Public Opinion Quarterly, Vol. 73, No. 1, Spring 2009, pp. 32–55

SOURCES OF ERROR IN COGNITIVE INTERVIEWS

FREDERICK G. CONRAD
JOHNNY BLAIR

Abstract Cognitive interviewing is used to identify problems in questionnaires under development by asking a small number of pretest participants to verbally report their thinking while answering the draft questions. Just as responses in production interviews include measurement error, so the detection of problems in cognitive interviews can include error. In the current study, we examine error in the problem detection of both cognitive interviewers evaluating their own interviews and independent judges listening to the full set of interviews. The cognitive interviewers were instructed to probe for additional information in one of two ways: the *Conditional Probe* group was instructed to probe only about what respondents had explicitly reported; the *Discretionary Probe* group was instructed to probe whenever they felt it appropriate. Agreement about problems was surprisingly low overall, but differed by interviewing technique. The Conditional Probe interviewers uncovered fewer potential problems but with higher inter-judge reliability than did the Discretionary Probe interviewers. These differences in reliability were related to the type of probes. When interviewers in either group probed beyond the content of respondents' verbal reports, they were prone to believe that the respondent had experienced a problem when the majority of judges did not believe this to be the case (false alarms). Despite generally poor performance at the level of individual verbal reports, judges reached relatively consistent conclusions across the interviews about which questions most needed repair. Some practical measures may improve the conclusions

FREDERICK G. CONRAD is with the University of Michigan's Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48104, USA. JOHNNY BLAIR is with Abt Associates, 4550 Montgomery Avenue, Suite 800 North, Bethesda, MD 20814-3343, USA. *The authors would like to thank the following people for their help on this project: Greg Claxton, Nadra Garas, Jane Joseph, Rachel Levenstein, Nileeni Meegama, Nick Prieur, Elena Tracy, and Sara Woldehana. In addition, the authors are grateful to Gordon Willis, Edward Blair, and two anonymous reviewers for valuable comments on earlier drafts of this article. The authors thank the Bureau of Labor Statistics for supporting several of the reported activities through a contract with Westat, Inc. and a subcontract with the University of Maryland. None of the opinions expressed here are those of the Bureau of Labor Statistics; they are those of the authors alone.* Address correspondence to Frederick Conrad; e-mail: fconrad@isr.umich.edu.

doi:10.1093/poq/nfp013 Advance Access publication April 2, 2009

drawn from cognitive interviews but the quality of the findings is limited by the content of the verbal reports.

Introduction

The survey enterprise rests on the assumption that respondents understand questions as intended and are able to answer them adequately. Pretests can expose respondents' problems interpreting questions and producing answers, and may point to potential solutions. Perhaps the most rapidly growing type of pretesting is cognitive interviewing (e.g., Lessler, Tourangeau, and Salter 1989; Willis, Royston, and Bercini 1991; Blair and Presser 1993; Willis, DeMaio, and Harris-Kojetin 1999; US Department of Commerce 2003; Willis 2005; Beatty and Willis 2007), a set of methods in which laboratory respondents are asked to provide verbal reports about their thinking while answering draft survey questions. This often involves the "think-aloud" method which has a long tradition in psychology (e.g., Ericsson and Simon 1993). By thinking aloud, respondents report on the mental processes used to produce an answer. Cognitive interviewers often probe for more extensive verbal reports than respondents spontaneously provide (Willis, DeMaio, and Harris-Kojetin 1999; Willis 2005; Beatty and Willis 2007). The resulting verbal reports (whether spontaneously produced think-alouds or responses to interviewers' probes) are, in effect, raw data that interviewers or other analysts listen to, interpret and otherwise examine for evidence of problems. Although cognitive interviews are usually carried out with only a small number of respondents, the results can point to potentially serious problems that would be certain to increase measurement error if not addressed.

Our concern in the current article is that even though cognitive interviewing is carried out to reduce overall measurement error, it involves its own type of measurement error. This error has at least two sources: (1) differences in how verbal reports are elicited may lead to variation in their quality and (2) the way the verbal report is interpreted may vary between judges so that a particular report may be taken to have different meanings. If so, the data are likely to be less reliable and less valid indicators of problems than one would want, given the method's impact on questionnaire development.

We report an experiment in which we measure the degree to which listeners (judges) agree about whether a particular verbal report reflects a response problem and, if there is agreement, the degree to which listeners concur about the type of problem. The cognitive interviewers conducted one of two types of interviews distinguished by the kinds of probes they were licensed to administer. One group of interviewers was instructed to probe about explicit evidence of problems in respondents' verbal reports whenever such evidence was present; the other was licensed to probe at their discretion which meant that they could explore the existence of problems they believed were present but for which

the current verbal report contained no explicit evidence. Our primary questions were (1) how much agreement is there overall about the content of particular verbal reports and (2) are agreement rates affected by the kind of probes that *elicit verbal reports*, specifically probes restricted to what the respondents have said versus probes that may reflect interviewers' hypotheses about the presence and nature of problems irrespective of overt evidence.

THEORY OF VERBAL REPORTS

In their landmark book, Ericsson and Simon (1993) provide a theoretical account of how verbal reports are produced and what such data can and cannot tell us about people's thinking. The key component of their theory is that people can only report accurately about thinking to which they have access, i.e., the mental states they store in working memory en route to completing the task. For example, when planning a move in chess, players anticipate a series of moves, temporarily storing each state of the chessboard in working memory; they are thus able to verbally report each state while it is stored. However, the thinking required for some tasks is not as accessible and so people have little or nothing to report. For example, when we recall something we know very well such as the password for an email account, the information just comes to mind. Schneider and Shiffrin (1977) refer to this as an *automatic process* and there is little that people can say about how they retrieve the information.

The same is true in cognitive interviews. Sometimes respondents have access to their thinking and sometimes they do not. For example, when answering a behavioral frequency question like "How many times have you consumed more than five alcoholic beverages in the last month?" the respondent may recall and count specific events, storing a description of the events and a running tally in working memory; this is exactly the kind of information that we should be able to report verbally and in fact think-alouds are often used to explore how people answer this type of question (e.g., Menon 1993; Bickart and Felcher 1996; Conrad, Brown, and Cashman 1998). However, many survey response tasks seem unlikely to create such reportable temporary memories. Consider how you would respond to a question to which you know the answer very well, like "Have you ever filed for bankruptcy?" The answer most likely just comes to mind and there is little you can say about the process beyond "I just know the answer." Similarly, respondents may not accurately report how they answer "why" questions (Wilson, LaFleur, and Anderson 1996) or how they answer questions soliciting a preference (Wilson and Schooler 1991) because the information on which they base their answer is not accessible. In sum, there is not a perfect fit between all survey response tasks and the think-aloud process (see Willis 2005, p. 55). Pressing respondents to produce verbal reports when they do not have access to their thinking may increase ambiguity of the reports or alter the process being reported, often referred to as "reactivity."

