Aspects of data quality in cognitive interviews: The case of verbal reports

Frederick G. Conrad
University of Michigan

Johnny Blair
Abt Associates

Abstract

Cognitive interview techniques are constructed from a menu of laboratory procedures leading to many disparate techniques. However, one common thread across techniques is that they all produce verbal reports. It stands to reason that different versions of cognitive interviewing produce data and decisions that vary in their quality, but very little empirical evaluation has been conducted.

We propose a research agenda for evaluating the quality of information produced by cognitive interview techniques. We propose that problem detection and, ultimately, problem repair are the fundamental purposes of the method. The quality of each should be assessed through standalone experiments that measure reliability and validity of potential problems and effectiveness of revisions in eliminating recurrence of problems.

We then illustrate the kind of research we advocate with a case study that compares quality of the verbal reports in two cognitive interview techniques. One technique represents the practices of experienced cognitive interviewers. The other technique constrains interviewer probing to explicit indications of problems in respondents' verbal reports. The results suggest that, in cognitive interviewing in general, verbal reports about answering survey questions are difficult to interpret consistently, raising concerns about their credibility. They further suggest that constraining probes to specific respondent indications of problems leads to fewer but more reliably identified problems.

XX.1 Introduction

Cognitive interviewing has been used extensively for over 15 years to pretest questionnaires. Practitioners use cognitive interview findings to decide whether or not to revise questions in many important survey instruments and, often, how to revise them. Considering the weight given to the results of cognitive interviews, evaluations of the method have been surprisingly few. The majority of studies that have been conducted compare a single version of cognitive interviewing to other pretest methods (e.g. Lessler, Tourangeau & Salter, 1989, Fowler & Roman, 1992; Presser & Blair, 1994; Rothgeb, Willis & Forsyth, 2001; see Tourangeau, Rips & Rasinski,. 2000, pp 331-332, for a review). While a valuable beginning, the results of such studies can only apply to the particular cognitive interviewing technique investigated.

Cognitive interviewing is actually a generic designation for a variety of loosely related procedures. Different practitioners combine these procedures in various ways to produce alternative cognitive interview techniques (Willis, DeMaio & Harris-Kojetin,. 1999). Because almost no research has been published that compares these different techniques, we know little more about their performance than when cognitive interviewing was introduced. Perhaps one reason there has not been much progress in this area is the lack of an overarching approach to such research. While Willis et al. (1999) raise fundamental issues for the evaluation of cognitive interviewing, most empirical studies have been carried out as the opportunities present themselves rather than by design.

In the current chapter, we propose a research agenda to specify some aspects of cognitive interviewing that need inquiry, and provide some suggestions as to how that inquiry might be undertaken. The goal of such an agenda is not to produce prescriptions for the practice of cognitive interviewing, although such prescriptions may come out of the research we are recommending. Our point is that evaluation of cognitive interviewing is in such an early stage that progress will be most likely if the issues warranting study and the approach to this work follow a single framework. In the second section of this paper, we specify such an agenda. In the third section, we report a case study that begins to address some of the issues on the agenda, namely those that concern the quality of verbal reports in cognitive interviews.

XX.1.1 Characteristics of a research agenda.

We propose that research on cognitive interviewing should be both empirical and theoretically grounded. In particular, methodological researchers will only ask certain questions if they consider the theory about how particular techniques should perform. For example, a technique that depends mainly on think aloud procedures will not be good at detecting difficulties and errors in recall because it is generally known that information about retrieval processes is not available to respondents (e.g. Ericcson & Simon, 1993). Second, we propose that examining the *components* that comprise particular cognitive interviewing techniques will provide more information about why particular techniques perform as they do than will examining the technique as a whole. How do different kinds of instructions to respondents affect the outcome of cognitive interviewing? How does the information elicited by generic think aloud procedures, in which the interviewer plays a relatively passive role, compare to the information provided in response to direct probes or in respondent paraphrasing (Willis, et al. 1999)? Are some respondent tasks better than others for explaining why a problem occurred or suggesting how to repair the question? And so on.

The common thread that connects cognitive interview techniques is that they all produce verbal reports. A useful research agenda, then, should include, though not be limited to, an assessment of the verbal reports produced by different techniques. We turn now to the theory of verbal reports and their use in survey pretesting.

XX.1.2 Verbal report techniques

The lynchpin of cognitive interviewing is people's ability to provide verbal reports about their thinking when answering questions (Lessler, at al., 1989; Tourangeau, et al., 2000; Willis, et al, 1999). The reports are elicited in different ways by interviewers and provided in different ways by respondents. For example, interviewers may simply ask respondents to report their thinking or may ask specific probe questions. These probes may be crafted prior to the interview or in response to respondent reports (Willis et al., 1999). Respondents may provide the reports concurrently, i.e. while formulating an answer, or retrospectively, i.e. after providing an answer. In all of these cases, the assumption of practitioners is that respondents are able to accurately verbalize some of the thinking that underlies their answers.

