
R: What do you mean by household furniture?
I: Okay. ah household furniture eh tables, chairs, footstools, sofas, china cabinets, utility carts, bars, room dividers, bookcases, desks, beds, mattresses, box springs, chests of drawers, night tables, wardrobes, and unfinished furniture. Do not include TV, radio, and other sound equipment, lamps and lighting fixtures, outdoor furniture, infants' furniture, or appliances.
R: Ask the question again *then*.
I: *Okay.* Has Kelly purchased or had expenses for household furniture?
R: No she hasn't.
I: Okay.

They did this reliably more often than interviewers in the mixed-initiative clarification groups (for 65.5% of the question-answer sequences), $F(1, 35) = 7.07$, $p = 0.012$, who were licensed to provide clarification even without an explicit request, as in this example:

I: How many hours per week does Mindy usually work at her job?
R: (long pause)
I: And by usually I mean fifty percent of the time or more, or the most frequent schedule during the past four or five months.
R: Uh, fifty hours.
I: Okay.

Interviewers trained to read scripted definitions presented exactly verbatim information almost always (for 83.9% of the question-answer sequences in which they provided clarification), as in the first example above. They did this reliably more often than

[2]In the transcribed excerpts, overlapping speech is enclosed in asterisks. A period between two spaces (.) represents a pause. A colon within a word indicates a lengthened sound. A hyphen at the end of a word indicates that the word was cut off. Question marks indicate rising intonation, and utterance-final periods indicate falling or flat intonation, regardless of whether the utterance is a question or an assertion.

interviewers who were allowed to use their own words (for 40.5% of the question-answer sequences), $F(1, 35) = 39.96$, $p < 0.0001$, as in this example:

I: Has Kelly purchased or had expenses for household furniture.
R: Uh- does a lamp is a lamp considered household furniture.
I: Okay no lamps are not considered to be household uh furniture according to our definition.
R: Okay then no.
I: No? Okay.

All this suggests that the interviewing techniques were implemented well enough to allow meaningful comparisons of respondent comprehension.

Comprehension accuracy

Overall, respondents comprehended the survey questions more accurately when they could receive clarification than when the interpretation was left up to them, $F(1, 107) = 45.53$, $p < 0.0001$. As Figure 1 shows, across all five interviewing conditions respondents' comprehension for straightforward scenarios was excellent; their answers almost always matched what the official definitions required. It was only for the complicated mapping scenarios that interviewing technique made a difference, interaction $F(1, 107) = 42.81$, $p < 0.0001$.

Focusing on the data from the four groups where interviewers provided clarification, responses for complicated mappings were reliably more accurate when interviewers provided mixed-initiative clarification than when they only responded to explicit requests for clarification, $F(1, 84) = 18.46$, $p < 0.001$. Responses were marginally better when interviewers used their own words to clarify question meaning rather than reading scripted definitions, $F(1, 84) = 3.26$, $p = 0.074$, but this may have been because of the substantial increase in accuracy in the conversational (paraphrased mixed-initiative clarification) case, interaction $F(1, 84) = 7.86$, $p < 0.006$.

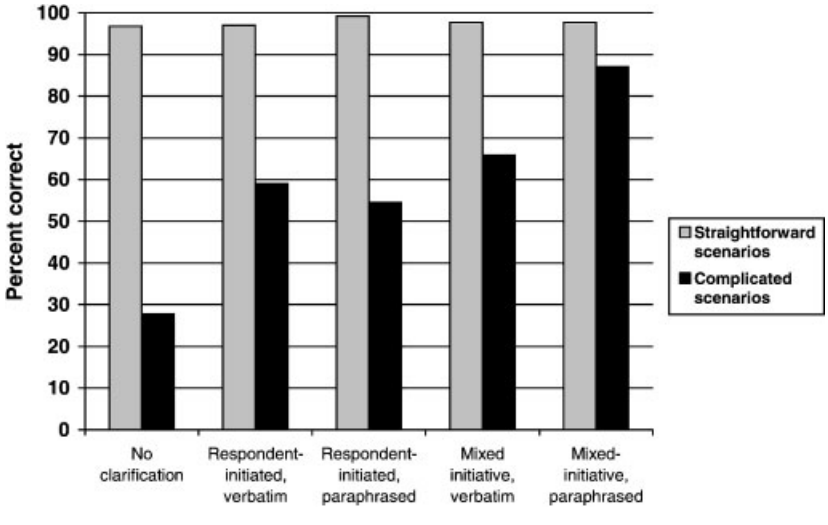

Figure 1. Comprehension accuracy, Study 1

Quality of clarification

The clarification that interviewers provided was highly accurate. For the 432 question-answer sequences where any clarification was provided, they presented completely accurate information in 411 cases (95.1% of the time). And inaccurate information did not always lead to inaccuracy; in the 20 cases where interviewers presented any inaccurate information, respondents still produced accurate answers for 15 cases, as in this interaction from a mixed-initiative paraphrased clarification interview:

I: Last week did Chris do any work for pay.
R: Mm:
I: And uh . by pay . we mean is he uh . uh does he receive uh monies for his work. Uh we do not include pay in kind such as food, lodging or- and lodging for work or expense accounts. But we do include um . is that what I said? Yes, uh . uh I read you the wrong . yeah, that's it, okay, yes, for pay. Mm-hm, that was right.
R: Okay.
I: Uh, did he do any work for pay?
R: Oh, it's a she,
I: *oh okay.*
R: *by the way.*
I: okay, she . did she do any work for pay.
R: She didn't receive any paychecks, but the work sh- for the work she does she receives a reduction in her tuition.
I: Oh.
R: Every year.
I: Okay, yes so we all- we uh count college assistantships and fellowships.
R: Okay.
I: as um income, so that would be yes then.
R: Okay.
I: And um and this would be on a regular basis, right?
R: Um . well she gets her tuitions reduced every year, it looks like *so*
I: *hmm* Well *I don't-*
R: *And she* works every week for ten hours.
I: Oh okay yes, on a regular basis that would be . correct then.
R: Okay.

Here the interviewer invented a definition for pay ('receiving monies for work') that was not part of the training materials, but eventually discovered the relevant issue (college assistantships) and provided accurate clarification.

Interviewers provided the information that respondents needed to hear for 81.5% of the question-answer sequences where clarification was provided: that is, they presented the components of the official definitions relevant to the fictional scenarios. But interviewers sometimes also presented unnecessary parts of definitions, telling respondents more than they needed to hear (presenting additional components of definitions beyond those strictly necessary). In this example from a scripted clarification interview, all the respondent really needs to hear is whether working on a continuing basis as a babysitter counts as a job, but the interviewer provides plenty of extra information:

I: Last week, did Pat have more than one job, including part-time, evening or weekend work?
R: U:h what do you mean by part-time?
I: All right a j- Or a job exists when there is a definite arrangement for regular work every week or every month, for pay or other compensation. Such as profits, anticipated profits, or pay in kind such as room and board. A formal, definite arrangement with one or more employers to work on a continuing basis for a specified number of hours per week or days per month, but on an irregular schedule during the week or month is also a job.
R: Okay. Now what was your question again?

And in 46 cases (of the 340 complicated scenarios where clarification was provided) interviewers provided only irrelevant parts of definitions—that is, they provided accurate but unhelpful clarification. In these cases, as one might expect, respondents were not much more accurate (15 out of 46 cases, 33%) than when no clarification had been given (27%).

The different interviewing techniques led respondents to behave differently as well. Respondents in the no-clarification group almost never requested clarification (for 1.6% of the questions), while respondents in the other groups explicitly requested clarification 32.9% of the time, $F(1, 106) = 34.77$, $p < 0.0001$. At a gross level, respondents recognized when they needed clarification: if we examine the four groups where clarification was available, respondents were far more likely to explicitly ask for clarification for complicated mapping scenarios (for 47.0% of the questions) than for straightforward mappings (for 15.8% of the questions), $F(1, 74) = 83.69$, $p < 0.0001$. But they did not request clarification every time they needed it; we know this from the fact that their comprehension accuracy was far from perfect. And getting unsolicited clarification from interviewers did not make respondents more likely to initiate clarification themselves (despite improving their comprehension); respondents in the mixed-initiative groups requested clarification for 34.7% of the questions, not reliably more often than respondents in the respondent-initiated clarification groups' 28.2%, $F(1, 74) = 1.47$, *ns*.

Interview duration

Interviews took longer when interviewers provided more clarification. Table 1 shows the median interview duration in minutes for all five types of interviewing. Focusing on the four groups where interviewers clarified question meaning, we see that mixed-initiative interviews, where interviewers provided unsolicited clarification, took longer than

Table 1. Median interview duration in minutes

	Initiative		
Type of clarification	No clarification	Respondent-initiated clarification	Mixed-initiative clarification
None	3.41 (min 2.48, max 5.99)		
Verbatim clarification		7.83 (min 4.00, max 13.95)	7.62 (min 3.95, max 14.90)
Paraphrased clarification		7.43 (min 2.75, max 10.90)	11.47 (min 6.10, max 35.44)