AMBIGUITY OF VERBAL REPORTS

There are several reasons why different listeners might interpret verbal reports differently. First, even verbally skilled respondents may not be able to articulate the kind of thinking involved in answering many survey questions if this is not accessible (Groves, Fultz, and Martin 1992). If an interviewer prompts the respondent to say something when relevant information is not available, a compliant respondent might say something vague, rather than saying nothing; the vaguer the statement the larger the number of possible interpretations, increasing the chances that some listeners might hear evidence of a problem in such reports and others might not. More detailed and concrete reports may well be more consistently interpreted.[1]

A second reason why verbal reports may be ambiguous concerns the interviewer's behavior, in particular the use of probes and prompts. To some listeners, the interviewer's probes and prompts may seem to affect what the respondent says, leading these listeners to dismiss apparent reports of problems as artifacts of the interviewers' behavior. To other listeners, such reports reflect thorough detective work by the interviewer and are credible indications of problems. For example, suppose the interviewer asks the respondent to paraphrase the question and, on the basis of the paraphrase, the respondent seems to misinterpret the question (cf. Groves, Fultz, and Martin 1992). Some listeners might hear legitimate evidence of a problem in what the respondent says while to others the report may sound as though the respondent is only having trouble paraphrasing and no trouble understanding the question.

A related potential source of ambiguity in verbal reports is the lack of probing. A respondents' verbal report may hint at the presence of problems but not be definitive. If the interviewer does not probe for more clarity about what the respondent has reported, different listeners may reach contradictory conclusions.

REACTIVE THINKING

The use of verbal report data to study cognition has been controversial in part because of concerns about *reactivity*, i.e., the possibility that thinking aloud can affect the process about which people are reporting (e.g., Russo, Johnson, and Stephens 1989; Schooler, Ohlsson, and Brooks 1993). In cognitive interviewing, some kinds of probes may react with the way respondents answer the question. For example, if an interviewer probes about a problem that he or she anticipated but did not actually observe (e.g., "Was the question clear?"), this could start the respondent thinking (aloud) about the ways in which the question might potentially be unclear, but from which he or she would not

1. Of course, a respondent who feels pressured to say something may fabricate a description of a problem. But unlike a vague example, this could well be consistently interpreted by different listeners if it is sufficiently concrete.

Table 1. Design of the Study

	Discretionary probe interviews	Conditional probe cognitive interviews
Conducting interviews	4 interviewers × 5 interviews = 20 interviews	4 interviewers × 5 interviews = 20 interviews
Interpreting verbal reports	Interviewers write narrative summary	Interviewers classify verbal reports into problem taxonomy
	2 transcribers map the narrative summaries into problem taxonomy	
	4 independent judges classify all verbal reports into problem taxonomy	

necessarily have suffered. Our main experimental manipulation was intended to vary the type of probes administered in different interviewing conditions to see if the resulting verbal reports lead to different conclusions about the presence or absence of problems.

Study Design

The current study was carried out to assess how the method of eliciting verbal reports might affect agreement among listeners and explore why judges disagree when they do. Eight cognitive interviewers each administered five interviews following one of two approaches to cognitive interviewing that differed primarily in the kind of probing behavior that was permissible. In one, experienced practitioners were free to explore potential problems as they saw fit and so we refer to this as *discretionary probe* cognitive interviewing. These interviewers were instructed to do what they ordinarily do on the assumption that this would involve a relatively high frequency of such probes. In the other variety of cognitive interviewing, interviewers' probing behavior was restricted to certain conditions, specifically situations in which respondents' verbal reports contained indications of problems. This meant these interviewers could not probe when the interviewers' intuitions or previous respondents' reports led them to suspect a problem. We refer to this as *conditional probe* cognitive interviewing.

The design of the study is presented in table 1. Four interviewers each conducted five discretionary probe interviews and four different interviewers each conducted five conditional probe interviews.[2] They administered a

2. The respondents were recruited by randomly sampling telephone numbers from local telephone directories in the College Park, MD area and contacting potential respondents until gender and

questionnaire of 49 questions borrowed from several draft questionnaires created at the Survey Research Center at the University of Maryland. Our basic unit of analysis is the question administration, i.e., each occasion on which one of the 49 questions is asked. Thus, there are 980 (i.e., 49 × 20) question administrations for each of the two types of cognitive interviews for a total of 1,960 observations. For each question administration, the interviewer indicated whether or not there was a problem.

In addition to the interviewers' problem evaluation, four independent judges listened to all of the recorded interviews and for each question the judges classified any problems they detected. Thus, we constructed a second data set from the judges' problem assessments. Each judge provided 980 problem assessments for each type of cognitive interviewing for a total of 7,840 (i.e., 4 × 980 × 2) observations.

DISCRETIONARY PROBES

We asked the experienced practitioners to conduct cognitive interviews as they ordinarily do under the assumption that at least some of the time they would probe about problems that they anticipated respondents might experience, based either on their expert assessment of a survey item or their experience in previous interviews, but not necessarily on the basis of the current respondents' verbal reports. These probes were, thus, administered at the interviewers' discretion. We allowed these interviewers to define "problem" as they saw fit and asked them to prepare written reports of the problems they detected for each question in each interview. We asked these interviewers to prepare written reports because they indicated this is their standard practice. What is probably not standard is to document each administration of each question; we requested reports in this format to enable analyses of individual verbal reports.

According to their written reports and a subsequent debriefing, all of these interviewers reviewed the questionnaire and prepared some probes prior to conducting any interviews. They reported that in the interviews they used a combination of planned and improvised probes to explore potential problems. Two of the four indicated that, over the course of the interviews, they were less likely to probe about problems they had already encountered than about unexplored problems. Despite our use of a single label, we recognize that what we call discretionary probing may actually differ to some extent across the four interviewers.

CONDITIONAL PROBES

We instructed the interviewers in this group to solicit concurrent verbal reports from respondents and then to focus their probing on behavioral evidence of

---

education quotas were met for the two experimental treatments (interviewing techniques). This led to comparable groups of respondents in the two treatments.

Table 2. Conditional Probes

C	Respondent cannot answer (possibly because the task is too difficult) or does not know the answer (when it would be reasonable to know it); respondent does not provide a protocol.
P	"What was going through your mind as you tried to answer the question?"
C	Answer after a period of silence.
P	"You took a little while to answer that question. What were you thinking about?"
C	Answer with uncertainty; this can include explicit statements of uncertainty or implicit markers such as frequent use of "um" and "ah," changing an answer, etc.
P	"It sounds like the question may be a little difficult. If so, can you tell me why?" "What occurred to you that caused you change your answer?" "You emphasized/ or you repeated [word]. Why was that?"
C	Answer contingent on certain conditions being met ("I'd say about 25 times if you don't need a super precise answer.")
P	You seem a little unsure. Was there something unclear about the question?
C	Erroneous answer; verbal report indicates misconception or inappropriate response process
P	Clarify respondent's understanding of particular term or the process respondent uses. Suppose the respondent's report suggests she misunderstood the word "manage." Probe this term. "So you don't manage any staff?"
C	Respondent requests information initially instead of providing an answer
P	"If I weren't available or able to answer, what would you decide it means?" Are there different things you think it might mean?" If yes: "What sorts of things?"

