Public Opinion Quarterly, Vol. 71, No. 2, Summer 2007, pp. 287–311

RESEARCH SYNTHESIS: THE PRACTICE OF COGNITIVE INTERVIEWING

PAUL C. BEATTY
GORDON B. WILLIS

Abstract Cognitive interviewing has emerged as one of the more prominent methods for identifying and correcting problems with survey questions. We define cognitive interviewing as the administration of draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends. But beyond this general categorization, cognitive interviewing potentially includes a variety of activities that may be based on different assumptions about the type of data that are being collected and the role of the interviewer in that process. This synthesis reviews the range of current cognitive interviewing practices, focusing on three considerations: (1) what are the dominant paradigms of cognitive interviewing—what is produced under each, and what are their apparent advantages; (2) what key decisions about cognitive interview study design need to be made once the general approach is selected (e.g., who should be interviewed, how many interviews should be conducted, and how should probes be selected), and what bases exist for making these decisions; and (3) how cognitive interviewing data should be evaluated, and what standards of evidence exist for making questionnaire design decisions based on study findings. In considering these issues, we highlight where standards

PAUL C. BEATTY National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Room 3218, Hyattsville, MD 20782. GORDON B. WILLIS National Cancer Institute, National Institutes of Health, 6130 Executive Blvd., MSC 7344, EPN 4005 Bethesda, MD 20892-7344. Earlier versions of this paper have benefited from the comments and encouragement of many people, including Duane Alwin, Bob Groves, Mick Couper, Roger Tourangeau, Norbert Schwarz, Jim House, Howard Schuman, and the anonymous reviewers. We also thank our many colleagues who have discussed various points raised here with us over the course of several years, especially Kristen Miller. Views expressed in this article are those of the authors, and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention, or of the National Cancer Institute, National Institutes of Health. Address correspondence to Paul C. Beatty; e-mail: pbb5@cdc.gov.

doi:10.1093/poq/nfm006 Advanced Access publication May 29, 2007.

for best practices are not clearly defined, and suggest broad areas worthy of additional methodological research.

Developing and evaluating questions has always been one of the key challenges for survey researchers. Beginning in the 1980's, cognitive interviewing has emerged as one of the more prominent methods for identifying and correcting problems with survey questions. Numerous academic survey centers, government agencies, and commercial research firms have incorporated cognitive interviews into their usual procedures for questionnaire development, and some organizations (e.g., the Census Bureau, National Center for Health Statistics, Bureau of Labor Statistics, Westat, RTI International, and Abt Associates) have created permanent questionnaire design laboratories to facilitate this practice. Cognitive interviewing was a prominent topic at the 2002 conference on Questionnaire Design, Evaluation and Testing, and the volume resulting from the conference describes several facets of this activity (Presser et al. 2004). In addition, a book by Willis (2005) contains an extensive review of the methodology and serves as a practical guide for carrying out cognitive interviewing projects.

In spite of these developments, there does not appear to be a succinct and commonly accepted definition of what cognitive interviewing[1] consists of, or consensus regarding its best practices (Presser et al. 2004). A basic definition proposed by Beatty (2003) that seems to reflect its most common application is that cognitive interviewing entails administering draft survey questions while collecting additional verbal information *about* the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends. But beyond this general categorization, cognitive interviewing potentially includes a variety of activities that may be based on different assumptions about the type of data that are being collected and the role of the interviewer in that process. For example, the verbal material generated by such interviews could consist of (1) respondent elaborations regarding how they constructed their answers, (2) explanations of what they interpret the questions to mean, (3) reports of any difficulties they had answering, or (4) anything else that sheds light on the broader circumstances that their answers were based upon. This material could be based on explicit follow-up questions (probes) from an interviewer, or based on general instructions to "think out loud" as much as possible. The interviewer herself could range from a relatively unskilled data collector to an expert investigator; the interview could be based on

1. This usage is to be distinguished from the "cognitive interview" described by Fisher and Gieselman (1992) used to extract information from event eyewitnesses.

a scripted protocol, be semi scripted, or largely improvised based on the issues that emerge from discussion. Analysis may be based on systematic review of interview transcripts, or entirely from notes taken during the interview [more detailed discussion of the potential varieties of cognitive interviewing practice are provided by Willis 2005 and Conrad and Blair 2004].

Given such variety, it may be difficult to understand what someone means when claiming to have conducted cognitive interviews. This is certainly a problem for consumers of cognitive interview findings, but a lack of consensus on objectives, procedures, or even general terminology can also inhibit methodological development. Providing a "foundation for optimal cognitive interviewing," as called for by Presser et al. (2004, p. 115) seems to require both an understanding of current cognitive interviewing practices, and an analysis of areas where best practices are unclear. This synthesis has been written to serve as a key element of that foundation, focusing on three key considerations: (1) what are the dominant paradigms of cognitive interviewing—what is produced under each, and what are their apparent advantages; (2) what key decisions about cognitive interview study design need to be made once the general approach is selected (e.g., who should be interviewed, how many interviews should be conducted, and how should probes be selected), and what bases exist for making these decisions; and (3) how cognitive interviewing data should be evaluated, and what standards of evidence exist for making questionnaire design decisions based on study findings.

Alternative Paradigms of Cognitive Interviewing: Thinking-aloud and Probing

All forms of cognitive interviewing entail administering survey questions to a participant[2] while collecting additional verbal information relevant to survey responses. Beyond that, practices seem to be based on two primary paradigms. One involves a cognitive interviewer whose role is to facilitate participants' verbalization of their thought processes, but to intervene as little as possible in generating this verbal information. The other involves an interviewer who guides the interaction more proactively, generally asking additional, direct questions about the basis for responses. The former paradigm is rooted in the think-aloud procedure, in which interviewers encourage participants to verbalize thoughts while answering questions (e.g., "tell me what you are thinking . . . how are you coming up with your

2. We use the term "participant" in place of the word "respondent" to distinguish between those answering questions in a cognitive interview and an actual survey (although we note that the term "subject" is also sometimes used to refer to the former (e.g., Willis 2005).

answer to this?"). The latter paradigm is rooted in the practice of intensive interviewing with follow-up probes (e.g., "Can you tell me in your own words what that question was asking?")

THE "PURE" THINK-ALOUD AND NON-INTERVENING COGNITIVE INTERVIEWER

Although the definition of cognitive interviewing provided earlier is quite broad, its original paradigm was both more specific and more explicitly psychological. The initial impetus for cognitive interviewing came from an interdisciplinary seminar on the Cognitive Aspects of Survey Methodology, often referred to as the first "CASM" meeting (and summarized in Jabine et al. 1984). Loftus (1984), elaborating upon ideas presented at this meeting, proposed that a psychological research technique known as *protocol analysis* could be adapted as a pretesting methodology for survey questions. The blueprint for this technique was developed by Ericsson and Simon (1980, expanded in 1993) and relies upon concurrent *think-aloud* reports. Think-aloud reports were used to yield insights into the thought processes involved in participants' completion of certain tasks in a laboratory setting. The validity of the procedure assumes that these reports are available, can be accurately reported, and that reporting them does not fundamentally change the activities that participants reported about. Ericsson and Simon argued that these assumptions can often be met, and Loftus suggested that think-alouds yielded information about how participants tended to retrieve memories (e.g., of medical visits). Such data could be used to develop questions that reflected these retrieval strategies. For example, she suggested defining the reference period of recall questions from a past date up to the present, rather than from the present backwards.