Ericcson and Simon (1993) pioneered the modern use of verbal reports, providing a theoretical account of how the method works, including its limitations. They have been concerned primarily with thinking aloud in which laboratory participants report on their thinking in a relatively undirected way, i.e. the experimenter may prompt the participant to keep talking but does not probe with substantive questions. The key component of Ericcson and Simon's theory is that people can only report on thinking to which they have access. By this view, they should be able to accurately report on processes that engage working memory in a series of discrete mental steps. For example when planning a move in chess, people must anticipate a series of moves and consequences, temporarily storing each one in working memory. In the survey response context, answering a behavioral frequency question by recalling specific events and counting each one involves similar temporary storage and thus should be amenable to verbal report techniques (e.g. Bickart & Felcher, 1996; Conrad, Brown & Cashman, 1998). Respondents are less able to report on questions that do not have this character such as answering "why" questions (Wilson, LaFleure & Samuels, 1996) or those soliciting a preference (Wilson & Schooler, 1991).

Despite the popularity of verbal report data for studying high-level cognition, the method has been controversial. The crux of the controversy surrounds the observation that thinking aloud can degrade the process about which people are reporting, a so called reactive effect (e.g. Russo, Johnson, & Stephens, 1989; Schooler, Ohlsson & Brooks, 1993). While in most reactive situations, thinking aloud degrades the task being studied, it has also been shown to improve its performance (Russo et al., 1989). If verbal reports degrade survey respondents' question answering performance in cognitive interviews, this could give the appearance of response difficulties where, in the absence of verbal reports, there actually are none. If verbal reports improve the question answering process, they could mask problems that would otherwise be present. Taken together, these findings suggest that verbal reports are fragile sources of data, sometimes valid but sometimes not, sometimes independent of the process being reported but sometimes not.

In most studies that have evaluated cognitive interviewing the typical measures are the number of problems and the type of problems. While these measures may help practitioners choose among different pretesting methods for particular purposes, they do not help assess the *quality of the information* about problems produced in cognitive interviews. Implicit in much of this research is the assumption that verbal reports – the primary information about problems – are high quality, i.e. veridical reflections of response problems. But these reports have seldom been treated as data in themselves and evaluated accordingly. This differs from how such reports are treated by cognitive psychologists (e.g. Ericsson & Simon, 1993) where the quality of verbal reports is routinely evaluated, for example, by computing agreement among coders.

While Willis and Schechter (1997) did directly test the quality of cognitive interview results, their focus was not on the verbal reports themselves, but on the use to which those reports were put. The authors measured the accuracy of cognitive interview results for five questions in predicting problems and the absence of problems in several field tests. The authors assessed these predictions by comparing the response distributions of two samples of respondents. One sample was presented originally worded questions and the other was presented questions revised on the basis of cognitive interviews. The predictions were confirmed for four out of the five questions and partially for the fifth. While verbal reports were, presumably, the raw data that informed the revision process, it is impossible to know how much the revisions also benefited from the skill of the designer(s). We advocate disentangling these issues and present a case study in section 3 that does this by addressing just the quality of verbal reports. The point for now is that verbal reports are data whose quality can be assessed just as the quality of data at other points in the survey process can be evaluated (e.g. Dippo, 1997; Esposito & Rothgeb, 1997).

XX.2 Research agenda

In this section, we sketch the broad features of a research agenda (see Figure 1). In general, we advocate conducting stand-alone experiments that directly compare different cognitive interviewing techniques. In order to compare techniques, they must each be well defined and have reasonably well specified objectives. By defining techniques in terms of their components (e.g. the type of probes permitted by the interviewer, the instructions to the respondent, the types of tasks required of the respondent, etc.) it is easier to design experiments that can explain differences in outcome between techniques. If only one component differs between techniques, the explanation for different results is likely to lie here. Existing techniques may differ on just one or two components and can be compared as is; alternatively a technique can be constructed by substituting a new component for one in an existing technique. Being clear on the objectives of a technique is important for assessing its success. For example, if the point is to detect as many potential problems as possible, one would apply a different set of criteria than if the point is to detect only the most severe problems[1].

Problem detection and problem repair are the essential objectives of pretesting. A cognitive interview technique is efficient if it is cost-effective in identifying question problems and providing useful information for their revision. Since detection and repair can be assessed in multiple ways, one concern is the choice of appropriate measures. Where possible, we advocate the use of multiple measures. Certainly for problem detection, one might count numbers of problems found, the nature of the detected problems, how reliable a technique is in uncovering problems and to what extent those problems are valid, i.e. occur in field settings, adversely affect respondent answers, etc. Other measures may also be useful.

Insert Figure 1 about here

The process of problem repair is harder to assess, since its effectiveness depends greatly on the skills of the personnel doing the revision. Certainly question repair is more likely to

[1] In the latter example, one would have to operationalize the measure of severity, which includes more than merely counting the number of problems.

succeed if it is based on information about why a problem occurred and the conditions under which it is most likely to occur. The main criterion in assessing question repair is whether or not a problem recurs in subsequent testing, either with cognitive interviews in the laboratory or in a field setting using other problem detection techniques.