NOTE.—C stands for "condition," P for "probe."

problems in those reports, i.e., evidence in respondents' speech and actions. They were instructed to probe only when the respondents' verbal reports corresponded to a generic pattern indicating a potential problem (e.g., an explicit statement of difficulty, or indirect indications such as a prolonged silence or disfluent speech). The conditions under which probing was permissible are displayed in table 2. When such a condition was met, interviewers were instructed to probe by describing the respondent behavior that suggested the possibility of a problem (e.g., "You took some time to answer; can you tell me why?").[3] Other than probing under these conditions, the interviewers were instructed not to play an active role.

The rationale for this restricted probing was to (1) assure that respondents' verbal reports reflected what is in their working memories and (2) reduce the chances of reactive effects. Only if the respondent had spontaneously

3. These interviewers were unconstrained in their choice of words and could converse with the respondent about the possibility of problems much as when cognitive interviewers administer *expansive probes* described by Beatty (Beatty, Schechter, and Whitaker 1997; Beatty 2004). Because these interviewers were not specifically instructed to conversationally expand their probes, we cannot evaluate the benefits of more versus less expansive probing.

reported her thinking can we be sure it is accessible to her, i.e., available in her working memory. If interviewers were to ask for more details when none are available, this could lead to reports that are rationalized, embellished, or even fabricated; the probes themselves might lead respondents to believe they have experienced a problem, or to make a comment that the interviewer interprets as indicating a problem. By remaining silent (not probing) in the absence of verbal evidence of a problem, interviewers are less likely to produce reactive effects.

This approach to probing is relatively uncommon in practice. If it were to be taught to experienced cognitive interviewers they would have to temporarily "unlearn" what they ordinarily do. We chose instead to teach the technique to relatively novice cognitive interviewers. A two-day training session covered both how to conduct the interviews and what qualifies as evidence of problems. The interviewers practiced the technique in mock interviews with each other. The instructor (one of the authors) provided feedback to the interviewers and determined when all four had grasped the probing technique.

The conditional probe interviewers directly entered problem codes into a database via a coding tool after each interview. They entered problems into a problem taxonomy developed by Conrad, Blair, and Tracy 1999. Coded problems were essential for most of the analyses reported below.

QUESTIONNAIRE

The topics covered by the 49 questions included nutrition, health care, AIDS, general social issues, and computer use. Twenty-five of the questions concerned the respondents' behaviors and 24 concerned their opinions.

INTERVIEWERS

The four discretionary probe interviewers each had more than five years of experience conducting cognitive interviews at different organizations within the Federal government and private sector survey research communities. Three of the four had doctoral degrees in psychology. This level of education seems to us to be typical of experienced practitioners of cognitive interviewing (cf. Groves et al. 2004, p. 246).

The four conditional probe interviewers were less experienced with pretesting questionnaires though all four worked in survey organizations in either the academic, commercial, or government sectors. Two of the four had some brief experience with production cognitive interviewing and the other two had been exposed to the theory behind the method and had completed relevant class exercises in a graduate level survey methodology course. In contrast to the discretionary probe interviewers, three of whom held doctoral degrees, three of

these interviewers held only bachelors degrees (though were pursuing masters degrees) and one held a masters degree.[4]

All eight cognitive interviewers were told that they were participating in a methodological study sponsored by a government agency.

DATA FROM THE COGNITIVE INTERVIEWS

In order to measure agreement between all of the interviewers and judges, the narrative reports created by the discretionary probe interviewers were classified into the same problem taxonomy into which the conditional probe interviewers directly entered the problems they had detected, thus providing comparable problem descriptions. The classification of problems in the narrative was essentially a transcription task. Two transcribers independently mapped the written problem reports to the problem taxonomy and then worked out any discrepancies together. Both transcribers had been introduced to cognitive interviewing in a graduate survey methodology course but neither had conducted cognitive interviews. They were given written definitions of the problem categories and a training session in use of the taxonomy that included practice coding until the authors judged them to be competent users of the taxonomy.[5]

It is possible that because the conditional probe interviewers were acquainted with this coding scheme when conducting their interviews, they could have identified different types of problems than the discretionary probe interviewers who were not familiar with the codes. In fact there was no difference in the proportions of different types of problems reported by the two groups (Conrad and Blair 2004).

In addition to the interviewers' own conclusions about the presence of problems, four independent judges coded the presence of problems in all 40 interviews using the same problem taxonomy. The judges had all been introduced to cognitive interviewing in graduate courses in survey methodology and, in one case, through employment in a government statistical agency.[6] The judges participated in the same training on the definition of problems (not on the interviewing technique) as did the conditional probe interviewers and

4. We considered crossing the factors of technique and experience/education so that both experienced and inexperienced interviewers would use both methods. However, this would have led to intractable problems. For example, even if experienced cognitive interviewers could be trained in the conditional probe technique, there would be no way to equalize the experience that they have with the two techniques.

5. The exact rationale for this problem taxonomy (see Conrad and Blair 1996, for a discussion) was not central to its use in the current study. Other problem taxonomies certainly exist (e.g., Forsyth, Lessler, and Hubbard 1992; Willis and Lessler 1999) and would have been appropriate here. The point is that one needs some set of problem categories in order to tally the problems. For example, one cannot count two verbal reports as illustrating the same problem without somehow categorizing those reports.

6. They were not questionnaire design experts, in contrast to some of the interviewers. There was no evidence, however, that this affected the results (discussed in the Results section).

had also been exposed to cognitive interviewing in the same graduate courses. By registering all problems in a single problem taxonomy, it was possible to measure agreement between each interviewer and each of the four judges as well as agreement between each pair of judges.

Results

DESCRIPTION OF THE DATA

Over the 40 cognitive interviews, interviewers administered questions on 1,960 occasions. Interviewers reported potential problems in 519 of these question administrations. The conditional probe interviewers detected 216 problems and the discretionary probe interviews detected 303 problems.

DIFFERENCES BETWEEN INTERVIEWING GROUPS

The two groups of interviewers behaved quite differently from one another, much as expected. Recall that the discretionary probe interviewers were instructed to use their ordinary technique but not specifically to use discretionary probes. We adopted this approach under the assumption that these interviewers would administer a large proportion of discretionary probes if left to their own devices. This was in fact the case.[7] Of the 987 probes that these interviewers administered, 41% were "context-free" in that they were used to explore problems that had not been explicitly indicated in the preceding discourse context. This is essentially what Willis (2005) refers to as "proactive probes." For example, if interviewers ask respondents what a particular word means or to paraphrase the survey question when there is no overt indication that respondents have misunderstood particular words or the overall intent of the question, such a probe is context-free. Such context-free probes can be formulated on the basis of the interviewer's assessment (e.g., while reviewing the questionnaire prior to conducting any interviews) or experience in previous interviews (e.g., a previous respondent exhibited difficulty with the question). Although such probes are not based on the current interview context, there may well be sound reasons for their administration. In the interview excerpt in online Appendix A, the interviewer probes in lines 15, 22–23, and 29–30 about problems for which there was no immediate evidence.