Early papers on cognitive laboratory methods (e.g., Royston et al. 1986) suggest that initial cognitive interviews were based heavily, if not exclusively, upon instructions to participants to think-aloud as they thought about and answered survey questions. In practice, this meant that cognitive laboratory participants were asked to report what they were thinking while answering, and interviewers simply reminded respondents to continue providing such information as necessary. Think-aloud responses were the dominant data produced in the interviews, and interviewer behavior was constrained accordingly. It is probably not surprising that this was the original paradigm of cognitive interviewing, since it closely followed the methodology of protocol analysis that served as its basis.

AN ALTERNATIVE PARADIGM: INTERVIEWERS ASKING DIRECT QUESTIONS TO EXPLORE THE SURVEY RESPONSE PROCESS

At some point, an alternative paradigm of cognitive interviewing emerged that expanded upon the use of "pure" think-alouds—in particular, allowing

for the addition of direct probing by the interviewer. Several precedents involving "intensive interviewing" had emerged independent of the first CASM meeting—for example Streett (1983) had described "frame of reference" questions about how respondents interpreted terms, which could be administered within pretests. In addition, Converse and Presser (1986) proposed "participating pretests," in which respondents would be specifically told that the interview is evaluative and asked to explain their answers; scripted and unscripted probes might be used. Both works cite Belson's (1981) work as a precedent, in which he probed about respondent interpretations of questions and the circumstances surrounding their responses in an effort to identify reporting errors. Apparently, the distinction between true think-aloud interviews and "intensive interviews" became blurred at some point, with both eventually falling under the header of "cognitive interviewing." It is easy to imagine how this could have occurred, especially given that intensive probes were often "cognitive"—addressing how terms were interpreted, how participants remembered certain facts, whether answers fit into available response categories, and so on.

This probing-based paradigm appears to have evolved gradually. Although some early descriptions of cognitive interviewing (Bercini 1992; Forsyth and Lessler 1991; Royston 1989) focus on think-alouds as the dominant component of cognitive interviewing, others (Royston and Bercini 1987; Willis, Royston, and Bercini 1991) suggested that both think-alouds and probing could be viable alternatives. Later, Willis (1994) proposed putting a greater emphasis on probing, primarily based on the observation that thinking-aloud seems awkward and burdensome for many participants. DeMaio and Rothgeb (1996) then proposed that cognitive interviews could include interviewer-guided activities such as probes about comprehension, confidence ratings, and requests to paraphrase questions in the absence of thinking-aloud. Several other articles suggest that the trend toward acceptance of such activities continued; Gerber and Wellens (1997) noted that cognitive interviewing had seemed to evolve from its original form to include "more probes and probes about meaning than was originally intended" (p. 35). O'Muircheartaigh (1999), along the same lines, suggests that cognitive interview practice had "diverged substantially" from the original paradigm derived from Ericsson and Simon. Willis, DeMaio, and Harris-Kojetin (1999), noting that cognitive interviews are often called "think-aloud interviews," recommended that the latter term should be used more sparingly because think-aloud protocols were not necessarily the dominant component of cognitive interviewing as it was then practiced.

In short, many practitioners of cognitive interviewing began to collect verbal material other than "pure" think-alouds and, at least in some cases, empowered the interviewer to guide interviews based on individual content. This is not to say that the use of think-alouds was abandoned, as virtually all descriptions of cognitive interviewing mention think-alouds as one possible

component (see DeMaio and Landreth 2004; Willis 2005), and some researchers (e.g., Conrad, Blair, and Tracy 2000) have continued to favor that approach. However, this new paradigm owed relatively little allegiance to the procedures for verbal protocol analysis proposed by Ericsson and Simon. Rather, it appears to have emerged for pragmatic reasons—some researchers gravitated toward this paradigm simply because it gave them additional useful information.

AN ASSESSMENT OF THE TWO PARADIGMS

The goal under both probing and think-aloud paradigms is to generate verbal information that is usually unseen in a survey interview in order to evaluate how well the questions are meeting their objectives. This puts interviews from both paradigms on an important common ground. Yet they are carried out differently and are based on different assumptions, which may have important implications regarding the nature of the data that they generate. Advocates of the think-aloud paradigm propose that it has several advantages. One is that the procedures are relatively standardized, reducing the chances that interviewers could introduce bias into the data collection process; another is that interviewers do not need to be knowledgeable about questionnaire design or the objectives of specific questions (Bolton and Bronkhorst 1996). Conrad, Blair, and Tracy (2000) note that interviewer probing can create artificiality (e.g., changing content and flow), whereas these problems are presumably avoided in think-aloud interviews. This artificiality is the major reason why Oksenberg, Cannell, and Kalton (1991) proposed that probing, when used, should follow only a few questions per interview. It is possible that apparent problems with a survey question could be products of a unique interaction between cognitive interviewers and participants, rather than "real" questionnaire problems (Beatty, Willis, and Schechter 1997).

Forsyth and Lessler (1991) and van der Veer, Hak, and Jansen (2002) propose an additional advantage: because think-aloud data are collected *during* the response process, they have a certain purity that probe responses (provided *after* responding) do not. However, a considerable body of research beginning with Nisbett and Wilson (1977) calls into question whether think-alouds are literal reflections of thought processes. More likely, they are re-constructions, although they are likely to reflect actual processes to some degree (Wilson, LaFleur, and Anderson 1996). Furthermore, as Willis (2004) notes, Ericsson and Simon (1980) did not insist upon exclusive use of thinking-aloud: their crucial point was that self-reported information should be in *short-term memory*. From that perspective, reports based on probes immediately following questions are probably not much different than think-aloud reports. Also, the practical value of think-aloud reports for questionnaire designers may not depend upon literal accuracy. For example, DeMaio, Ciochetto, and Davis (1993)

determined through verbal protocols that some dietary recall questions forced participants to think chronologically, which was difficult for many. Revised questions focused on *what* was eaten rather than *when* it was eaten, freeing participants to use whatever recall strategy worked best for them. Whether or not the participants reported literal thought processes, the material they provided highlighted why one version of the question was difficult, and suggested a viable alternative.

The think-aloud procedure generates information that can be useful to questionnaire designers, can be administered with only modest training, and appears to avoid problems of artificiality that could be created through probing. However, Willis (2005) also reviews evidence that some cognitive interview participants perform think-alouds poorly. Some psychological researchers acknowledge variation in ability to perform this task, but suggest that it does not seem to be correlated with any other observable variables (van Someren, Barnard, and Sandberg 1994). The extent to which this varying ability limits the effectiveness of think-alouds for evaluating questionnaires is not completely clear.

In addition to overcoming this potential weakness, advocates of the more probing-centered paradigm claim that probing offers several advantages. For example, Willis (1994, 2005) suggests that probing provides focus to the participant's behavior. Given that participants may diverge onto irrelevant tangents when relying entirely upon general instructions to articulate thought-processes, he suggests that carefully selected probes help to focus attention on pertinent issues. Of course, such probing requires interviewer judgment regarding both what the most pertinent issues are, and what probes are most appropriate to return attention to those points. This is important, because it is not the use of probes per se that regains control of the interview, but an interviewer skilled at using the "right" probes. The implication of Willis' suggestion is that interviewers should retain discretion over interview content; however, they may also require special expertise to wield that discretion effectively.