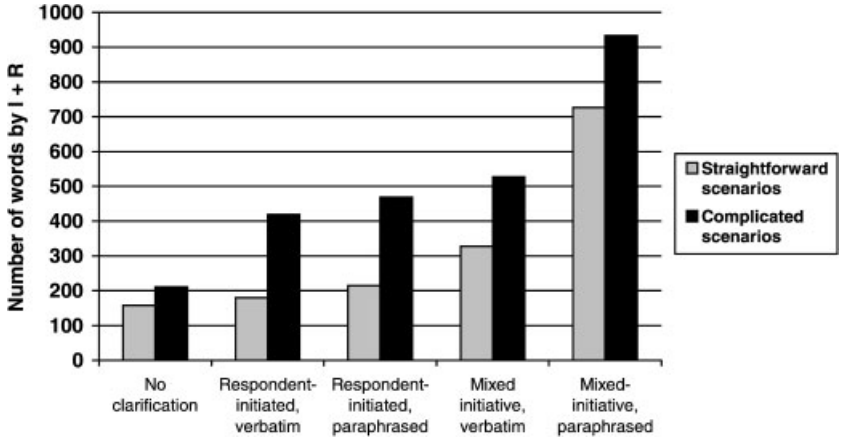

Figure 2. Interview duration (total words per question by interviewer and respondent), Study 1

interviews with only respondent-initiated clarification, $F(1, 83) = 36.32$, $p < 0.001$. Interviews took reliably longer when interviewers used their own words to clarify the questions rather than reading the script, $F(1, 83) = 17.41$, $p < 0.001$, but as the table shows, the largest increase in duration was for the mixed-initiative paraphrased case, interaction $F(1, 83) = 10.09$, $p = 0.002$.

Part of the increased interview duration for interviews with clarification was not merely because interviewers spent more time clarifying complicated mappings; they also spent a great deal of additional time discussing straightforward mappings, which really did not need to be discussed. Figure 2 plots how much time was spent in all five kinds of interviews on complicated versus straightforward mappings, as measured by the number of words spoken by interviewer and respondent together per question. As the figure shows, when interviewers provided clarification only when explicitly requested, the amount of discussion of straightforward mappings was not much greater than in strictly standardized interviews, while the amount of discussion of complicated mappings increased. When interviewers could also provide unsolicited clarification, the amount of 'unnecessary' time spent on straightforward mappings increased.

Across all five interviewing groups, then, we see an emerging pattern: response accuracy for complicated mappings increases with interview duration. The duration of the interviews conducted with intermediate levels of clarification was more than twice that of the strictly standardized interviews, and the duration of the most 'conversational' (mixed-initiative paraphrased) interviews was more than three times that of the strictly standardized interviews. In fact, there is a strong linear relationship ($r = 0.98$) between interview duration and response accuracy. For each additional minute that interviewers and respondents spent on clarification, there was a 7% gain in accuracy. This suggests that more clarification to respondents improves response accuracy more, but at a linear increase in interview duration, and thus in survey costs.

STUDY 2

Study 1 demonstrated that when and how interviewers provide clarification can indeed affect the accuracy with which respondents comprehend survey questions about facts and

behaviours. Comprehension accuracy was poorest for the most strictly standardized interviews; it improved when interviewers provided clarification at respondents' request; and it improved further when interviewers also provided clarification at their own initiative. Comprehension accuracy was also better when interviewers paraphrased their clarification rather than reading it verbatim. All this suggests that in standardized interviews, how interviewers behave after reading the questions can have a powerful effect on data quality, particularly when respondents' circumstances do not map onto survey definitions in a straightforward way.

What Study 1 does not demonstrate is how the findings generalize to real-survey settings. Study 1 trained interviewers to administer distinctly different kinds of interviewing, so that we could tease apart factors that are theoretically independent (initiative in clarification and manner of clarification). But interviewing in the real world is unlikely to be so neat. First, organizations that consider themselves to subscribe to standardized practice can vary substantially in what they train their interviewers to do, from forbidding clarification to encouraging interviewers to provide definitions whenever respondents request them (Viterna & Maynard, 2002). Second, studies of interviewer behaviour within organizations that train interviewers to be strictly standardized have demonstrated that interviewers deviate from strict standardization some percentage of the time (see, e.g. Bradburn, Sudman and Associates, 1979; Mangione, Fowler, & Louis, 1992; Morton-Williams, 1979). How do the interviewing techniques actually implemented by standardized interviewers in real-survey organizations affect comprehension accuracy?