We refer to a second type of probe as "context-based" because these probes are developed and administered in reaction to something that the current respondent has uttered—or has not uttered in the case of pauses—in the process of answering the question. These probes, essentially, are requests for

7. Two coders jointly labeled each conversational turn in all of the interviews. The turn labels included prompts to keep talking, probes, context (or conditions) for probes, and explicit statements of problems. The two coders reached agreement on all assigned codes.

the respondent to elaborate about a problem that was possibly indicated in the immediately preceding discourse. An illustration appears in online Appendix B. In this exchange from a discretionary probe interview, the interviewer administers several context-based probes, lines 11–12. The interviewer probes about a potential indication of a problem in the respondent's speech ("I noticed you paused there for a second after I, I asked you that. Was there something unclear in that question?"). Such probes may be generic (e.g., "Can you tell me more about that?") or question-specific (e.g., "Why do you think your answer could be either 'somewhat urgent' or 'extremely urgent'"?). Our notion of context-based probes corresponds closely to Willis's (2005) reactive probes.

Of all the probes administered by discretionary probe interviewers, only 13% (128) were context-based. In contrast, the conditional probe interviewers primarily administered probes of this form, much as instructed. Although they administered only 236 probes overall, 61% (144) were based on context. The remainder consisted of context-free probes (5%) and generic prompts (34%) such as "Please keep talking" and "How did you come up with that answer?". Thus, the two interviewing groups differed in both the number of probes they administered (the discretionary probe group administered 4.2 times as many probes as the conditional probe group) and the types of probes they administered (discretionary probe interviewers posed more context-free probes than any other kind while conditional probe interviewers asked more context-based probes than any other kind).[8]

We now turn to several analyses that allow us to evaluate our primary questions about agreement in interpreting respondents' verbal reports in cognitive interviews. The first analysis is concerned with agreement between pairs of listeners and how the type of probes that interviewers administer might affect agreement. The second set of results focuses on those cases where the judges disagree with one another; we ask whether this disagreement occurs because the judges interpret the same words differently or because they focus on different parts of the respondent's verbal report. The third set of results concerns the accuracy of interviewers' problem detection treating the majority opinion of judges as a validation standard. In the fourth and final analysis, we ask whether across multiple interviews it is clear which questions require revision,

8. Note that this pattern is somewhat at odds with the pattern reported by Beatty (2004) and Beatty, Schechter and Whitaker (1997). In Beatty's study of probing by three cognitive interviewers, the probe category that most closely corresponded to our context-free probes ("Traditional Cognitive Probes") was the least frequently used; the categories "Confirmatory" and "Expansive" probes that are similar in spirit to our context-based probes were the most frequently used. The interviewers were likely to be more similar to our discretionary probe than conditional probe interviewers so this pattern reverses what we observed. It is difficult to know what is responsible for these differences across studies and this warrants further investigation. However, for current purposes, we are more concerned with how the type of probe affects agreement rates than in the generality of the pattern.

even when there is disagreement about the implications of particular verbal reports.

Agreement about the content of verbal reports: Our primary concern is the way in which the interviewer's behavior might affect agreement about the implications of individual verbal reports and so we focus on differences between the two interviewing groups. However, to provide context for such analyses, we first examine overall agreement, irrespective of the technique used by any one cognitive interviewer. The main finding is that agreement is not strong. The average kappa score across interviewer–judge pairs for decisions about the presence or absence of a problem was only .30 (fair agreement[9]). In those cases where there was agreement that a problem was present the agreement about the type of problem (based on the Conrad, Blair, and Tracy taxonomy) was $\kappa =$ .46 (moderate). While this is reliably higher than agreement about the presence of a problem ($t[26] = 3.058$, $p < .005$) it is still disturbingly low considering that it is based on pairs of judgments about the same verbal reports. Agreement was equally low when computed between pairs of judges. The average kappa for the overall problem decision, i.e., whether or not there was a problem, was .32 (fair). The average kappa score was also fair (.40) for decisions about the particular problem category within the taxonomy. At the very least, these low agreement scores suggest that verbal reports—the raw data from cognitive interviews—may be more ambiguous than is commonly assumed.

It is possible that some verbal reports are particularly likely to lead to disagreement. One hypothesis is that the way verbal reports are elicited affects the range of possible interpretations. More specifically, if the interviewer probes in a way that may seem to focus the respondent on a potential problem, then a subsequent verbal report that indicates the presence of a problem may be dismissed by some listeners while others might accept it as evidence of a problem. In contrast, other probes may lead to less ambiguous verbal reports. To examine this, we compare agreement for verbal reports elicited by discretionary probe and conditional probe interviewers whose probing techniques differed in much this way.

Agreement did in fact differ between the two cognitive interviewing techniques. Average kappa scores for interviewer–judge pairs were reliably higher for conditional probe interviews (.38) than for discretionary probe interviews

9. For the evaluation of observed kappa values, we use the following categories presented in Everitt and Haye (1992), p. 50:

Kappa	Strength of agreement
.00	Poor
.01–.20	Slight
.21–.40	Fair
.41–.60	Moderate
.61–.80	Substantial
.81–.00	Almost perfect

(.25), $t\,[26] = -3.68$, $p < .001$. While both kappa scores indicate "fair" agreement, there seems to be something about the verbal reports elicited with the conditional probe method that makes it somewhat easier for listeners to agree whether or not they contain evidence of problems. The pattern for inter-judge (as opposed to interviewer-judge) agreement was similar (average $\kappa = .36$ for conditional probe and .27 for discretionary probe interviews) but the difference was only marginally significant ($t\,[10] = -1.97$, $p < .07$).

To confirm that the difference in agreement for the interviews conducted by the two groups of interviewers was in fact due to differences in probes, we partitioned the question administrations into those that contain context-based probes and those that contained context-free probes within each interviewing group. We then selected all cases for which at least one judge detected a problem and computed the proportion of these for which two or more judges detected a problem, i.e., on which there was agreement. These proportions are higher for context-based than context-free probes for conditional probe interviews, .57 versus .35 ($\chi^2\,[1] = 9.74$, $p = .002$) and for discretionary probe interviewers (though not significantly), .59 versus .51 ($\chi^2\,[1] = 1.29$, n.s.). Because the type of probe on which agreement is higher (context-based) comprises a higher proportion of all probes in the conditional probe than discretionary probe interviews, this pattern could account for the higher overall kappa scores in the former than in the latter interviews. Apparently when context-based probes uncover potential problems, they are more uniformly compelling to judges (though far from perfect) than are context-free probes.