Another potential advantage of probing is that theoretically it should not interfere with the actual process of responding, whereas thinking-aloud might. Although probe responses are likely to be quite similar to think-aloud reports (see above), procedures for obtaining them are different; in the think-aloud case, participants at least attempt to provide some verbal information *during* the response process. Although Ericsson and Simon (1980) claim that thinking-aloud probably does not interfere with the response process, Russo, Johnson, and Stephens (1989) found that it had an impact on the accuracy of various mental computations; furthermore, Willis (1994) argues that thinking aloud is likely to increase the effort spent on creating a response, which has an unknown impact on the response process. Probing may create less interference than thinking aloud, while still capturing information stored in short-term memory. (However, as noted previously, probing may interfere

with the usual flow of the interview, potentially compromising the realism of questionnaire administration in a different manner.)

Perhaps the strongest justification for the probe-based paradigm is that it generates verbal material that questionnaire designers find useful, but that may not emerge unless a cognitive interviewer specifically asks for it. As Willis (2004) observes, think-aloud procedures were originally proposed to shed light specifically on *retrieval* processes. It is unclear whether think-aloud responses are useful for assessing comprehension problems, inadequate response options, or other questionnaire issues, whereas it is generally straightforward to probe directly about such matters. Similarly, Conrad, Blair, and Tracy (2000) note that think-alouds alone sometimes suggest a problem with a question but do not provide enough information to diagnose what the problem is. Probe responses might help to fill in this gap.

Although it is useful to present these paradigms as distinct, the boundaries between practices are probably not as precise as they once were, with both allowing for some degree of probing, albeit with some reservations. Advocates of the original paradigm (Conrad, Blair, and Tracy 2000) have conceded that probing makes important contributions, and advocates of the alternative (Beatty 2004) have acknowledged that probing can shape interview content in some undesirable ways. However, an important distinction remains regarding the expected role of the interviewer. The original paradigm employs an *unobtrusive* cognitive interviewer, who relies on standardized think-aloud protocols and possibly scripted probes. The alternative paradigm employs an *active* cognitive interviewer who is given more latitude to explore topics as they emerge within interviews. Thus, the practical decision has moved from whether or not to allow probes, to *how much* probing is appropriate, and whether this probing should be standardized or determined by interviewer judgment (or to what extent).

At this point it would be relatively easy to conclude, as does Willis (2005), that "In practice, think-aloud and verbal probing actually fit together very naturally" (p. 57), and that it is appropriate to call a truce such that practitioners adopt both methods. However, the appropriate relative weight to be put on either procedure may depend on specific factors relevant to testing. Optimal cognitive testing procedures may vary depending upon the age and cognitive ability of participants, topic of the questionnaire, or intended mode of administration. For example, Redline et al. (1998) evaluated a self-administered paper instrument with both think-aloud and purely retrospective approaches. They found that the methods produced similar results, except that participants with low educational levels tended to miss skip patterns under the think-aloud approach. This suggests that think-alouds interfered with normal navigation among some participants, making the method less desirable; however, this problem might not exist with computer-assisted instruments that handle skip logic. Furthermore, the increased prevalence of web surveys puts increased emphasis on

self-administered visual modes. While a probing paradigm seems more common for interviewer-administered surveys, several researchers have considered whether different approaches might be best for self-administered questionnaires or web surveys (Bates and DeMaio 1989; Dillman and Redline 2004; Redline et al. 1998; Schechter, Blair, and Vande Hey 1996). It may be that navigational and other creative decisions involved in completing web surveys are more appropriately evaluated through think-alouds, and the need for such assessments could spur a resurgence of that technique (Willis 2005).

Specific Parameters in Design and Implementation: Who to Interview, and How to Probe

Although Forsyth and Lessler (1991), Willis (1994), and DeMaio and Rothgeb (1996), among others, have contributed significantly to establishing general descriptions of cognitive interviewing, the literature is generally not very detailed concerning many specifics regarding design and implementation of studies based upon this method. There does not appear to be general consensus regarding issues such as sample sizes needed for adequate testing, participant selection, the ideal background and training of interviewers, and choice of cognitive probes [although Willis (2005) does provide recommendations in these areas]. This section reviews some considerations along these lines, and suggests considerations that are useful in making decisions about specific approaches.

WHO TO INTERVIEW, AND HOW MUCH INTERVIEWING TO DO

Cognitive interviewing literature has paid surprisingly little attention to issues of appropriate composition and size of cognitive interviewing samples. Practitioners generally acknowledge that participants are chosen by convenience and that such samples are "not designed to be representative [of any larger population], but to reflect the detailed thoughts and problems of the few respondents who participate in [cognitive interviews]" (DeMaio et al. 1993). One clear consequence of such sampling is that cognitive interview practitioners cannot directly determine the extent of questionnaire problems in the population—they can only identify question characteristics that are believed to pose problems with some unspecified frequency. Other than that, the specific guidance that is available advocates demographic variety of respondents, and that participants should include people relevant to the topic of the questionnaire being tested (Willis 1994, 2005).

Still, some sampling considerations could help to strengthen claims that a reasonably thorough effort has been made to identify the most pressing problems with a questionnaire. For example, participants can be selected to cover as much of a questionnaire's conceptual terrain as possible. If questionnaires include skip patterns that lead to various branches, the

sample should be sufficiently diverse to explore as many of these different paths as possible. Whatever topic the questions focus on (e.g., health insurance), the sample should cover a variety of circumstances relevant to that topic (e.g., people with a variety of health insurance situations, including some with no insurance at all). Within those parameters, it also seems desirable to select participants representing some demographic variety. This does not ensure "representativeness," but casting as wide a net as possible over varying circumstances maximizes the chances that discovery will be effective. Similarly, interviewing in multiple locations could improve the variety of circumstances that are captured in testing (Miller 2002).

It is also not clear whether cognitive interviewers select samples of adequate size. Current practices seem based on the assumption that the most critical questionnaire problems will be revealed by a small sample of relevant participants. However, Blair et al. (2006) found that in one study, significant questionnaire problems were uncovered even after 50 or more cognitive interviews, and that some of the most serious problems (as judged by external reviewers) were not identified until relatively late in the process. General guidance calls for cognitive interviews to be conducted in "rounds" that mostly commonly range between 5 and 15 interviews, which are ideally repeated following efforts to revise questions and eliminate problems (Willis 1994, 2005; McColl 2001). This iterative approach seems useful, but it is not clear how many rounds are usually conducted given time and resource constraints. Even under ideal circumstances, this approach still leaves open the question of how researchers can determine that they have conducted enough rounds of interviews to warrant stopping the process. Some qualitative researchers make such decisions based on the idea of *category saturation* (Strauss and Corbin 1990). Put simply, this means that the researcher identifies groups of people most relevant to the study and conducts interviews with members of each until they yield relatively few new insights. While the total sample sizes generally used in cognitive interviewing might fall short of those required to reach the point where insights actually stop emerging, the general principle of operating based on diminishing returns may be useful. On the whole, it seems unlikely that typical sample sizes currently used for cognitive interviewing are sufficient to provide truly comprehensive insight into the performance of a questionnaire, and additional standards are needed to determine optimal sizes.