In the remainder of section 2, we provide more detail about the kind of research that we believe is necessary in order to understand and, ultimately, improve cognitive interviewing techniques. In certain instances, we note that both the techniques studied and the methods used to study them may differ from common practice. This is a natural consequence of *studying* as opposed to *using* a method.

XX.2.1 Defining a cognitive interview technique

Despite variation in current practice (e.g. Blair & Presser, 1994; Willis et al., 1999), cognitive interview techniques share a few basic features that need to be covered in a definition. The interviews involve a one-on-one interaction between an interviewer and a respondent. There may be a protocol for interviewers to follow and, if there is, the degree to which interviewers are expected to adhere to it may vary. Prior to the interview the interviewer typically gives the respondent a general description of what will happen in the interaction; and the interviewer has decided how she will conduct the interview. What actually takes place in the interview may deviate from these intentions.

In the interview, respondents answer questions and produce other verbal reports. The interviewer asks the questions, probes and selectively follows up on what the respondent says. Additionally, the interviewer can clarify or amend instructions to the respondent. Finally, the interviewer can summarize her understanding of a verbal report. The substance of the interview is a discourse between interviewer and respondent, which is typically either summarized in notes or recorded verbatim.

A definition should answer three basic questions:

- What are the interviewer and respondent instructed to do?
- What are the data from the cognitive interview?
- How are these data analyzed?

XX.2.1.1 Interviewer and respondent instructions. Interviewers can instruct respondents to report their thinking either concurrently or retrospectively; or respondents can be instructed simply to answer the questions and respond to probes. Sometimes the respondent completes the interview without comment or interruption, after which the interviewer probes or asks other debriefing questions.

The degree to which the interviewer takes an active role can vary. The most passive role is to read the questions, and possibly prompt the respondent to keep thinking aloud, leaving it to him to produce verbal reports in accordance with the initial instructions. A more active role can involve myriad behaviors. For example, interviewers can administer scripted probes prepared prior to the interview which can be general and apply to any question, e.g. “In your own words what does this question mean to you?” Probes can also be question-specific, such as “What do you think the phrase ‘health care professional’ means in this question?” Alternatively, probes can be improvised during the interview in response to something the respondent says, or based on an idea that occurs to the interviewer. Combinations of these options are also used.

Information about respondent instructions can be obtained from a written script of instructions, the transcript of the interviewer’s introduction to the session, or the interview transcript. Since instructions may be changed or supplemented during the interview, an examination of the interview transcript is recommended.

XX.2.1.2 Cognitive interview data. Cognitive interviews produce verbal reports and we consider these to be the raw data for later analysis. The response processes that these reports describe are affected by the instructions to respondents and by the interviewer behaviors. As far as we are aware, little attention has been given to what actually constitutes verbal reports in cognitive interviews. Verbal reports certainly include respondent think-aloud statements and answers to probes (Willis et al., 1999). They probably, but not necessarily, include other respondent remarks. But should verbal reports include interviewer statements, such as recapitulation by the interviewer of her understanding about what a respondent said, or the interviewer's summary of a potential problem? When the interviewer takes a very passive role, this is not an issue. The more active the interviewer role, the more ambiguity there is about what comprises the "verbal reports."

The definition of verbal reports is not typically a concern in everyday *use* of cognitive interviews and, in fact, may not be essential. As Tourangeau et al. (2000), concurring with Conrad and Blair (1996), note "the conclusions drawn from cognitive interviews are only loosely constrained by the actual data they produce [p. 333]." However, in *methodological* research such definition is essential; otherwise, it is hard to know what information is produced in the interviews and what information was produced in the accompanying processes, e.g. discussion by the research team.

From our perspective, methodological research on cognitive interview techniques benefits from analyses that separate the process of eliciting verbal reports (data collection) from their interpretation (data analysis). This approach permits assessing how well suited a technique is to each task. Again we note that this recommendation may depart from common pretest procedure; but our goal is to promote evaluation of the method, not make prescriptions for its practice. Eliciting verbal reports, interpreting them, and, finally, applying those interpretations to question revision each involve different processes. A careful analysis of a technique will isolate one process from the others. It may well be that two techniques use identical processes to elicit reports– e.g. classical think aloud procedures. But interpretation of those reports may differ, e.g. one technique may rely on interviewer notes or even recollections of what took place in the interview, while the second method employs some sort of detailed analysis of interview transcripts.

Similarly, in one technique, question revision might be based mainly on the interviewer's judgment about the reason that a question caused problems for some respondents; another technique might rely more on follow-up probes to try to get respondents to articulate reasons for problems they appeared to have. Clearly, there are many possible approaches available to the cognitive interviewer/analyst. It seems prudent, whenever possible, to disentangle the effects of such alternative approaches.