Why might this be? We propose that it is not just the probes but the respondent's reaction to them that contributes to different levels of agreement under the two types of cognitive interview. When interviewers probe about a particular respondent utterance in order to determine if it indicates a problem, the respondent's reply to the probe should lead to a relatively clear-cut problem judgment because the respondent can clarify whether his or her initial utterance did or did not suggest a problem. However, probes that are *not* clearly tied to something the respondent has already said may produce less definitive results. Suppose the interviewer asks the respondent what a particular term means even though the respondent has not indicated any confusion. If the respondent's answer to this probe indicates possible confusion, this may be hard for listeners to evaluate. Has the interviewer uncovered an actual problem or introduced one? Different listeners may hear such an exchange differently, which would lead to low agreement.

Consider the following example[10] of a context-based probe followed by a description of a problem. The respondent indicates a possible problem in lines

10. 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 ("that-") indicates that the word was cut off. Question marks indicate rising intonation, and utterance-final periods indicate falling or flat intonation, regardless of whether the utterance is a question or an assertion.

8–9 and 13–14, namely that he hasn't heard of any drug use in the neighborhood so he doesn't know how to answer. The interviewer then probes (line 15) for more information about why the respondent does not know how to answer. The respondent answers, providing a description of the problem, in lines 16–18. Here he indicates that because he has not heard of any drug abuse in the neighborhood he does not know what kind of answer the interviewer is looking for—whether it would be appropriate to answer "don't know" or "not an urgent problem."

01 I: For each problem listed below, please indicate
02 to me whether it is a very urgent problem, a
03 somewhat urgent problem, a small problem,
04 or not a problem at all in your neighborhood.
05 Drug abuse.
06 R: Mm.
07 I: Tell me what you're thinking.
08 R: Well, I haven't heard of any. Well, I haven't
09 heard of any drug use in my neighborhood.
10 I: Okay.
11 R: Um. So um.
12 I: Okay.
13 R: In my neighborhood. Um so um. I don't know
14 how to answer that. I don't know.
15 I: Why don't you know how to answer it?
16 R: Just because since I: don't hear of ANY I
17 don't know like what answer you're looking for
18 because
19 I: What-whatever it means to you. Again the
20 question for each problem listed below please
21 indicate to me whether it is a very urgent problem,
22 a somewhat urgent problem, a small problem or not at
23 problem at all in your neighborhood and the first one.
24 R: Oh it's a very urgent problem.

In contrast, consider this example of a context-free probe about meaning followed by a problem description. The respondent provides an adequate answer (line 5) without evidence of a problem. The interviewer then asks the respondent to paraphrase the question (lines 6–7). This is followed by the respondent's description of a possible problem (lines 8–12), i.e., confusion and uncertainty about how to answer.

01 I: Okay would you say that getting AIDS is an
02 extremely serious threat to a person's health,

03 a very serious threat, somewhat serious or
04 not too serious?
05 R: Um . extremely serious.
06 I: Okay and what were you what do you think
07 that question was asking?
08 R: Um I don't know. I'm kind of confused.
09 Maybe extremely or not extremely. Like
10 depending on like the size of the person or
11 like the health condition they're in. Like how
12 healthy they are. Like their age.

The question is whether problem descriptions of the first sort, lines 16–18 after the context-based probe are more convincing than problem descriptions of the second kind, i.e., lines 8–12 after a context-free probe. To explore this possibility, we computed the proportion of question administrations involving these two interaction patterns in which at least one judge identified a problem.

Looking first at the discretionary probe interviews, 79.3% (23 out of 29) of the question administrations in which a respondent problem description is elicited by a context-based probe were considered by at least one judge to reflect a problem. In contrast, only 53.8% (14 out of 26) of the administrations in which a respondent problem description is elicited by a context-free probe were judged to indicate the presence of a problem (χ^2 [1] = 5.91, $p < .05$). (Note that there were relatively few of the latter type of interactions because, while context-free probes were frequent in discretionary probe interviews, they were often not followed by any indication of a problem.) The conditional probe interviews show a similar pattern but the difference is not reliable: in 84.4% (27 out of 32) of the question administrations in which the respondent's problem description was preceded by a context-based probe, at least one judge concluded there was a problem while in 80% (4 out of 5) of the question administrations in which the respondent's problem description was preceded by a context-free probe, at least one judge perceived a problem (χ^2 [1] < 1, n.s.).

It is possible that respondents who believe they have provided an adequate answer and are then asked what they think the question means interpret the probe as an indication that they have not answered adequately. This could have been the case in lines 8 and 9 in the last example. By following Grice's (1975) cooperative principle, which enables listeners to infer certain of the speaker's intentions in everyday conversation, respondents might reasonably assume that the interviewer would only have asked the question about meaning if she had reason to doubt the respondent's interpretation. Thus, it would be reasonable for the respondent to revise her interpretation from what she originally thought, which could account for the confusion expressed in lines 11–15. To some listeners (judges) such a reactive process may be at work while to others no such distortion has occurred. Whatever the exact mechanism, the current data

provide clear evidence—albeit based on a small sample—that respondents' reports of problems are differentially persuasive to listeners when they follow different types of probes.

Sources of disagreement: While the two interviewing techniques lead to different levels of disagreement, both produce more disagreement than one would expect. While any disagreement is troubling, some types of disagreement may be of greater concern than others. More specifically, it would be disturbing if a pair of listeners reached different conclusions about the presence or category of a problem in exactly the same respondent utterances. This would speak to an inherent ambiguity in verbal reports in cognitive interviews—at least those obtained with the cognitive interviewing techniques that we investigated—and would raise serious doubts about relying solely on such data for improving questionnaires. It would be less worrisome if a pair of listeners disagreed about whether there is a problem or what kind of problem because they focus on different parts of respondents' verbal reports. One listener might home in on the first few words and another may give more weight to the latter part of the respondent's utterance. By this view, discrepancies in problem judgments result because listeners analyze different parts of the verbal report, i.e., different data.

To examine this further, we focused our analysis on agreement between judges about the type of problem. When the judges classified a problem into the problem taxonomy, they were also required to list the evidence on which they based the problem decision, i.e., to enter the relevant part of the verbal report, either verbatim or in their own words. This provided data about which words and paralinguistic information in respondents' verbal report contributed both to the judges' problem–no problem and problem type decisions, and makes it possible to assess whether pairs of judges considered the same or different parts of the verbal report. We selected a simple random sample of approximately 20% of eligible question administrations for each of the interviewing techniques for a total of 133 cases. We cross-tabulated the assignment of problem codes (same or different) and citation of evidence (same or different) by pairs of judges for all question administrations in the sample in which at least one pair of judges determined there was a problem. It was self-evident whether problem codes were the same or different. However, classifying the cited evidence as "same" or "different" required that we determine if the gist of the evidence cited by pairs of judges was the same or different. This is because judges were not required to literally transcribe the evidence and so could refer to the same verbal content using different words.