COGNITIVE INTERVIEWERS AS DATA COLLECTORS OR INVESTIGATORS

Tucker (1997)—in a position largely consistent with the original paradigm discussed earlier—calls for much greater standardization of cognitive interview procedures. Without this, he argues that "effective manipulation [of variables] will be impossible . . . the notion of falsifiability has no meaning . . . [and] the conditions necessary for generalizability will

be absent" (p. 72). Conrad and Blair (1996) similarly argue that "rigorous experimental methods that rely on objective quantifiable data" are preferable to "largely impressionistic methods" that they suggest are generally used in cognitive laboratories (p. 8). The methodological research agenda proposed by Conrad and Blair (2004) to evaluate the quality of verbal report data is based primarily on experimentation across alternative, clearly defined techniques; this would require relatively high standardization as well. Under these perspectives, creative contributions from interviewers that lead to nonstandardized behavior are undesirable and should be minimized. Interviewers would primarily serve as data collectors, and researchers would ideally determine issues they wished to probe about in advance. This puts a relatively high investigative burden on the front-end of the process (i.e., before data collection).

An alternative perspective is that interviewers themselves would serve as investigators, potentially making decisions about content and scope *during* data collection. Such interviews might have an exploratory character, being more attuned toward generating ideas about potential problems than determining their extent in the population. For example, Willis (1994) compares cognitive interviewers to "detectives" who rely at least partially upon improvisation in looking for clues about questionnaire problems. In subsequent work, he draws an analogy between cognitive and clinical interviews, which may be guided by intuition, experience, and flexibility (Willis 2004, 2005). This perspective forgoes consistency across interviews in favor of freedom to explore issues that emerge in discussions with participants. Its advantage is that it allows interviewers to explore issues that might have been missed through more tightly scripted interviews.

Beatty (2004), for example, reported results from a test of the following question: "In the past 12 months, how many times have you seen or talked on the telephone about your physical, emotional, or mental health with a family doctor or general practitioner?" One participant answered "zero," but answers to other questions suggested that he had received medical care recently. In response to a series of follow-up probes, the participant eventually explained that his answer was accurate because the question referred to *talking on the phone*—the word "seen" had been lost within the verbiage of the question. In subsequent interviews, interviewers found that this mistake was commonly made. The identification of this problem emerged specifically from interviewer improvisation based on careful attention to responses. Furthermore, this probing provided insight into the particular causes of the problem.

However, the decision to rely upon this sort of *emergent probing* places considerable trust in the interviewer's ability to notice potential problems and choose appropriate follow-up probes. There does not appear to be sufficient guidance about what sort of training or background is most appropriate to prepare such interviewers for this role, though Willis (2005)

describes a developmental training process that emphasizes study of questionnaire design, observation of cognitive interviews, and "on-the-job" practice that is critiqued by experienced interviewers. Clearly, it is important for such interviewers to understand potential types of cognitive or communicative errors that could affect the accuracy of survey responses, and to have familiarity with various options for eliciting useful verbal material from participants. A thorough grounding in survey methodology would probably be useful. Interviewers would also need to understand the measurement objectives of the questions being tested.

In contrast, cognitive interviewers functioning largely as data collectors would not require this level of expertise. Like traditional survey interviewers (e.g., see Fowler and Mangione 1990), their role would be primarily to read pre-determined questions and follow instructions accurately. Detailed knowledge of survey errors and measurement objectives might be useful, but would not be necessary. Interviewers would not need to understand why think-alouds or probes were being administered, although they would need to exercise judgment in recognizing when participants had provided adequate think-aloud or probe responses.

Clearly, the use of interviewer-investigators requires a highly specialized work force that is likely to be more expensive and difficult to assemble than the alternative. Studies based on this paradigm are also likely to involve higher levels of interviewer variation and improvisation. In order to fully assess whether insights gleaned from such efforts outweigh the drawbacks of expense and lack of generalizability, additional research could explore what interviewers actually do under various paradigms of practice and how conclusions are reached. The results would help to make better choices about the costs and benefits of various approaches to cognitive interviewing.

In particular, cognitive interviewers increasingly face the need to conduct cross-cultural and multi-lingual testing, where monolingual staff simply cannot conduct the interviews. It will be necessary to establish means for either quickly training new cognitive interviewers to be proficient, or else to develop standardized testing protocols that require lower levels of proficiency. Kudela et al. (2004) describe such a cross-cultural cognitive testing project. Researchers were able to coordinate cognitive interviewing of a tobacco use questionnaire across several Asian languages as well as English, relying on relatively quickly trained interviewers who applied a standardized protocol (that is, where probes were completely pre-scripted). The overall results were somewhat reassuring; at least in part, a common set of problems emerged across cultures, languages, and sets of interviewers, indicating that some problems with survey questions appear to be universal, and that the separate cognitive interviewing teams independently identified these defects.

WHAT TO ASK: THE SELECTION OF PROBES

As discussed earlier, most recent conceptualizations of cognitive interviewing involve probing to some degree. If the interviewer is also an investigator, then she may select some of these probes herself; if a data collector, then the probes may be selected for her. But either way, someone must choose what probes are asked. Although the cognitive interviewing literature of the past 20 years has provided many examples of possible probes (e.g., Bercini 1992; DeMaio and Rothgeb 1996; Forsyth and Lessler 1991; Willis 1994), it is not clear whether particular probes are likely to be most effective for various purposes. Willis (1994) suggested that probes should not suggest a "correct" answer, a principle that also applies to survey questions. Foddy (1998) concluded that specific probes such as "what does [term] mean to you" are more successful than general ones such as "what were you thinking when you first answered the question?[3] Beatty (2002) found that participants answered probes about the meaning of terms differently when they were administered alone than they did within the context of a particular survey question. In general, however, these sorts of recommendations appear to be uncommon.

Cognitive interviewers may be able to obtain some guidance about how to choose "good" probes from literature on qualitative interviewing, which may include lessons on what to ask, how to ask it, and how to make sense of narrative data. For example, Weiss (1994) suggests that interviewers generate narrative by asking about specific events rather than generalized experience. Holstein and Gubrium (1995) encourage interviewers to be on the lookout for "confusion, contradictions, ambiguity and reluctance" as signs that "meanings are being examined, reconstituted, or resisted" (p. 79). In the case of cognitive interviews, such instances might call for additional probing. Variants of qualitative interviewing are also employed by anthropologists, and some guidance may be obtained from that field as well. For example, Gerber (1999) notes that anthropologists can explore whether terms are "culturally inappropriate" for a particular population. But rather than simply asking a participant what a term such as *self-reliance* means, an anthropologist might explore its meaning in different contexts, e.g., with regard to child rearing, older family members, or welfare recipients. This might suggest that general cognitive interview probes such as "what does this term mean to you" might be less effective than specific ones exploring how a term is used in a participant's life. Figure 1 suggests a probing classification scheme that attempts to systematically organize the

3. It should also be noted that in this study, general probes were administered before specific probes for most questions. It is possible that the performance of specific probes was enhanced by "priming" from the general probes. It is also possible that results would vary depending on the criteria used to determine that probes were successful.