Study 2 examined this head on by asking a small sample of interviewers from the same mainstream survey organization used in Study 1 to carry out interviews as they ordinarily do. Respondents in the laboratory answered the same questions from ongoing surveys used in Study 1 on the basis of the same fictional scenarios. Thus, we could characterize where interviewers from this organization fall within the continuum of possible interviewing techniques, and get an idea of how this is likely to affect comprehension accuracy in surveys as they are currently administered. Of course, this is only one organization among many; the effects might be different elsewhere. Also, interviewers instructed to 'do what they ordinarily do' may well act more as they think they ought to than as they actually do day to day. Nonetheless, such a study provides initial evidence on how the findings from Study 1 are relevant to real-world surveys.

Materials

Respondents were asked the same questions used in Study 1, counterbalanced following the same scheme. Respondents answered the questions on the basis of the same fictional scenarios used in Study 1. As in Study 1, scenarios were not available to the interviewers, so interviewers did not know what the correct answers to the questions were, and they could not predict the correct answers from previous interviews. As in the earlier study, for each respondent 50% of the scenarios had complicated mappings and 50% had straightforward mappings, and the order of presentation was counterbalanced according to the same scheme used in Study 1.

Participants

The 11 interviewers were professional Census Bureau interviewers (10 female, one male) at the Hagerstown, MD telephone facility who had not participated in Study 1. They

averaged 20.6 months interviewing experience, ranging from 5 to 45 months. Each interviewer telephoned two paid respondents in the Bureau of Labor Statistics laboratory in Washington, DC (except for one interviewer who only telephoned one respondent).

The 21 respondents were recruited from ads in the *Washington Post*. They came from a range of demographic backgrounds; 10 were female and 11 were male; seven were black, 12 were white, and two were Asian; and they ranged in education from high school diplomas to graduate degrees, averaging 16.3 years of education. The demographics of the group were much the same as for the respondents in Study 1, and so it is reasonable to make comparisons across the groups.

Interviewer training

All interviewers were trained on the key survey concepts, using group discussion and a quiz; the same training for survey concepts was used as for Study 1. But this time there was no particular training on interviewing technique; interviewers were asked to conduct the interviews exactly as they ordinarily do. Note that official practice at this facility is not entirely standardized, according to the very strictest criteria (Fowler & Mangione, 1990). Although in written training materials (US Department of Commerce, 1994) interviewers are instructed to read questions exactly as worded and to use only non-directive probes, video training materials allow interviewers to clarify questions at the respondents' request. Interviewers routinely reported that they could clarify concepts when asked to.

Results

Degree of standardization

Before assessing comprehension accuracy, we examined what kinds of interviews this sample of interviewers gave. The authors coded transcripts of the 21 audio-recorded interviews following Study 1's scheme.

Overall, of the 254 question-answer sequences coded, only 50 (19.8%) deviated in any way from the strictest version of standardization. Thus, in the main, interviewers in this setting seem to conform to the kinds of practices advocated by Fowler and Mangione (1990). But there was substantial variability among interviewers in their practices. One of the 11 interviewers followed strictest standardization to the letter, never deviating in any way from strict practice in either of the two interviews she carried out. Four of the 11 were highly standardized, providing definitions in response to requests for clarification or asking informative follow-up questions no more often than once per interview. At the other extreme, three interviewers deviated from standardization for at least four of 12 questions in each interview, up to as many as six questions. The others fell in between.

The most frequent deviation was asking unscripted clarification questions (34/50 cases). Consider this example, where after the respondent answered the question ambiguously, the interviewer probed in a way that reveals what the official definition counts as a bedroom.

I: How many bedrooms are there in this house.
R: Originally two, but they're using three
I: Okay but there were two rooms designed specifically for bedrooms?
R: Right.
I: Okay.

In the following example, the interviewer uses follow-up questions to hone in on the various features of the official definition, and essentially tells the respondent what the answer ought to be:

I: Does anyone in this household have a business or a farm.
R: Um, Harry does, but he ju- I mean- just works as a gardener though.
I: Okay so is it it's um like a business of his?
R: Uh he just does it uh on the side ever since he retired?
I: Okay, so um like so um do you classify is a- as a business I mean does he have like a listing in a- in a classified section in the telephone or *is it on the side*
R: *No, no*, it's just something that he does on the side.
I: So it's- I mean so it's technically it's not really a business then is that what you're saying, it's just something he *does-*
R: *It's* not a business, but he did- did
I: Is *this-*
R: *consider it* work, I'd consider it work though
I: Okay, but not a business?
R: Yes.
I: Okay.