A coder compared the evidence cited by pairs of judges in all the sampled question administrations. The coder was instructed to treat two pieces of evidence as the same if (1) the "topic" of the evidence was the same, (2) the substance of the evidence was the same, and (3) the actor (interviewer or respondent) was the same. If the comparison failed to meet any of these three

criteria, then the two pieces of evidence were classified as different. For example, if both pieces of evidence concerned the reference period of the question, the topic was the same, but if one concerned the length of the reference period and the other concerned unfamiliar terminology, the topic was the same but the substance of the evidence was different. If both pieces of evidence noted the respondents' inability to recall events from such an old reference period, the substance of the problem was the same. If, however, one piece of evidence described the respondent's failure to notice the dates bounding the reference period and the other noted that the reference period was too long for accurate retrieval, the substance of the evidence differed. Finally, if one piece of evidence described a problem based on something the respondent said and the other piece was an interviewers' description of a problem, the actors were different. The coder could refer to a transcript of the question administration to help pinpoint which segment of the verbal report the judges were referring to in the evidence they listed. Agreement was scored liberally so that if one judge listed more than one piece of evidence, the evidence listed by another judge could only mismatch one time.

A question administration was eligible for this analysis if at least two judges decided it involved a problem. First, we focus on the agreement patterns irrespective of the interviewing technique. In the majority of cases (63.2%), the judges agreed about the type of problem and cited the same evidence in support of those problems (χ^2 [1] = 57.59, $p < .000$).[11] Our focus, however, is on the cases where the judges disagreed. In those cases they most often (22.6%) cited different parts of the respondent's report as evidence of the problem.

In the interview excerpt presented in online Appendix C, one judge identified a "temporal" problem while the other identified a "memory" problem. The judge who identified the temporal problem cited as evidence an utterance about the large number of dates to search through (line 21). The judge who identified the memory problem cited an utterance whose gist was that the number of categories was too large to be easily kept in mind (line 25). We cannot know if this disagreement came about because the judges simply did not attend to the line in which their counterpart identified a problem or because they attended to these lines but concluded they did not indicate a problem. The important point is that their disagreement was due to their focus on different pieces of evidence.

The judges considered the same evidence to indicate different problems in only a small percent (7.5%) of question administrations. While it is encouraging that the same words are rarely interpreted differently, it is troubling that there was *any* disagreement about the nature of problems indicated by the same words. This serves as a reminder that while verbal reports are certainly data,

11. It may seem counter-intuitive that judges agreed on problem type in 70% of the question administrations yet produced low kappa scores for this judgment in the full data set. This just reflects the way kappa is calculated. Because it adjusts for chance agreement, kappa can be quite low even when percent agreement is relatively high.

they are subject to more variable evaluation than are quantitative data. Finally, in a small percentage of cases (6.8%), judges interpreted different evidence as indicating the same type of problem. This type of disagreement could be relatively benign in practice where only one judge (usually the interviewer) analyzes the verbal reports: if two judges have actually identified the identical problem—but it is just evident in different conversational turns—repairing the question based on either judge's observation would lead to the same revised question. However, it is also possible that the two judges have identified different problems which just happen to be of the same type, e.g., both involve the respondent's lack of familiarity with a technical term, but the two judges have observed this type of problem for different terms. In this case, repairing the problem found by one judge would leave the other problem intact.

Disagreement and validity. Whatever the reason two listeners disagree about the presence of a problem, they cannot both be correct with respect to some validation standard. Thus, the erroneous judgment either "detects" a problem that is not actually present—a *False Alarm*—or overlooks a problem that is present—a *Miss*; accurate judgments detect a problem that is in fact present—a *Hit*—or determine there is no problem when in fact there is none—a *Correct Rejection*. (We have borrowed these analytic categories from signal detection theory, e.g., Green and Swets 1966.) It is conceivable that because of the prevalence of context-free probes in the discretionary probe interviews these interviewers produced more False Alarms than did those using the conditional probe technique. When cognitive interviewers administer context-free probes, they seem to be inquiring about a particular problem, possibly encountered in a previous interview. As a result they may be more inclined to interpret the subsequent verbal report as confirming the problem than when they do not probe in this way. Such a tendency could be related to a generally lower threshold for what counts as a problem and thus greater sensitivity to problems. If so, discretionary probe interviewers should rarely miss actual problems and should "false alarm" relatively often. Alternatively, the hypothesis-driven approach behind context-free probes could increase an interviewer's sensitivity to those problems he or she anticipates, leading to a relatively high proportion of hits, and reduce the interviewer's sensitivity to those problems he or she does not expect, leading to a relatively high miss rate.

To explore this, we derived a validation standard for each question administration based on the majority opinion of the four judges: if three or more judges agreed that a verbal report reflected a problem, we considered an interviewer to be correct when he or she judged there to be a problem in that report (a Hit) and incorrect when he or she did not (a Miss); note that Hits and Misses are complementary in that they jointly account for all of the cases in which a problem was indicated by the standard. When fewer than three judges determined that a problem was evident, we considered an interviewer to be correct when he or she concluded that the verbal report displayed no evidence of problems

Table 3. Percent Correct Rejections, False Alarms, Misses, and Hits for Discretionary Probe Interviews and Conditional Probe Interviews

	Correct rejection	False alarm	Miss	Hit
Discretionary probe	71.7	28.3	23.1	76.9
	(665)	(263)	(12)	(40)
Conditional probe	81.4	18.6	15.7	84.3
	(756)	(173)	(8)	(43)
All interviews	76.5	23.5	19.4	80.6
	(1421)	(436)	(20)	(83)

NOTE.—Number of question administrations within parentheses.

(a Correct Rejection) and incorrect when he or she judged there to be a problem (a False Alarm); note also that Correct Rejections and False Alarms are complementary in that they jointly account for all of the cases in which the standard indicated there was no problem. This is a "validation standard" with respect to each question administration in our study, not with respect to actual problems found in subsequent field administration.[12] Because the criterion for the presence of a problem within a question administration is quite stringent (a majority of judges), there are relatively few cases for which the standard indicated a problem and therefore relatively few Hits and Misses.[13]

The signal detection rates are presented in table 3. Discretionary probe interviewers produced nearly 10% more false alarms than did the conditional probe interviewers (28.34 versus 18.62%), $t(979) = 5.11$, $p < .001$. This is consistent with what is expected given the differences in probing behavior. In contrast, there is no reliable difference in the Miss rates for the two groups of interviewers ($t(979) < 1$, n.s.); although if there is any difference (and more power would be required to detect it), it appears to be in the direction of more Misses for the discretionary than conditional probe cognitive interviewers. This is certainly at odds with an increased sensitivity explanation and suggests that discretionary probe interviewers may be somewhat blinded to the presence of unanticipated problems because they tend to focus their search on problems they expect to find.

12. At least three studies (Willis and Schechter 1997; Fowler 2004; Blair et al. 2007) have been conducted that assess the validity of cognitive interview results in production interviews. All three studies found that in production interviews respondents did experience some of the problems anticipated by the cognitive interviews and in two of the studies (Willis and Schechter 1997; Fowler 2004), questions that were revised on the basis of the pretest led to lower levels of problems than did the originally worded questions.

13. If there was no data from a coder for a particular question administration, the validation criterion was two out of three coders' identification of a problem. Data were never missing from more than one coder.