XX.2.1.3 Data analysis. From our perspective, there are at least two stages involved in analyzing verbal reports for methodological purposes. First, the verbal reports have to be interpreted and, where problems exist, coded into problem categories. If the verbal reports are not classified in some way then evaluation is restricted to the verbatim content of the reports. We note that the coding of problems may not be necessary, or even useful, in practice but it is necessary for evaluation purposes. Second, the coded reports are counted and the tallies for techniques are compared.

The methodologist should design the specific codes to distinguish and capture the kinds of information of interest in the particular study. Our focus below is on problems so we have primarily coded problems. Other methodologists might wish to study the connection between verbal reports and possible solutions to problems, so they might use codes for potential repairs. The coding scheme should be exhaustive and well defined. That is, it should be possible to assign any problem (or whatever the topic of interest) to a code. We advocate assigning a problem to one and only one code. Clearly, a particular verbal report may indicate multiple problems – and these should be coded individually – but a single problem should be uniquely classified so that it is tallied just once.

XX.2.2 What is the technique intended to accomplish?

The cognitive interview goals used by practitioners vary considerably (Blair & Presser, 1993 ; Willis et al., 1999; Tourangeau et al., 2000). In addition to generally uncovering question flaws, some practitioners may want use the method primarily to confirm their intuitions about possible question problems, while other practitioners may seek information from cognitive interviews to aid problem repair. Still others may wish to determine whether questions are problematic for subgroups of respondents with certain demographic or behavioral characteristics. And so on.

These purposes and most others depend on problem identification. If a technique is weak on problem identification, it will not (at least on its own) provide much value in achieving these other goals. A minimum quality requirement is for cognitive interviews to produce data that credibly identify problems. We take problem identification to be the logical starting point for research to evaluate cognitive interviewing techniques.

XX.2.3 How is the success of a technique measured?

XX.2.3.1 Classifying and counting problems. Problem classification and problem counting are closely related. First, if the goal is to count unique problems or recurrences of the same problem, then some description of the type of problem is required to distinguish one problem from another. Second, if the classification scheme is designed to be exhaustive, then the list of classes themselves will aid in identifying a verbal report's evidence of a problem. Thus, it is unsurprising that many of the attempts at formal analysis of cognitive interview techniques involve problem classification schemes (See Tourangeau et al. 2000, pp. 327-328, for a discussion of many of the coding schemes that have been used.).

XX.2.3.2 Thresholds for problem acceptance. A problem report can be accepted based solely on the judgment of the interviewer. But this is not the only possible criterion. In summarizing 'best practices' in cognitive interviewing Snijkers (2002) notes that frequently the interviewer and someone who observed the interview meet to discuss the interview results. But he does not address how possible differences in their assessments are resolved. The implication is that two listeners can lead to better problem detection than just one.

A minimum requirement for treating a verbal report as evidence of a problem is that at least one interviewer/analyst concludes that a *problem* exists. However, it may be of methodological interest to examine the impact of higher thresholds, i.e. agreement among more than one judge. One issue on the proposed research agenda is to examine how the number and types of problems change when different amounts of agreement between analysts are required in order to accept the evidence.

XX.2.3.3 Reliability and validity in problem detection. Whether agreement is measured between two or more judges, it is a clear way to assess the reliability of problem detection. Do different judges reviewing the same verbal report agree on whether it suggests a problem or not? If they agree a problem exists, do they agree on the type of problem? Another way to assess the reliability of a technique is to compare the problems found on multiple administrations (or tests) of the method. By this view, a cognitive interviewing technique is successful to the extent that it turns up similar problems across different interviews.

To help explain why particular techniques are more or less reliable in either sense, one can examine the interaction between interviewers and respondents. For example, it may be that different types of probes lead to answers that are more or less reliably interpreted.

Similarly, validity can be defined in several ways, each of which entails a different view of what makes a detected problem "real." In one view, problems are real only if they demonstrably lead to incorrect data in field data collection. By extension, a reported problem is valid the more probable its occurrence in any given interview. A problem will rarely affect all respondents; most will affect only some portion of respondents. But if a potential problem detected in a cognitive interview does not affect any respondents in field administration of the questionnaire, it cannot be considered valid. Yet another sense of problem validity is severity – how large is the measurement error produced by the problem.

XX.2.4 Stand-alone experiments

Methods research on cognitive interviews has been of two types: "piggybacking" the evaluation of a technique onto a production survey (e.g. Willis & Schechter, 1997) or conducting a stand-alone experiment (e.g. Presser & Blair, 1994), which can be a laboratory experiment and/or a field experiment. We favor the latter approach in order to exercise control over the variables of interest. In stand-alone experiments, the researcher can often determine the sample size. As has been advocated in the evaluation of usability testing methods in Human-Computer Interaction (see Gray & Salzman, 1997, pp. 243-244) we endorse using larger samples than are typical in production pretesting. In the case study described below, 8 interviewers each conducted 5 cognitive interviews for a total of 40 interviews – probably four or more times the number that is typical in production use of the method. This was sufficient for us to carry out some analyses but not all; the ideal sample size really depends on the questions one is asking. Larger samples not only increase the power of the subsequent analysis, but also provide better estimation of the probability that respondents will experience a particular problem.