In addition to using unscripted follow-up questions, interviewers deviated from strict standardization by providing paraphrased (never verbatim) official definitions (seven cases of accurate paraphrasing and four cases of inaccurate paraphrasing), rewording questions in order to probe further, and explicitly telling respondents what the answer should be, as in 'That would be two bedrooms, then'.

Most of these deviations from strictest practice conformed with the spirit (if not the letter) of what interviewers had been trained to do; recall that they are encouraged to clarify meaning when respondents request it. Interviewers (and respondents) did fairly well at recognizing when clarification was needed; of the 50 cases where they deviated from strictest practice, 34 (68%) occurred when mappings between questions and circumstances were complicated, as compared with 16 (32%) when the mappings were straightforward. To the interviewers' credit, the information they provided was highly accurate; 45 of the 50 deviations from strict standardization were entirely accurate and could not be seen as misleading the respondents.

Comprehension accuracy

As in Study 1, respondents' answers matched the official definitions virtually perfectly for straightforward mappings (98% accuracy). But their response accuracy for complicated mappings was quite poor, 36%. If we compare these accuracy rates to Study 1, we see that the pattern is not reliably different than for the strictly standardized interviews in that study, and that accuracy is far poorer than in cases where interviewers had been trained to provide clarification. This confirms, on the one hand, that interviewers at this organization are performing comparably to strictly standardized interviewers; unfortunately, it also demonstrates that the interviewers' performance does not lead to good comprehension of question concepts for complicated mapping scenarios.

A more detailed examination of comprehension accuracy shows that interviewer deviations from strict standardization led to substantial improvement in accuracy (see Figure 3). In fact, the 79% accuracy rate for complicated mappings when interviewers deviated is comparable to the 87% rate in Study 1's mixed-initiative paraphrased interviews.

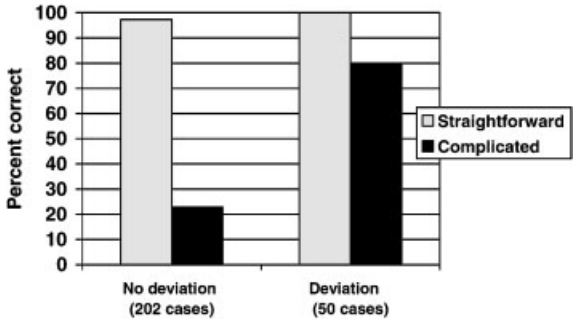

Figure 3. Comprehension accuracy when interviewers deviated from strictly standardized practice, Study 2

This finding is supported from several other views of the results. If we examine average comprehension accuracy for each interviewer, we see that interviewers who deviated more from strict standardization elicited more accurate responses for complicated mappings, $r(11) = 0.76$, $p = 0.007$. If we examine how often interviewers deviated from strict standardization on a question-by-question basis, we see that comprehension accuracy for complicated mappings was greater for those questions on which interviewers deviated more, $r(12) = 0.59$, $p < 0.05$. Of course, the sample is quite small and we should be careful about generalizing, but the overall picture is consistent with the finding from Study 1: greater deviation from strict standardization can lead to improved comprehension accuracy for complicated mappings.

Interview duration

Interviews took a median of 4.26 min, ranging from 2.95 to 8.18 min. They did not take reliably longer than the strictly standardized interviews in the Study 1 (3.41 min, ranging from 2.48 to 5.99); both standardized interviews and the interviews in Study 2 were much quicker than the mixed-initiative paraphrased interviews (11.47 min, ranging from 6.10 to 35.44 min). The total number of words per question-answer sequence used by interviewers and respondents when mappings were straightforward (234) and complicated (334) in Study 2 is relatively close to the number of words in the strictly standardized interviews in Study 1 (158 for straightforward mappings and 211 for complicated mappings); however interviewers and respondents used far fewer words in Study 2 than in the mixed-initiative paraphrased interviews of Study 1 (727 for straightforward mappings and 933 for complicated mappings). If we break down these word counts further, we see that what really took time was interviewers' deviations from strict standardization (see Figure 4). For those complicated mapping cases where interviewers did not deviate from strict standardization, the interviewers took no longer than for straightforward cases.

So the emerging picture is that comprehension accuracy can improve substantially when interviewers deviate from strict standardization, but this deviation takes time. This provides more naturalistic evidence for the tradeoff between comprehension accuracy and interview efficiency noted in Study 1.

Interestingly enough, interviewers with the most experience (measured in months) produced the least accurate responses for complicated mappings, $r(11) = -0.63$, $p = 0.038$. This may be related to the (non-significant) trend for more experienced interviewers to deviate from strict standardization less often, $r(11) = -0.51$, $p = 0.112$.