	Proactive Administration (initiated by the interviewer/researcher)	**Reactive Administration** (triggered by subject behavior)
Standardized Construction (constructed *prior* to the interview)	(1) Anticipated Probes	(3) Conditional Probes
Non-standardized Construction (constructed *during* the interview)	(2) Spontaneous Probes	(4) Emergent Probes

Figure 1. Model of verbal probing in the cognitive interview (from Willis 2005).

various major categories of probes, and to indicate the circumstances under which each is potentially most beneficial.

This model of probing distinguishes two major dimensions: (1) whether probes are searching in nature (proactive) as opposed to responsive (reactive), and (2) whether they are fashioned ahead of the interview (standardized) or during its course (nonstandardized). The 2×2 combination of these two major dimensions produces four key probing variants:

(1) *Anticipated* probes are those that are scripted, or at least roughly configured, based on the anticipation of a problem with the question.
(2) *Spontaneous* probes are flexible in that they are not scripted in advance. However, these probes are not based on any particular response from participants—they derive from interviewers who decide to search for potential problems on their own initiative.
(3) *Conditional* probes have been introduced by Conrad and Blair (2004); these are pre-determined probes that are applied only if triggered by particular participant behaviors (e.g., if a participant hesitates, the protocol calls for the interviewer to say: "You took a little while to answer that question—what were you thinking about?").
(4) *Emergent* probes are flexible, unscripted, and reactive; the interviewer selects such probes in response to something that a participant says (e.g., something that indicates an apparent problem).

Willis (2005) discusses the various situations that might give rise to the use of each probe category, the relative benefits and drawbacks of each, and the relative benefits of structured versus flexible approaches to probing. In particular, he proposes that proactive varieties of probing are most useful when problems have been anticipated, yet the participant gives no indication of having the problem until probing is done (i.e., to detect cases of

"Silent Misinterpretation", as phrased by DeMaio and Rothgeb 1996). Reactive probing is presumably most appropriate in the opposite case: where problems were unanticipated, yet the participant does indicate some difficulty (e.g., a long silence). It is likely that multiple varieties of probing are appropriate within the same interview, depending on the mix of problems either expected or encountered.

To lend more specificity to these arguments, researchers might put this notion to the test. This could be done by paying careful attention to the nature of probing that is either proactive or reactive—or structured versus unstructured—within cognitive interviewing studies, mainly through careful review of cognitive protocols and interview recordings. Given the sets of (anticipated) probes that were both fashioned and administered, which resulted in the detection of apparent problems, and how often? Conversely, how often did subjects indicate the presence of problems on their own, how often were these followed up by reactive forms of probing, and what types of problems were then identified? A compilation of such basic data would be extremely useful in establishing best practices.

Evaluating Evidence from Cognitive Interviews

Whether cognitive interviews are conducted based on a fairly standardized protocol or with greater interviewer flexibility, the major product is still verbal text that needs to be evaluated to determine whether or not a question poses a problem for respondents. One advantage of more standardized protocols is that they generate data more amenable to systematic coding and analysis. For example, Conrad and Blair (1996) propose that verbal protocols be coded in a table with "types of problems" on one axis (lexical, temporal, logical, etc.), and "response stage" (understanding, task performance, and response formatting) on the other. Once criteria are established to indicate the presence of a problem, cognitive interview data can be objectively coded to determine if the criteria are met.

As noted earlier, some practitioners propose that additional, unscripted probing from a skilled investigator brings enough additional material to the surface to justify the lack of standardization. However, the resulting variation in data across questions and interviewers can complicate analysis. Interview content can vary considerably, with some questions or issues given relatively little attention and others pursued in depth. The presence or absence of verbal reports can be attributable to either questionnaire problems or interviewer discretion, making it difficult to establish objective criteria of problems. Nevertheless, we suggest that analysis can be based on whether apparent problems can be logically attributed to question characteristics. For example, consider the tested survey question "Thinking about your physical health, which includes physical illness and injury, for how many days during the

past 30 days was your physical health not good?" Schechter, Beatty and Willis (1998) concluded that respondents in general have a difficult time answering this by the following process of reasoning:

(1) Observing a response problem: several participants could not provide codeable responses (a number between zero and thirty), even when probed.
(2) Considering the specifics: some participants indicated that there was no way to answer the question ("a day is part good and part bad—you can't characterize it as one or the other"); others complained about response task difficulty ("I don't do bookkeeping on this") especially given complicated health pictures.
(3) Identifying question characteristic that create the problem: the question assumes that a "day" is a reasonable metric, but it may not be for people with varying-quality days.
(4) Evaluating generalizability: it seems reasonable that this problem could recur for people with multiple health problems difficult to keep track of, and for respondents with health that varies throughout the day.

A claim that this process found "proof" of the problem would overstate the evidence. However, a reasonable case could be made that respondents in similar circumstances would have similar difficulties responding, and that the difficulties are caused by a faulty assumption about the way individuals can characterize their health. Note also that the evidence is not linked to the *number* of participants who report a particular problem. Whether it takes many or a few participants to construct such an argument, it needs to be evaluated based on logical merits. It is conceivable that a solid argument about a questionnaire problem could be constructed around a single case, or that such an argument might fail to materialize around several cases.

ERROR IN COGNITIVE INTERVIEW ANALYSIS

As useful as cognitive interviewing may be, it can still lead to conclusions that are incomplete, misleading, or incorrect. There are several possibilities for error: cognitive interviews could identify problems that would not turn out to be "real" in surveys; cognitive interviews could fail to identify problems that exist in actual survey administration; and cognitive interview findings might be inconsistent when conducted by independent groups of researchers. The first two could be considered problems with the validity of the method and might be respectively classified as *errors of commission* and *errors of omission*. The third could be considered to be a problem with the reliability of the method.

In defending themselves against errors of commission, practitioners assume that *cognitive interviewing finds problems that will carry over to*

actual surveys. Unfortunately, there is often no obvious way to verify that hypothesized problems are "real." Logical arguments may have to do, and researchers will have to determine for themselves whether they find such arguments to be meritorious. However, there have been at lease two attempts to verify that cognitive interview findings were borne out by field data (Beatty, Fowler, and Cosenza 2006; Willis and Schechter 1997). Both studies administered several questions and revisions based on cognitive interview findings in a split ballot. Arguably, some revisions produced more plausible statistics (e.g., regarding hours spent doing "strenuous physical activity" per day). Such validation is potentially useful, but expensive.

As for *errors of omission*, there is no reasonable way that cognitive interview practitioners could claim to have found all problems with a questionnaire. As noted earlier, the method can make no claims that it has represented the population as a whole. Its usefulness is based on the assumption that the most egregious problems will become evident in most groups of participants who are reasonably appropriate to the topic of the survey, and interviewing often concludes based on a subjective judgment that interviews are yielding diminishing returns. There is always the possibility that one additional interview could yield a significant new insight, or that an additional interviewer would be more likely to notice additional problems. By the same token, claims that a questionnaire has "no problems" are impossible—the strongest claim that could be made is that no problems have (yet) been discovered.