While we recognize the central role of cognitive interview data in problem repair, we think evaluations of cognitive interview techniques should separately assess the quality of the data produced by the techniques and the use of those data in question revision.

XX.3. Case Study

In the following section we describe a study in which we examined two variants of cognitive interviewing. The study is one attempt to gather information about several, though by no means all, of the items on the research agenda presented in the previous section. We report the study here primarily to illustrate the kind of evaluation we advocate, rather than as the final word on either of the versions we examined or on cognitive interviewing in general. In particular, the study is an example of a stand-alone experiment, carried out for methodological rather than survey production purposes.

The study produced thousands of observations of quantitative and qualitative data. Based on these data, the study illustrates the use of agreement measures to assess the interpretation of verbal reports, though it does not make use of validity measures. In addition, it illustrates the use of interaction analysis to explore the types of probes used and their relation to the kinds of problems identified. Finally, the study evaluates two versions of cognitive interviewing to test the effects of varying particular features of the general method rather than to evaluate these versions per se. One could vary the features in other ways or vary other features.

Because this was a methodological study and not a production application of cognitive interviewing, certain aspects of the way we used the method may depart from its use in production settings. For example, the number of interviews (40 total) may be larger than is common and the way that problems were reported (written reports for each interview and codes in a problem taxonomy) may differ from what seems to be typical (written reports that summarize across interviews).

We recruited eight interviewers and asked each to conduct five cognitive interviews with a questionnaire constructed from several draft instruments created by clients of the Survey Research Center at the University of Maryland. The questionnaire contained 49 substantive questions about half of which were factual and half opinion questions. The topics included nutrition, health care, AIDS, general social issues and computer use. The interviewers were told that they were participating in a methodological study sponsored by a Federal agency.

Four of the interviewers were experienced practitioners of cognitive interviewing. Each had more than five years of experience at different organizations within the Federal government and private sector survey 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. This group conducted cognitive interviews using whatever method they ordinarily use. We will refer to the procedures they used as "conventional" cognitive interviewing.

The remaining four interviewers were less experienced with pretesting questionnaires though all four worked in survey organizations, in either the academic, commercial or Federal sector. Two of the four had some experience with production cognitive interviewing and the other two had been exposed to the theory behind the method and had completed relevant class exercises as part of their masters level training in survey methodology. In contrast to the conventional cognitive interviews, three of whom held doctoral degrees, three of these interviewers held only bachelors degrees and one held a masters degree. This level of experience and education seemed typical to us of junior staff in survey research centers – staff that typically do not conduct cognitive interviews. This group of interviewers was trained to use a version of cognitive interviewing in which the types of probes and the circumstances of their use were restricted. One consequence of restricting the set of probing conditions, relative to conventional cognitive interviewing, was to simplify the interviewers' task by reducing the amount of experience and judgment required to know when and how to probe. We refer to the current procedure as the "conditional probe" technique.

Ideally, we would have crossed the factors of technique and experience/education to disentangle any effects of one factor on the other. This would have meant that, in addition to the two groups who differed on both factors, we would have instructed inexperienced interviewers to carry out conventional cognitive interviews and trained experienced interviewers to follow the conditional probe method, thus creating a 2 x 2 design. However it was not feasible to test these additional groups. First it seemed unlikely to us that experienced cognitive interviewers could "unlearn" their regular technique, developed over many years, for the duration of the study[2]. Interference from the old technique on the new one would have made the results hard to interpret. Moreover, they would not have been highly experienced with this particular method[3]. Second, because there is no industry wide consensus about what constitutes cognitive interviewing, we were not able to define the conventional method well enough to train inexperienced interviewers in its use. Even if we had been able to train them in conventional cognitive interviewing they would not have had the education or years of experience that the conventional cognitive interviewers had.

XX.3.1 The particular cognitive interviewing technique(s).

XX.3.1.1 Instructions to interviewers. Because of the lack of consensus about conventional practice, we asked the experienced practitioners to conduct cognitive interviews as they ordinarily do, allowing them to define "the method" – both the procedure for conducting the interviews and the criteria for what constituted a problem. We did not provide them with an interview protocol or instructions for respondents and we did not require that they examine the questionnaire ahead of time – if they did not ordinarily do so. We asked them to prepare written reports of problems in each question of each interview. We did not define "problem" for them but instructed them use whatever criteria they used in ordinary practice. We required reporting at this level (individual questions on individual interviews) in order to compare the interviewers' judgments about each particular verbal report to those of coders listening to the same verbal

[2] We acknowledge that in some organizations, experienced cognitive interviewers are routinely asked to modify their practice. However, the success with which they do this is an empirical question about which we have little relevant data.