Problems across interviews: The focus of our analyses to this point has been the quality of information in respondents' individual verbal reports produced by individual question administrations. But these data can also be analyzed at a higher level in which the focus is not the individual question administration but the survey question, taking into account all verbal reports elicited by each question. This level of analysis is particularly relevant to the practical rationale for cognitive interviewing, namely to identify questions requiring revision.

The results of the first level of analysis indicate that respondents' verbal reports are sometimes ambiguous with respect to the presence of problems. Yet over multiple verbal reports, i.e., reports from several interviews, different judges might reach roughly the same conclusions about particular questions' need for revision even if the judges disagree about the implications of specific individual verbal reports. The reasoning is that over a set of interviews, seriously flawed questions should produce more evidence of problems than questions without flaws. While any one piece of evidence may be inconclusive, the body of evidence may convince the judges that problems exist. In contrast, if another question does not elicit evidence of problems across the interviews, the judges are unlikely to conclude that it needs repair.

For this second level of analysis, we computed a "problem score"—the number of problems detected—for each question for each judge and compared those scores across judges. The problem score was calculated for each question by summing the problems detected by each of the four judges across all 40 cognitive interviews and separately for the 20 discretionary probe and the 20 conditional probe interviews. The problem score is not itself an aggregate judgment about which questions require revision but a measure of the information (sum of problems found in individual verbal reports) likely to inform an aggregate judgment; it serves here as a proxy for the judgment that might occur about particular questions in an actual pretest.

Because the problem scores are simple sums, they are larger for the 40 interviews (median problem score = 21.91) than for the 20 discretionary probe (median problem score = 13.89) or the 20 conditional probe interviews (median problem score = 10.63). High agreement at the question level would take the form of highly correlated problem scores across judges. The average correlation of problem scores for pairs of judges over the 40 interviews is .71, $p < .01$. The judges agree reasonably well about which questions most need revision, although their agreement for the full set of questions is still not high, especially considering that their judgments are based on far more observations than is typical in production cognitive interviewing.

When we examine problem scores produced by the two interviewing techniques, the correlations are lower, .54, $p < .01$ and .66, $p < .01$ for discretionary probe and conditional probe interviews, respectively. At this level of analysis, the two techniques identify many of the same questions as candidates for repair: 7 of the 10 items with the highest problem scores in the discretionary probe

interviews were also among the top 10 in the conditional probe interviews. However, this also means that the particular cognitive interviewing technique led to different conclusions about which questions most needed revision 30% of the time. This is a fairly high discrepancy rate considering that these were the questions with the *highest* problem scores by one method.[14]

Note that the problem score correlations increase as the number of interviews increases from 20 to 40. Presumably, the more verbal reports collected the less important any one of them becomes in the judges' decision about which items need repair. In a study conducted to specifically examine the impact of sample size on cognitive interview pretest findings, we (Blair et al. 2006) found support for the observation that as sample size increases so does the number of problems detected.

This set of findings is mixed news for cognitive interviewing practitioners. While it suggests that many of the same questions are likely to be identified as most needing revision irrespective of exactly how the interviews are conducted and who listens to the interviews, the agreement is not especially high, despite unusually large sample sizes. On the positive side, increasing the sample size does improve agreement. Moreover, in practice there are usually layers of discussion and decision made prior to actual question revision; cognitive interview results contribute to the conversation but are usually not the entire story.[15]

Discussion

Cognitive interviewing can provide a window onto respondents' thinking and thus the problems they encounter when answering survey questions. While we have argued that the results of cognitive interviews are not always definitive, there is no doubt that cognitive interviewing reveals legitimate problems often enough to add value to the survey development process. Our concern here is that as cognitive interviews are sometimes used in practice, they may (1) flag "problems" that really are not problems and (2) identify different problems depending on who interprets the verbal reports. Some minor revisions to cognitive interviewing procedures may reduce these concerns—we discuss two such modifications below. However, the strong implication of the current study is that there is a tradeoff between the number of apparent problems, both real and spurious, that are detected and the precision of the problem detection process, i.e., the percentage of valid problems. Leaving interviewers to probe

14. Even though the two cognitive interviewing methods did not select exactly the same questions for revision, the types of problems contributing to the 10 highest problem scores were similar across the methods: overwhelmingly, the problems involved various types of comprehension difficulties, consistent with other research findings about cognitive interview-identified problems (Presser and Blair 1994; Conrad and Blair 1996). This result emerges even though there is some variation in the cognitive interview techniques used across these studies.

15. The exact relation between cognitive interviewing results and questionnaire revision is worthy of study (see Conrad and Blair 2004) but beyond the scope of the current paper.

at their discretion results in a larger number of problems being identified at the cost of more false alarms; constraining interviewer probing produces fewer false alarms but detects fewer actual problems. To the extent that researchers modify questionnaires based on spurious problem reports—false alarms about the presence of a problem—they have been ill served by the pretest results. On those occasions when researchers modify questionnaires in the absence of actual problems, this can be costly for obvious reasons.

If different interviewers or analysts listening to recorded interviews detect different problems, any one interviewer may miss legitimate problems that another interviewer might detect. Just as answers in production survey interviews should be independent of which interviewer asks the questions, so the detection of problems in cognitive interviewing should be independent of who interprets the verbal reports. Although it cannot correct the fundamental problem of ambiguity in verbal reports, the use of multiple reviewers who work together to reach consensus on problem identification may to some extent mitigate the weakness in cognitive interview data. This does not obviate the fundamental need to improve the individual verbal reports on which all else rests. However, beyond improving the analysis of individual verbal reports (increasing agreement and weeding out false alarms), our question level results suggest that increasing the number of reports for each question can improve identification of questions that need revision. It is probably wise to strive for both improved verbal reports and larger numbers of cognitive interviews.

Because interviewers' behavior can affect the quality of verbal reports (for better or worse), users of cognitive interviewing must carefully consider what behaviors to promote and what behaviors to discourage. To this end, constrained probing behavior might be worth exploring. In the current study, relatively unconstrained context-free probing turned up more potential problems but lowered agreement among listeners while the relatively constrained context-based probing turned up fewer but more reliably identified problems. If one desires more confidence in the results of cognitive interviewing, then increasing the proportion of context-based probes would be desirable. On the other hand, it is possible that very experienced cognitive interviewers may turn up subtle but important problems if they are free to probe when there is no overt indication of a problem, so there could be a cost to prohibiting all such probing. This debate will have to play out in subsequent research.

A related tension exists between exploration and confirmation. A cognitive interview can be an opportunity to go beyond one's intuitions, enabling the discovery of unanticipated problems. Alternatively it can be a test bed for one's hypotheses about the presence of problems, enabling the empirical evaluation of one's suspicions about the pitfalls of a questionnaire. But, it may not often be feasible for both. The purpose of cognitive interviewing is likely to differ at different stages of questionnaire development. With an early draft of a questionnaire, the objective is generally to identify all potential problems even at the risk of some false alarms. In the final stages of questionnaire development,

the objective usually shifts to one of confirming that repairs have been effective, in which case a higher criterion for what counts as a problem seems appropriate.