Finally, there is always the possibility that independent groups of cognitive interviewers might not reach the same conclusions regarding a particular questionnaire (see DeMaio and Landreth 2004; Forsyth, Rothgeb, and Willis 2004). However, it would probably not be unusual for different groups to discover different insights, especially if interviewers were operating under an "investigator" paradigm. Under both the investigator and data-collector interviewer paradigms, there is an element of chance regarding who is interviewed, meaning that different insights might emerge from interviews, especially in earlier stages of a project. Differences could also be a function of varying backgrounds and sensitivities to various sorts of problems. It seems unlikely that cognitive interviewing would generate reliable findings in the sense that survey researchers might use the term (i.e., each set of interviews identifies the same set of problems with questions). When findings are different, yet not necessarily contradictory, this may indicate that no one set of findings is complete, and they need to be examined more closely. Findings that are difficult to reconcile might indicate either faulty reasoning by analysts, or that interviewing has not yet yielded an adequate understanding of responses associated with a question. The former case calls for a closer look at the data, while the latter indicates a need for continued data collection.

CONCEPTUALIZING THE PRODUCT OF COGNITIVE INTERVIEWING

Central to method evaluation is the choice of the particular outcome measure we choose to evaluate. Implicit in many discussions regarding cognitive interviewing is the assumption that it should help researchers develop *measurably better* survey questions—that is, it should be able to identify and eliminate problems until researchers have homed in on an "ideal" question wording. If held to such a standard, it is not clear that cognitive interviewing is successful. However, an alternative view is that cognitive interviewing should simply provide questionnaire designers with insights about the consequences of various questionnaire design decisions. These findings may not always point to a clearly superior version of a question. Rather than attempting to find the "right" way to ask a survey question, cognitive interviewing may be more suited to helping researchers assess *tradeoffs*—the advantages and disadvantages of asking questions in a certain manner.

For example, consider the question: "Are you currently being treated by a doctor for arthritis?" Cognitive interviewing has suggested that this question is simple for participants with either serious arthritis or no arthritis at all. However, it can be complicated for participants whose circumstances may not qualify as "current treatment" (e.g., those who had seen a doctor for arthritis pain over a year ago). Does this qualify as a "problem" with the question? If so, it might make sense to add more specific language than "current treatment." However, doing so would make the question longer, more burdensome, and potentially confusing to respondents who did not have difficulty with the original version. Cognitive interviewing may be useful simply because it provides information to make such design decisions as logically as possible—indeed, it may be the most efficient method available for illuminating such issues. In that light, cognitive interviewing may be less suited to finding the "best" questions than to guiding "best informed" design decisions.

Maximizing the quality of insights derived by cognitive interviewing calls for more clearly established best-practices, which calls for additional research. Clearly, experimentation on variations of cognitive interview practice (or comparisons with other forms of pretesting) is one possibility. In a straightforward experiment, researchers decide upon a particular variant of cognitive interviewing, manipulate a characteristic of interest, and determine whether this manipulation affected some dependent measure. For example, the researchers may provide identical instructions to two sets of interviewers with different levels of experience, and evaluate whether the number or type of problems that they identified varies. Conrad and Blair (2004) document a case study of this type, finding (among other things) that a more "conventional" version of

cognitive testing identified more problems than a restricted "conditional probe" version applied by less experienced interviewers; however, in the conventional probe version, there was less overall agreement with coders about the presence of problems.

Such experiments might be informative about the effect of manipulating particular variables, which could be especially useful for evaluations of relatively structured forms of cognitive interviewing. However, it is possible that the manipulations may have little resemblance to actual practice (a problem noted by Conrad and Blair 2004). This could be a particular problem in efforts to evaluate more qualitative forms of cognitive interviewing—i.e., those in which interviewer freedom to explore in an unpredicted (and possibly unpredictable) manner is a fundamental attribute of the method. It is difficult to use experiments to compare the yield of alternative qualitative methods, because the level of control required for such an evaluation makes the methods non-qualitative by definition. To the extent that cognitive interviewing is qualitative in nature, the usefulness of experimental manipulations of methodology may be limited by artificiality.

Another option for research is to study activities already conducted by cognitive interviewers. Such studies relinquish most or all experimental controls in order to maximize realism. For example, Presser and Blair (1994) compared results of cognitive interviewing with results from other forms of pretesting the same instrument; DeMaio and Landreth (2004) compared cognitive interview results produced by three separate research teams; Beatty (2004) studied what cognitive interviewers on one particular project actually did. All studies identified differences in results across methods or interviewers; however, the lack of experimental control makes it difficult to determine what specific characteristics of the study implementations were driving these differences. In fact, inherent differences in cognitive interviewing practice make it difficult for studies such as these to make firm conclusions about what the method does or does not accomplish in general. The same limitation applies to split-ballot studies (Beatty, Fowler, and Cosenza 2006; Willis and Schechter 1997) comparing response distributions from original and re-worked survey questions. The studies provide evidence that cognitive interviewing as practiced yielded useful results, but not all cognitive interviewing studies necessarily have equal utility. They also do not identify which attributes of cognitive interview practice were responsible for the usefulness of the results. Such studies can contribute an understanding of what one particular *variant* of cognitive interviewing produced—but only if the study makes clear exactly what was done in the cognitive interviews.

NEXT STEPS TOWARD DEVELOPING BEST PRACTICES

To close, we consider where the practice of cognitive interviewing seems the least well-specified, and suggest four areas where additional work should be particularly fruitful.

(1) *Determining optimal sample sizes for cognitive interviews*. Recent work by Blair et al. (2006) should be replicated using different instruments, types of questions, groups of interviewers, and variants of cognitive interviewing practice. One potentially useful variation would be to employ an iterative testing approach, based on rounds of testing with questionnaire revisions between rounds. This approach is arguably accepted as an ideal practice, and it would be useful to see whether revised questionnaires are in fact "better," and how rates of problem identification decline across revisions. Along the same lines, it would be useful to see how many interviews are required for independent groups of interviewers to reach consensus that the most significant questionnaire design issues had been identified.

(2) *Stronger guidance regarding data collection decisions*. As mentioned previously, the prevalence of web and mixed-mode surveys call for revisiting decisions about the proper balance between thinking-aloud versus probing. Further studies involving self-administration that directly compare think-alouds with various probing approaches should follow Redline et al. (1998) in assessing a range of relevant behaviors, such as usefulness of the verbal reports in identifying various conceptual and navigational problems on the same instrument. Similarly, additional research should evaluate the efficacy of various probing choices given various types of questions, modes, and participants. Although some guidance for selecting probes is available (see Willis 2005), little empirical evidence is available to demonstrate which design decisions yield better data for questionnaire design decisions. For example, analyses described by Foddy (1998) and Beatty (2004) examined the relationship between interviewing probing and participant reports. Insights into the actual workings of cognitive interviewing projects could suggest which approaches seem to be most effective.