[3] Another approach to increasing the expertise of interviewers using the conditional probe method would have been to give inexperienced interviewers substantial practice with the conditional probe method prior to the study. However, it would have been impractical to give them several years of experience, which is what would have been required to match their experience to that of the conventional interviewers.

report.[4] According to their written reports and a subsequent debriefing, these interviewers 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 already discovered problems than novel ones. While we treat this as a single "conventional" method, we recognize that each interviewer might approach the data collection task somewhat differently.

The conditional probe interviewers were introduced to the method in a two-day training session. They were instructed in both how to conduct the interviews and in how to classify problems identified in the interviews. For the interviewing technique, they were instructed to solicit concurrent verbal reports from respondents and to focus their probing on behavioral evidence of problems in those reports[5]. They were instructed to intervene 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). 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?"). Other than probing under these conditions, the interviewers were not to play an active role. 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 essentials of the probing technique.

The restrictions on probing were motivated by ideas and findings in the psychological literature on verbal report methods. First, people can only provide accurate verbal reports about

[4] One interviewer indicated that this reporting procedure departed from her more typical practice of summarizing across interviews. The departure concerned her because it treated each interview in isolation; she indicated that she typically varied her interviews on the basis of what she had learned in previous ones.

the content of their working memory (Ericsson & Simon, 1993), and thus may sometimes legitimately have nothing to report. This is especially likely when respondents retrieve the information on which their answers are based from long term memory. This type of retrieval usually occurs automatically (*e.g.* Shiffrin and Schneider, 1977) in the sense that people do not actively control the process and thus are not aware of its details. Probing respondents for more details when none are available could lead to rationalized or embellished reports. Under these circumstances, the conditional probe interviewers were instructed to do nothing[6]. However, if a respondent is able to report something and the report contains a hint of a problem then, presumably, reportable information exists which the respondent has not clearly articulated. In this case, probing should clarify the initial report without encouraging respondents to embellish it.

A further impetus to experiment with restricted probing was the set of findings about reactivity mentioned in the introduction, i.e. the observation that thinking aloud can distort the process about which respondents are reporting. Reactive effects seem particularly likely when the response task is difficult because pressing respondents to report may demand mental resources that would otherwise have been devoted to responding. By instructing interviewers to remain silent when respondents give no suggestion of problems, we believed the interviewers would be less likely to contribute to reactive effects.

In addition to practice with the probing technique, the conditional probe interviewers were taught to identify problems through (1) an introduction to well known types of problems,

[5] The written materials used to train the conditional probe interviewers are available from the authors.
[6] Note that the inability to report on a process does not mean it is free of problems. It simply means we do not have any data about the process.

e.g. double-barreled questions, (2) discussion of 12 detailed definitions for each category in a problem taxonomy they would later use, and (3) practice identifying problems in audiotaped mock cognitive interviews. The mock interviews illustrated respondent behaviors for which probing was and was not appropriate (see Conrad, Blair & Tracy, 1999). The instructor provided feedback including pointing out missed problems and determined when the interviewers had grasped the problem definitions.

XX.3.1.2 Instructions to Respondents. The conventional cognitive interviewers did not indicate *a prioi* how they typically instruct respondents so we relied on the interview transcripts to determine what they actually did. The transcripts showed substantial variation in wording and content between these interviewers. Two of the conventional interviewers said that although the questions would be administered as in a "real" interview, they were not as interested in the answers as in how the respondent came up with their answers. Two interviewers mentioned that respondents should tell the interviewer about any difficulties encountered. Three of the interviewers also gave some variant of think aloud instructions. All four conventional cognitive interviewers said that the purpose of the interview was to learn about comprehension problems before the survey went into the field.

The conditional probe interviewers were trained to provide think aloud instructions to respondents that closely followed those of Ericsson and Simon (1993, p. 376) but they were not given an exact script to present to respondents. In general they were instructed to encourage respondents to report everything that passed through their heads while answering each question and to do so without planning what to say after thinking. The interview transcripts confirmed that all four interviewers did this reasonably consistently.

XX.3.1.3 Data from the cognitive interviews. Each of the eight interviewers (four conventional, four conditional probe) conducted five interviews. All 40 interviews (20 per type of cognitive interviewing technique) were audio recorded. The conventional cognitive interviewers each wrote a narrative report listing the problems they identified in each administration of each question. We used this reporting format, instead of summarized reports, in order to measure agreement between interviewers and other analysts about the presence of problems in particular administrations of a question. Problems identified in the written reports were later classified into a taxonomy of problems (see Conrad & Blair, 1996) in what was essentially a transcription task. The main types of problems are lexical (primarily issues of word meaning), logical (both logical connectives like “and” and “or” as well as presupposition), temporal (primarily reference periods) and computational (a residual category including problems with memory and mental arithmetic). Two transcribers independently mapped the written problem reports to the problem taxonomy and then worked out any discrepancies together[7]. 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 an oral introduction to the taxonomy. The exercise was discussed with both transcribers until they seemed competent to carry out the task.