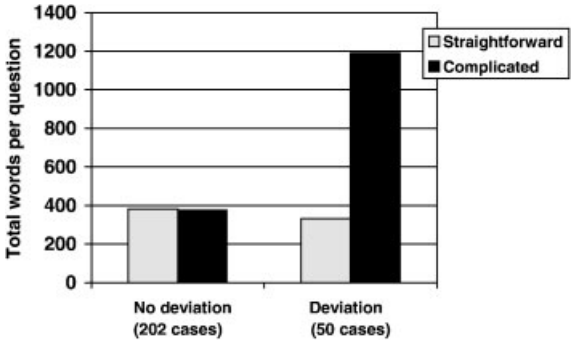

Figure 4. Interview duration (total words per question by interviewer and respondent) when interviewers deviated from strictly standardized practice, Study 2

CONCLUSIONS

Taken together, the results of Studies 1 and 2 confirm that what interviewers do after they read the questions in standardized surveys about facts and behaviours can drastically affect whether respondents understand question concepts as the survey designers intend. The corollary is that respondents answering seemingly unambiguous questions about facts and behaviours can interpret question concepts quite differently than intended when their circumstances do not match the prototypical situations one might first think of (see also Conrad & Schober, 2000; Schober, 1998; Schober & Conrad, 1997, 2002). For more protoypical situations, in contrast, response accuracy can be excellent.

Study 1's findings go beyond earlier work (Conrad & Schober, 2000; Schober & Conrad, 1997) by demonstrating the effects of particular interviewer behaviours: initiating and paraphrasing clarification. Training interviewers to initiate clarification whenever they feel it is necessary, rather than relying on respondents to ask for clarification, promotes the most accurate respondent comprehension; respondents can not be relied on to ask for clarification at all times when they need it. Training interviewers to clarify in their own words rather than relying on a script may also have some benefits. But the clear interaction between these two factors shows that the real benefit is in training interviewers to provide clarification in their own words whenever they deem it necessary.

Study 2 demonstrates that the factors manipulated in Study 1 are dimensions on which interviewing varies in an actual survey setting. At least at one survey centre, some interviewers follow strictly standardized procedure, and others provide clarification both after respondents ask for it and also by asking pointed follow-up questions that provide definitional information. When clarification is provided, it is likely to be paraphrased rather than verbatim. And, as in Study 1, when interviewers deviate more from the most strictly standardized procedures, respondents comprehend and answer survey questions more accurately, as Suchman and Jordan (1991) have proposed should be the case. On average, however, current practice at Study 2's survey facility is far closer to strictly standardized than to conversational interviewing, and as such it leads to poorer comprehension accuracy than it might.

All in all, the data show that allowing interviewers to use some of the collaborative resources of ordinary conversation (providing respondent-initiated or scripted clarification) is better than denying all of them (strictly standardized interviewing), but even better is allowing interviewers to collaborate more as they do in spontaneous conversation

(mixed-initiative paraphrased clarification). As such, the data are consistent with arguments that language comprehension—both in spontaneous and scripted situations—is part of a collaborative process (Clark, 1996; Clark & Wilkes-Gibbs, 1986). Under a collaborative view of language use, in contrast to an autonomous view (Schober & Clark, 1989), the meaning of a speaker's utterance is not inherent only in the speaker's words; the same words can mean different things to different listeners, and an addressee has not finished interpreting a particular utterance until both speaker and addressee have agreed that the utterance has been understood sufficiently for current purposes. Speakers in ordinary conversation do not rely on scripts to clarify their meaning; they can adjust their utterances to fit their addressees' needs and provide clarification whenever they believe their addressees need it, even if their addressees have not requested clarification. Although survey interviews differ in important ways from spontaneous unscripted conversations (Houtkoop-Steenstra, 2000; Schaeffer, 2002; Schober, 1999; Schober & Conrad, 2002), survey respondents seem to bring the same interpretive resources to the survey setting as they do to other settings.

Some researchers (e.g. Tourangeau, Rips, & Rasinski, 2000, pp. 57–59) have asserted that a collaborative view of survey interactions leads to the untenable position that the meanings of words would have to be renegotiated each time they are used. This is not what we are proposing; our data, in fact, show that for straightforward scenarios respondents understood survey concepts as intended without any clarification. Obviously, under the right circumstances shared background knowledge about the basic semantics of a language can lead to successful comprehension without extra collaborative moves (depending on one's theory of semantics). Our results refine a collaborative view of language use by demonstrating, in a substantive domain of application, *when* extra collaboration leads to improved comprehension: when respondents' (addressees') circumstances do not align prototypically with the survey designers' (authors') definitions.