(3) *Studying in greater depth what actually happens in cognitive interviews*. Although experimentation can be useful, we believe that greater methodological strides can be made through detailed analysis of *processes* in particular cognitive interviewing projects. That is, analysis could focus on what was done in a particular project, and how certain design decisions or interviewer actions led to certain conclusions, without necessarily attempting to control how studies were carried out. What such studies would lack in control they would

make up for in realism. Additional research on previously conducted cognitive interviewing studies may be increasingly feasible with the forthcoming launch of the Q-Bank database, which will document which questions were tested, methods used in the testing, and findings from cognitive interviewing projects conducted in various federal agencies (Miller 2005).

(4) *Encouraging enhanced documentation of procedures.* Perhaps the most significant impediment to the development of best practices is a lack of shared understanding of what cognitive interviewing projects have actually entailed. Given the large variety of practices covered under "cognitive interviewing," better documentation is essential. Cognitive interview practitioners should agree upon a common set of key parameters that are specified within any cognitive testing report—such as the number of interviews, interviewers, and rounds of interviewing; the relative mix of think-aloud versus probing, the nature of probing (concurrent versus retrospective, whether proactive, reactive, or both); the recruitment methods, and so on.

Conclusion

Although Sirken and Schechter (1999) have argued that cognitive laboratory testing has "forever changed the attitude that questionnaire design is simply an art rather than a science" (p. 4), the truth may be that there is as much art as science in cognitive interviewing. That does not necessarily diminish its value. Nor does it necessarily matter whether the method is true to its explicitly psychological roots—as Presser (1989, p. 35) concluded, if its application "does nothing more than increase the resources devoted to pretesting, it will have been all to the good." The practice of cognitive interviewing has resulted in a considerable expansion in the time and energy devoted to developing large-scale survey questionnaires over the past two decades. However it is implemented, cognitive interviewing can put useful information into the hands of researchers who need to make difficult choices. Hopefully this review provides some perspective about the varieties of activities that are encompassed under the term, which may help researchers to understand—and to further explore—what is actually being produced in cognitive interview studies. More importantly, it will hopefully foster further discussions about best practices among practitioners. With such continued discussion, researchers should be better equipped to create questions that are clear, that pose memory and recall tasks that respondents can reasonably be expected to accomplish, and that allow respondents to express their answers accurately.

References

Bates, Nancy, and Theresa J. DeMaio. 1989. "Using Cognitive Research Methods to Improve the Design of the Decennial Census Form." *Proceedings of the U.S. Bureau of the Census Annual Research Conference* 267–77.

Beatty, Paul. 2002. "Cognitive Interview Evaluation of the Blood Donor History Screening Questionnaire." In *Final Report of the AABB Task Force to Redesign the Blood Donor Screening Questionnaire*. Report submitted to the U.S. Food and Drug Administration.

_____. 2003. Answerable Questions: Advances in the Methodology for Identifying and Resolving Questionnaire Problems in Survey Research. Doctoral Dissertation, the University of Michigan *In Dissertation Abstracts International* 64(09): 3504A.

_____. 2004. "The Dynamics of Cognitive Interviewing." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

Beatty, Paul, Floyd J. Fowler, and Carol Cosenza. 2006. "Do Questionnaire Design Recommendations Lead to Measurable Improvements? Some Experiments with Alternate Versions of Complex Survey Questions." *Proceedings of the Q2006 European Conference on Quality in Survey Statistics.* Available online at: http://www.statistics.gov.uk/events/q2006/downloads/T04_Beatty.doc.

Beatty, Paul, Gordon Willis, and Susan Schechter. 1997. "Evaluating the Generalizability of Cognitive Interview Findings." In *Statistical Policy Working Paper No. 26: Seminar on Statistical Methodology in the Public Service*. Washington, DC: Statistical Policy Office, U.S. Office of Management and Budget.

Belson, William. 1981. *The Design and Understanding of Survey Questions*. Aldershot, England: Gower.

Bercini, Deborah H. 1992. "Pretesting Questionnaires in the Laboratory: An Alternative Approach." *Journal of Exposure Analysis and Environmental Epidemiology* 2:241–8.

Blair, Johnny, Frederick Conrad, Allison Ackermann, and Greg Claxton. 2006. "The Effect of Sample Size on Cognitive Interview Findings." Paper presented at the American Association for Public Opinion Research Conference, Montreal, Quebec, Canada.

Bolton, Ruth N., and Tina M. Bronkhorst. 1996. "Questionnaire Pretesting: Computer Assisted Coding of Concurrent Protocols." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, ed. Norbert Schwarz, and Seymour Sudman. San Francisco: Jossey-Bass.

Conrad, Frederick, and Johnny Blair. 1996. "From Impressions to Data: Increasing the Objectivity of Cognitive Interviews." 1996 *Proceedings of the Section on Survey Research Methods*, Vol. 1, pp. 1–9. Alexandria, VA: American Statistical Association.

_____. 2004. "Data Quality in Cognitive Interviews: The Case for Verbal Reports." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

Conrad, Frederick, Johnny Blair, and Elena Tracy. 2000. "Verbal Reports Are Data! A Theoretical Approach to Cognitive Interviews." In *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference*. Washington, DC: Office of Management and Budget.

Converse, Jean M., and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Survey Questionnaire*. Newbury Park, CA: Sage.

DeMaio, Theresa J., Susan Ciochetto, and Wendy Davis. 1993. "Research on the Continuing Survey of Food Intakes by Individuals." *1993 Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 1021–6. Alexandria, VA: American Statistical Association.

DeMaio, Theresa J., and Ashley Landreth. 2004. Cognitive Interviews: Do Different Methods Produce Different Results? In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

DeMaio, Theresa J., Nancy Mathiowetz, Jennifer Rothgeb, Mary Ellen Beach, and Sharon Durant. 1993. *Protocol for Pretesting Demographic Surveys at the Census Bureau*. Washington, DC: U.S. Bureau of the Census.

DeMaio, Theresa J., and Jennifer Rothgeb. 1996. "Cognitive Interviewing Techniques: In the Lab and in the Field." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, ed. N. Schwarz, and S. Sudman. San Francisco: Jossey-Bass.

Dillman, Don A., and Cleo D. Redline. 2004. Concepts and Procedures for Testing Paper Self-Administered Questionnaires: Cognitive Interview and Field Test Comparisons. In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

Ericsson, K. Anders., and Herbert A. Simon. 1980. "Verbal Reports as Data." *Psychological Review* 87:215–51.

_____. 1993. *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

Fisher, Ronald P., and R. Edward Geiselman. 1992. *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Thomas.

Foddy, William. 1998. "An Empirical Evaluation of In-Depth Probes Used to Pretest Survey Questions." *Sociological Methods and Research* 27:103–33.

Forsyth, Barbara H., and Judith T. Lessler. 1991. "Cognitive Laboratory Methods: A Taxonomy." In *Measurement Error in Surveys*, ed. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman. New York: Wiley.

Forsyth, Barbara H., Jennifer M. Rothgeb, and Gordon Willis. 2004. Does Questionnaire Pretesting Make a Difference? An Empirical Test Using a Field Survey Experiment. In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

Fowler, Floyd J., and Thomas W. Mangione. 1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park, CA: Sage.

Gerber, Eleanor R. 1999. "The View from Anthropology: Ethnography and the Cognitive Interview." In *Cognition and Survey Research*, ed. Monroe G. Sirken, Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur, and Roger Tourangeau. New York: Wiley.