The conditional probe interviewers directly classified any problems they detected for a given administration of a question into the problem categories in the taxonomy. They were required to choose a single category for a particular problem but could code more than one

[7] We did not compute any sort of reliability measure for the transcription task primarily because it did not involve judgment about the nature of problems – this had already been done by the interviewers in their reports – but as a matter of translating the interviewers’ descriptions to those in the taxonomy. In addition, because the transcribers worked jointly to resolve differences, their work was not independent.

problem per question. They were provided with written definitions of the problem categories, an oral introduction, and were given coding practice until they were competent users of the taxonomy.

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

In addition to the interviewers' own judgments about the presence of problems, four coders coded the presence of problems in all 40 interviews using the same problem taxonomy. They participated in the same training as did the conditional probe interviewers on the definition of problems (not on the interviewing technique) and had classroom exposure to cognitive interviewing. This made it possible to measure agreement between each interviewer and each of the four coders as well as agreement between each pair of coders.

We use agreement measures here to assess the quality of the data on which suggested revisions are based. Coding agreement in the current study measures the quality of the information that serves as *input* to the decision about revision. If two coders do not agree on the presence of a problem or its identity, the respondent's verbal report is ambiguous and thus less definitive about the need for revision than one would hope. Of course, in practice, researchers mull over the results of cognitive interviews before deciding whether to revise particular questions and how to revise them, but their decision making is constrained by the information produced in the cognitive interview, i.e. it's hard to make a good decision based on murky information.

XX.3.3 Key Findings

XX.3.3.1 Number and type of problems. Because cognitive interviews have been found to be sensitive primarily to problems concerning comprehension (e.g. Presser & Blair, 1994), we would expect more lexical and logical than temporal and computational problems in the current study. This is in fact what we observed. Over all 40 cognitive interviews, interviewers identified .13 lexical and .11 logical problems per question versus .02 temporal and .04 computational problems per question. This serves as a general check that procedures were relatively similar to those used in other studies (if not in actual practice).

None of these patterns differed substantially between the two types of cognitive interviews. However, conventional cognitive interviewers reported 1.5 times as many potential problems as did the conditional probe interviewers. If more problems are detected with one technique than another, this could indicate that the larger number refers entirely to actual problems and the smaller number reflects missed problems. Alternatively, the larger number could include reported problems that are not actually problems ("false alarms") and the smaller number, therefore, would be the more accurate one. And the truth could be somewhere in between if one technique promotes false alarms and the other tends to miss problems. This could be most clearly disentangled if we had a validity measure, i.e. an objective measure of whether a reported was in fact a problem for that respondent. In the absence of such a measure, reliability provides some indication of how much stock to place in reports of problems.

XX.3.3.2 Agreement measures. Problem detection was surprisingly unreliable when measured by agreement between interviewers and coders about the presence of problems in the *same* verbal report. That is, when an interviewer and coder listened to the same verbal report, they often reached different conclusions about whether or not it indicated the presence of a problem. The average kappa score for all interviewer-coder pairs for the judgment that a particular question administration did or did not indicate a problem was only .31 ("fair" agreement according to Everitt & Haye, 1992, p.50). This low agreement rate cannot be attributed to the complexity of the coding system since the judgment about the presence or absence of a problem did not involve the specific categories in the coding system. In fact, agreement on the particular problem category in the taxonomy for those cases where an interviewer and coder agreed there was a problem was reliably higher, .43 ("moderate" agreement according to Everitt & Haye, 1992, p. 50), than their agreement that a problem simply was or was not present. However, even this score is disturbingly low considering, again, that the interviewers and coders were interpreting the same verbal reports, and considering that problem reports in cognitive interviews are used to justify changes to questions in influential surveys. At the very least, these low agreement scores suggest that verbal reports – the raw data from cognitive interviews – are often ambiguous.

Although agreement is low, it is reliably higher for conditional probe than conventional cognitive interviews. For the simple judgment about whether or not a problem was indicated by a particular verbal report, the kappa score for conditional probe interviews is .38 but only .24 for conventional cognitive interviews, a reliable difference (see Table XX.1). When there is agreement that a verbal report indicates a problem, agreement about the particular problem category shows a non-significant advantage for the conditional probe interviews, kappa = .47, over the conventional interviews, kappa = .39.

Insert TableXX.1 about here

One might argue that the general advantage for the conditional probe interviews is due to the greater similarity in experience and training between coders and conditional probe interviewers than between coders and conventional cognitive interviewers. However, if that were the case, the kappa scores would be lower for pairs of conventional cognitive interviewers and coders than for pairs of coders. But this was not the case. Average kappas were statistically equivalent for interviewer-coder and coder-coder pairs interpreting the conventional cognitive interviews (see Table XX.2 for inter-coder agreement scores).