The results also clearly demonstrate that improved accuracy from extra collaboration comes at the cost of increased interview duration; greater accuracy comes at a greater cost, as we have also found in a telephone survey of a national sample (Conrad & Schober, 2000) and in computer-administered self-interview questionnaires (Schober, Conrad, & Bloom, 2000). We propose that there is always a tradeoff between the need for accurate data and the costs of getting them. The tradeoffs are particularly important to consider given Study 2's finding of how substantially individual interviewers and examiners can vary in how they implement standardized procedures (see also Marlaire & Maynard, 1990; Viterna & Maynard, 2002).

Several important caveats are in order. First, how these findings apply to real-world interviewing settings depends on the frequency of complicated mappings between respondents' circumstances and survey definitions. If almost all respondents interpret a concept the same way, and if the circumstances on which they are basing their answers align with the survey designers' notions, then leaving the interpretation up to respondents will work. For two of the three domains tested here, housing and purchases, we have evidence that in a national sample survey, complicated mappings are frequent enough that leaving the interpretation up to respondents can lead to substantial misinterpretation (Conrad & Schober, 2000). But this varies from question to question; for some question concepts, interpretation is much more uniform (and correct) than for others. So to judge the generality of these findings, it is necessary to know much more about the range of possible interpretations and circumstances within a population than survey designers or most social researchers usually know, even from extensive pretesting.

Second, our studies examine questions about verifiable facts and behaviours. It is unknown how these findings apply to questions about attitudes or opinions, where the 'true value' a question is designed to tap may not have even existed prior to the question's being asked. Certainly there is the danger that clarifying the meaning of terms like 'abortion', 'race', or 'happiness' in opinion questions might bias responses. On the other hand, if different respondents interpret such terms quite differently, then measures of attitudes and opinions may not be measuring what they intend to be measuring.

Third, our claim is that our findings about comprehension of wording in surveys about facts and behaviours should extend to other domains that use standardized wording and interviewing procedures, like standardized tests, psychological assessments, experiment instructions, and user interfaces. So, for example, in standardized intelligence tests like the WAIS-III, how examiners choose to probe and query, and the kinds of feedback they give that might provide evidence of how examinees should comprehend questions, ought to have a measurable impact on test scores, as preliminary evidence suggests (Condon & Schober, 2002). But, of course, the extent of the generalizability of our findings to other domains that rely on standardized language remains to be shown empirically.

One might ask whether the interpretive variability our respondents demonstrated could be resolved simply by wording questions better. Perhaps the potentially dangerous prospect of empowering interviewers to clarify as they see fit could be avoided simply by making sure that questions really are intelligible to the majority of respondents, through more extensive pretesting procedures. The trouble with this solution is, first, that questions cannot get much simpler than 'Last week, did you do any work for pay?'; any additional clarification added to handle the kinds of complicated situations that respondents face ends up looking a lot like interviewing with clarification. Worse, such a solution simply increases the length of a survey for all respondents, because everyone gets the clarification every time; longer surveys may be less likely to be completed.

The degree to which respondents' conceptions of standardized wording overlap with researchers' conceptions may, of course, vary in different domains or for different populations of respondents. If conceptual mismatches are known to be rare, or if the need for precision is not pressing, then the extra costs of clarification may not be justified. On the other hand, if the frequency of mismatches is unknown, or if they are known to be frequent, then encouraging researchers to clarify may be a good idea. Just how much leeway researchers should be given would then depend on how certain one needs to be that respondent comprehension—and thus the data—are accurate.

ACKNOWLEDGEMENTS

We thank Susan Brennan, Kimberly Clark, Cathy Dippo, Laura Hahn Lind, Kate Richardson, Clyde Tucker, Ron Tucker and Eduardo Vega for assistance and advice. This material is based upon work supported by the National Science Foundation under grants No. SBR-9730140 and ISS-0081550 and by the Bureau of Labor Statistics. This paper reflects the opinions of the authors and not the Bureau of Labor Statistics.

REFERENCES

Barrow, H. S. (1993). An overview of the use of standardized patients for teaching and evaluating clinical skills. *Academic Medicine, 68*, 443–453.