Gerber, Eleanor R., and Tracy R. Wellens. 1997. "Perspectives on Pretesting: 'Cognition' in the Cognitive Interview?" *Bulletin de Methodologie Sociologique* 11:18–39.

Holstein, James A., and Jaber F. Gubrium. 1995. *The Active Interview*. Thousand Oaks, CA: Sage.

Jabine, Thomas B., Miron L. Straf, Judith M. Tanur, and Roger Tourangeau, ed. 1984. *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, DC: National Academy Press.

Kudela, Martha S., Kerry Levin, Margaret Tseng, May Hum, Susie Lee, Ching Wong, Susie McNutt, and Dierdre Lawrence. 2004. *Tobacco Use Cessation Supplement to the Current Population Survey Chinese, Korean, and Vietnamese Translations: Results of Cognitive Testing*. Final Report submitted to the National Cancer Institute, Rockville, MD.

Loftus, Elizabeth F. 1984. "Protocol Analysis of Responses to Survey Recall Questions." In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, ed. Thomas B. Jabine, Miron L. Straf, Judith M. Tanur, and Roger Tourangeau. Washington, DC: National Academy Press.

McColl, Elaine. 2001. "Protocol for Cognitive Testing of Global Health Status Rating Items." Unpublished manuscript.

Miller, Kristen. 2002. "The Role of Social Location in Question Response: Rural Poor Experience Answering General Health Questions." Paper presented at the American Association for Public Opinion Research Conference held in St. Pete Beach, Florida, May 2002.

_____. 2005. "Q-Bank: Development of a Tested-Question Database." *2005 Proceedings of the Section on Government Statistics*, pp. 1352–9. Alexandria, VA: American Statistical Association.

Nisbett, Ronald E., and Timothy D. Wilson. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84:231–59.

O'Muircheartaigh, Colm. 1999. "CASM: Successes, Failures, and Potential." In *Cognition and Survey Research*, ed. Monroe G. Sirken, Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur, and Roger Tourangeau. New York: Wiley.

Oksenberg, Lois, Charles F. Cannell, and Graham Kalton. 1991. "New Strategies for Pretesting Survey Questions." *Journal of Official Statistics* 7:349–65.

Presser, Stanley. 1989. "Pretesting: A Neglected Aspect of Survey Research.'" In *Conference Proceedings: Health Survey Research Methods*, ed. Floyd J. Fowler, Rockville, MD: National Center for Health Services Research.

Presser, Stanley, and Johnny Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" In *Sociological Methodology*, ed. Peter V. Marsden. Vol. 24, pp. 73–104. Washington, DC: American Sociological Association.

Presser, Stanley, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Eleanor Singer. 2004. "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly* 68(1): 109–30.

Presser, Stanley, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. 2004. *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: John Wiley and Sons.

Redline, Cleo, Richard Smiley, Meredith Lee, Theresa DeMaio, and Don Dillman, D. 1998. "Beyond Concurrent Interviews: An Evaluation of Cognitive Interviewing Techniques for Self-Administered Questionnaires." *1998 Proceedings of the Section on Survey Research Methods*, pp. 900–5, Alexandria, VA: American Statistical Association.

Royston, Patricia N. 1989. "Using Intensive Interviews to Evaluate Questions." In *Conference Proceedings: Health Survey Research Methods*, ed. Floyd J. Fowler, Rockville, MD: National Center for Health Services Research.

Royston, Patricia N., and Deborah Bercini. 1987. "Questionnaire Design Research in a Laboratory Setting: Results of Testing Cancer Risk Factor Questions." *1987 Proceedings of the Section on Survey Research Methods*, pp. 829–33. Alexandria, VA: American Statistical Association.

Royston, Patricia N., Deborah Bercini, Monroe Sirken, and David Mingay, D. 1986. "Questionnaire Design Research Laboratory." *1986 Proceedings of the Section on Survey Research Methods*, pp. 703–7, Alexandria, VA: American Statistical Association.

Russo, J. Edward, Eric J. Johnson, and Debra L. Stephens. 1989. "The Validity of Verbal Protocols." *Memory and Cognition* 17:759–69.

Schechter, Susan, Paul Beatty, and Gordon B. Willis. 1998. "Asking Survey Respondents About Health Status: Judgment and Response Issues." In *Cognition, Aging, and Self-Reports*, ed. Norbert Schwarz, Denise Park, Barbel Knauper, and Seymour Sudman. Philadelphia, PA: Psychology Press.

Schechter, Susan, Johnny Blair, and Janet Vande Hey. 1996. "Conducting Cognitive Interviews to Test Self-Administered and Telephone Surveys: Which Methods Should We Use?" *1996 Proceedings of the Section on Survey Research Methods*, pp. 10–7. Alexandria, VA: American Statistical Association.

Sirken, Monroe, and Susan Schecheter. 1999. "Interdisciplinary Survey Methods Research." In *Cognition and Survey Research*, ed. Monroe G. Sirken, Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur, and Roger Tourangeau. New York: Wiley.

Strauss, Anselm, and Juliet Corbin. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Newbury Park, CA: Sage.

Streett, Anitra Rustemeyer. 1983. "Investigating Respondents' Interpretations of Survey Questions." In *Approaches to Developing Questionnaires*, ed. Theresa J DeMaio. Statistical Policy Working Paper 10. Washington DC: Statistical Policy Office, U.S. Office of Management and Budget.

Tucker, Clyde. 1997. "Methodological Issues Surrounding the Application of Cognitive Psychology in Survey Research." *Bulletin de Methodologie Sociologique* 11:67–92.

van der Veer, Kees, Tony Hak, and Harrie Jansen. 2002. "The Three-Step Test Interview (TSTI): An Observational Instrument for Pre-testing Self-Completion Questionnaires." Paper presented

at the Questionnaire Development, Evaluation, and Testing Conference in Charleston, South Carolina.

van Someren, Maarten W., Yvonne F. Barnard, and Jacobijn A.C. Sandberg. 1994. *The Think-Aloud Method: A Practical Guide to Modelling Cognitive Processes*. San Diego, CA: Academic Press.

Weiss, Robert S. 1994. *Learning from Strangers: The Art and Method of Qualitative Interviewing Studies*. New York: The Free Press.

Willis, Gordon. 1994. "Cognitive Interviewing and Questionnaire Design: A Training Manual." Cognitive Methods Staff Working Paper Series, No. 7. Hyattsville, MD: National Center for Health Statistics.

_____. 2004. "Cognitive Interviewing Revisited: A Useful Technique, in Theory?" In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

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

Willis, Gordon, Theresa 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*, ed. Monroe G. Sirken, Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur, and Roger Tourangeau. New York: Wiley.

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

Willis, Gordon, and Susan Schechter. 1997. "Evaluation of Cognitive Interviewing Techniques: Do the Results Generalize to the Field?" *Bulletin de Methodologie Sociologique* 11:40–66.

Wilson, Timothy D., Suzanne J. LaFleur, and D. Eric Anderson. 1996. "The Validity and Consequences of Verbal Reports About Attitudes" In *Answering Questions: Methodology for Determining Cognitive Processes in Survey Research*, ed. Norbert Schwarz and Seymour Sudman. San Francisco, CA: Jossey-Bass.