All measurement activities involve some error and cognitive interviewing is no different. The goal for practitioners, based on the current results, is to design cognitive interviewing procedures that maximize problem detection while minimizing the error inherent in the technique.

Supplementary Data

Supplementary data are available online at http://poq.oxfordjournals.org/.

References

Beatty, Paul. 2004. "The Dynamics of Cognitive Interviewing." In *Methods for Testing and Evaluating Survey Questionnaires*, eds. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer, pp. 45–66, New York: Wiley.

Beatty, Paul, Susan Schechter, and Karen Whitaker. 1997. "Variation in Cognitive Interviewer Behavior: Extent and Consequences." Proceedings of the Section on Survey Research Methods, American Statistical Association, 1064–8.

Beatty, Paul C., and Gordon B. Willis. 2007. "Research Synthesis: The Practice of Cognitive Interviewing." *Public Opinion Quarterly* 71:387–12.

Bickart, Barbara, and E. Marla Felcher. 1996. "Expanding and Enhancing the Use of Verbal Protocols in Survey Research." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, eds. Norbert Schwarz, and Seymour Sudman, pp. 115–42. San Francisco, CA: Jossey-Bass.

Blair, Johnny, Allison Ackermann, Linda Piccinino, and Rachel Levenstein. 2007. "Using Behavior Coding to Validate Cognitive Interview Findings." Proceedings of the American Statistical Association: Section on Survey Research Methods.

Blair, Johnny, Frederick Conrad, Allison Ackerman and Greg Claxton. 2006. "The Effect of Sample Size on Cognitive Interview Findings." Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Blair, Johnny and Stanley Presser. 1993. "Survey Procedures for Conducting Cognitive Interviews to Pretest Questionnaires: A Review of Theory and Practice." Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association, 370–75. Alexandria, VA: American Statistical Association .

Conrad, Frederick, and Johnny Blair. 1996. "From Impressions to Data: Increasing the Objectivity of Cognitive Interviews." Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association, 1–10. Alexandria, VA: American Statistical Association.

Conrad, Frederick G., and Johnny Blair. 2004. "Aspects of Data Quality in Cognitive Interviews: The Case of Verbal Reports." In *Methods for Testing and Evaluating Survey Questionnaires*, eds. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer, pp. 67–88. New York: Wiley.

Conrad, Frederick G., Johnny Blair, and Elena Tracy. 1999. "Verbal Reports are Data! A Theoretical Approach to Cognitive Interviews." *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, Tuesday B Sessions, Arlington, VA, 11–20.

Conrad, Frederick G., Norman R. Brown, and Erin Cashman. 1998. "Strategies for Estimating Behavioural Frequency in Survey Interviews." *Memory* 6:339–66.

Ericsson, Anders and Herbert Simon. 1993. *Protocol Analysis: Verbal Reports as Data*. 2nd ed. Cambridge, MA: MIT Press.

Everitt, Brian. S., and Dale F. Haye. 1992. *Talking about Statistics: A Psychologist's Guide to Data Analysis*. New York: Halsted Press.

Fowler, Floyd J. Jr. 2004. "The Case for More Split Sample Experiments in Developing Survey Instruments." In *Methods for Testing and Evaluating Survey Questionnaires*, eds. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer, pp. 173–88. New York: Wiley.

Forsyth, Barbara H., Judith T. Lessler, and Michael L. Hubbard. 1992. "Cognitive Evaluation of the Questionnaire." In *Survey Measurement of Drug Use: Methodological Studies*, eds. C. Turner, J. Lessler, and J. Gfroerer. Rockville, MD: US Department of Health and Human Services.

Green, David A. and, John M. Swets. 1966. *Signal Detection Theory and Psycho-Physics*. New York: Wiley.

Grice, Paul H. 1975. "Logic and Conversation." In *Syntax and Semantics: Volume 3, Speech Acts*, eds. P. Cole, and J. L. Morgan, pp. 41–58. New York: Academic Press.

Groves, Robert M., Floyd J. Fowler Jr., Mick. P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. New York: Wiley.

Groves, Robert M., Nancy H. Fultz, and Elizabeth Martin. 1992. "Direct Questioning about Comprehension in a Survey Setting." In *Questions about Answers: Inquiries into the Cognitive Bases of Surveys*, ed. J. Tanur, pp. 49–61. New York: Sage.

Lessler, Judith, Roger Tourangeau, and William Salter. 1989. "Questionnaire design in the cognitive research laboratory: Results of an experimental prototype." *Vital and Health Statistics*, Series 6, No. 1 (DHHS Pub. No. PHS 89-1076). Washington, DC: US Government Printing Office.

Menon, Geeta. 1993. "The Effects of Accessibility of Information in Memory on Judgments of Behavioral Frequencies." *Journal of Consumer Research* 20:431–40.

Presser, Stanley, and Johnny Blair. 1994. "Do Different Pretest Methods Produce Different Results?" *Sociological Methodology* 24:73–104.

Russo, Jay, Eric Johnson, and Deborah Stephens. 1989. "The Validity of Verbal Protocols." *Memory and Cognition* 17:759–69.

Schneider, Walter, and Richard M. Shiffrin. 1977. "Controlled and Automatic Human Information Processing: 1. Detection, Search, and Attention." *Psychological Review* 84:1–66.

Schooler, Jonathan W., Stellan Ohlsson, and Kevin Brooks. 1993. "Thoughts Beyond Words: When Language Overshadows Insight." *Journal of Experimental Psychology: General* 122:166–83.

US Department of Commerce. 2003. Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses. US Census Bureau, Economics and Statistics Administration.

Willis, Gordon B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.

Willis, Gordon B., Theresa J. DeMaio, and Brian Harris-Kojetin. 1999. "Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques." In *Cognition and Survey Research*, eds. M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau, pp. 133–53. New York: Wiley.

Willis, Gordon B., and Judith Lessler. 1999. *The BRFSS-QAS: A Guide for Systematically Evaluating Survey Question Wording*. Rockville, MD: Research Triangle Institute.

Willis, Gordon B., Patricia Royston, and Deborah Bercini. 1991. "The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires." *Applied Cognitive Psychology* 5I:251–67.

Willis, Gordon B., and S. Schechter. 1997. "Evaluation of Cognitive Interviewing Techniques: Do the Results Generalize to the Field?" *Bulletin de Methodologie Sociologique*, 55(June):40–66.

Wilson, Timothy, Susanne J. LaFleur, and D. Eric Anderson. 1996. "The Validity and Consequences of Verbal Reports about Attitudes." In *Answering Questions: Methodology for Determining the Cognitive and Communicative Processes in Survey Research*, eds. Norbert Schwarz, and Seymore Sudman, pp. 91–114. San Francisco: Jossey-Bass.

Wilson, Timothy, and Jonathon Schooler. 1991. "Thinking Too Much: Introspection can Reduce the Quality of Preferences and Decisions." *Journal of Personality and Social Psychology* 60:181–92.