Beatty, P. (1995). Understanding the standardized/non-standardized interviewing controversy. *Journal of Official Statistics, 11*, 147–160.
Bradburn, N. M., Sudman, S., and Associates. (1979). *Improving interview method and questionnaire design: Response effects to threatening questions in survey research.* San Francisco: Jossey-Bass.
Butcher, J. N. (1972). *Objective personality assessment: Changing perspectives*. New York: Academic Press.
Clark, H. H. (1996). *Using language*. Cambridge, MA: Cambridge University Press.
Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1–39.
Condon, S. P., & Schober, M. F. (2002). Abstract of the Psychonomic Society 43rd Annual Meeting.
Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64*, 1–28.
Fisher, R. P., Geiselman, R. E., & Raymond, D. S. (1987). Critical analysis of police interview techniques. *Journal of Police Science and Administration, 15*, 177–185.
Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* Newbury Park, CA: SAGE Publications, Inc.
Grudin, J. (1989). The case against user interface consistency. *Communications of the ACM, 32*, 1164–1173.
Gulliksen, J., & Sandblad, B. (1995). Domain-specific design of user interfaces. *International Journal of Human-Computer Interaction, 7*(2), 135–151.
Gwyer, P., & Clifford, B. R. (1997). The effects of the cognitive interview on recall, identification, confidence, and the confidence/accuracy relationship. *Applied Cognitive Psychology, 11*(2), 121–145.
Houtkoop-Steenstra, H. (2000). *Interaction and the standardized survey interview: The living questionaire.* Cambridge, UK: Cambridge University Press.
Huffcutt, A. I., & Woehr, D. J. (1999). Further analysis of employment interview validity: a quantitative evaluation of interview-related structuring methods. *Journal of Organizational Behavior, 20*, 549–560.
Kahn, R., & Cannell, C. (1957). *The dynamics of interviewing: Theory, techniques, and cases*. New York: John Wiley & Sons.
Mangione, T. W., Fowler, F. J., & Louis, T. A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics, 8*, 293–307.
Marlaire, C. L., & Maynard, D. W. (1990). Standardized testing as an interactional phenomenon. *Sociology of Education, 63*, 83–101.
Meehl, P. (1947). The dynamics of 'structured' personality tests. *Journal of Clinical Psychology, 3*, 296–303.
Memon, A., Holley, A., Milne, R., Koehnken, G., & Bull, R. (1994). Towards understanding the effects of interviewer training in evaluation of the cognitive interview. *Applied Cognitive Psychology, 8*, 641–659.
Mishler, E. G. (1986). *Research interviewing*. Cambridge, MA: Harvard University Press.
Morton-Williams, J. (1979). The use of 'verbal interaction coding' for evaluating a questionnaire. *Quality and Quantity, 13*, 59–75.
Pangalo, G. J. (1993). Designing the user-interface. *Computers in Industry, 22*, 193–200.
Pangaro, L. N., Worth-Dickstein, H., Macmillan, M. K., Klass, D. J., & Shatzer, J. H. (1997). Performance of 'standardized examinees' in a standardized-patient examination of clinical skills. *Academic Medicine, 72*(11), 1008–1010.
Schaeffer, N. C. (2002). Conversation with a purpose–or conversation? Interaction in the standardized interview. In D. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 95–123). New York: Wiley.
Schober, M. F. (1998). Conversational evidence for rethinking meaning. *Social Research, 65*, 511–534.
Schober, M. F. (1999). Making sense of questions: an interactional approach. In M. G. Sirken, D. J. Hermann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research* (pp. 77–93). New York: John Wiley & Sons.
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-Steenstra, 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., & Bloom, J. E. (2000). Clarifying word meanings in computer-administered survey interviews. In L. R. Gleitman, & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 447–452). Mahwah, NJ: Lawrence Erlbaum Associates.

Smith, V. L., Kassin, S. M., & Ellsworth, P. C. (1989). Eyewitness accuracy and confidence: within-versus between-subject correlations. *Journal of Applied Psychology, 74*, 356–359.

Speak, M. (1967). Communication failure in questioning: errors, misinterpretations, and personal frames of reference. *Occupational Psychology, 41*(4), 169–181.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85*, 232–253.

Suchman, L., & Jordan, B. (1991). Validity and the collaborative construction of meaning in face-to-face surveys. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 241–267). New York: Russell Sage Foundation.

Tanaka, T., Eberts, R. E., & Salvendy, G. (1991). Consistency of human-computer interface design: quantification and validation. *Human Factors, 33*(6), 653–676.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York: Cambridge.

U.S. Department of Commerce. (1994). *Current Population Survey interviewing manual* (CPS-250). Washington, DC: Bureau of the Census.

Viterna, J. S., & Maynard, D. W. (2002). How uniform is standardization? Variation within and across survey research centers regarding protocols for interviewing. In D. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 365–397). New York: Wiley.

Wechsler, D. (1997). *WAIS-III administration and scoring manual: Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wright, O. R. (1969). Summary of research on the selection interview since 1964. *Personnel Psychology, 22*, 391–413.