Insert Table XX.2 about here

Irrespective of interview type, one might expect interviewer-coder agreement to be lower than coder-coder agreement because interviewers had more information available than did coders – audiotaped interviews in the case of coders versus audiotapes as well as interview notes and memories by the interviewers. Interviewers may have taken into account non-verbal information like respondents' facial expressions and gestures, not available to the coders. However, these differences in available information did not affect agreement. There was no difference in average kappas for interviewer-coder pairs and coder-coder pairs.

Pairwise agreement scores clearly indicate that these verbal reports were hard for different listeners to interpret in the same way. A further indication that verbal reports are inherently ambiguous is evident when we raise the threshold for accepting problems. In particular, the number of problems identified by at least one coder (though possibly more) is .42 problems per question. The number identified by at least two coders was only .09 problems per question. This figure drops even further to .02 problems per question when the threshold is at least 3 coders. And practically no problems, .004 problems per question, are detected by all four coders.

XX.3.3.3 Interviewer-respondent interaction. It is possible that the low agreement in interpreting verbal reports can be understood by examining the interaction between interviewers and respondents, since it is this interaction that produces the reports. At the very least, examining this interaction should help to document what actually happens in cognitive interviews. To address this, all 40 interviews were transcribed and each conversational turn was assigned a code to reflect its role in the interaction. These interaction codes should not be confused with problem codes: the interaction codes were assigned to each statement in the interview whereas problem codes were assigned to each question in the interview. The particular interaction codes for respondent turns included, among other things, potential indications of a problem (e.g. long pauses, disfluencies, and changed answers) and explicit respondent descriptions of a problem. Interviewer turns were coded, among other things, as probes about a problem that was expressed – at least potentially – in an earlier respondent turn and probes about a problem that was not expressed – even potentially – in a prior respondent turn.

Over all of the interviews, conventional cognitive interviewers probed 4.2 times as often as conditional probe interviewers. However, conditional probe interviewers probed about an explicit respondent utterance that potentially or explicitly indicated a problem 4.8 times as often

(61% versus 13%) as conventional cognitive interviewers. In the following example, a conventional probe interviewer asks about the respondent's silence[8].

I: Transportation of high-level radioactive wastes, means that state and local governments may have to spend money on things like emergency preparedness. Do you think that the federal government should reimburse the states and localities for costs like these?
R: Um: . I don't want to say no, because I think ultimately states:, should handle their own business, unless it's a federal issue . . .
I: I noticed you paused there for a second after I, I asked you that. Was there something unclear in *that question*?
R: *Um yeah* I would, unless I wasn't hearing the question, completely clearly, I, I wasn't sure WHO was producing the waste. Which that may have been the case I was thinking back to what the question asked, what you know who, who was producing the waste whether it was federal government or state.
I: Alright.

In contrast, in the following exchange, a conventional cognitive interviewer probes about the meaning of several terms in the question without any indication from the respondent that these terms are misinterpreted or are causing difficulty.

I: In general, do you think the police in your neighborhood treat cases of domestic violence seriously enough when they are called, or do you think the police are more likely to treat domestic violence as just a family matter?
R: [sighs] I don't, I haven't had much experience with the police in my area and I don't know anybody who's been abused, so I don't know how it's handled in my area.
I: Okay.
R: I know in some areas it's handled, differently, so.
I: Okay and how, how do you define "differently?"
R: I have heard of people getting slapped on the wrist for domestic violence and just let off the hook saying like "don't do this again." And I've also heard of cases where people have gone to jail for quite a while as a result of domestic abuse.
I: Okay. So tell me in your own words what you think THAT question is asking.
R: It basically seemed to be asking, um, "do the cops take it seriously or, or do they take it kind as leniently and none of their business."
I: Okay, so as a family matter, um, as, as "just a family matter", what does "just a family matter" mean, to you?

[8] 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.

R: It seems to me as something that, they don't think is in their authority to handle and also that it's not severe enough that it would warrant their involvement.
I: Okay.

Exchanges like the second that focused on possible meaning-related problems were quite frequent in the conventional cognitive interviews but relatively rare in the conditional probe interviews. In particular, 36% of conventional cognitive interviewers' probes concerned respondents' understanding of specific terms when respondents had not given any verbal evidence of misunderstanding these terms. Conditional probe interviewers administered this type of probe under these circumstances only 5% of the time that they probed.

Such differences in the type of interaction could be related to differences in levels of agreement for the two types of interviews. 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; the initial utterance either did or did not indicate a problem. However, probes that are not clearly tied to something the respondent said earlier 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 up until this point. If the respondent's answer to this probe indicates possible confusion, it may be hard for listeners to evaluate this. Has the interviewer uncovered an actual problem or introduced one? For example, the respondent may have understood the question well enough in context to accurately answer the question but not well enough to provide the relatively formal definition that the probe requests. Different listeners may hear such an exchange differently, which would lead to low agreement.

XX.3.4 Conclusions from Case Study

We have presented this study primarily as an example of the kind of evaluation research we advocate in section 2. One methodological lesson from the case study is that, even though it was hard to find four experienced cognitive interviewers, future studies should involve even more interviewers. Without larger numbers of interviewers, idiosyncratic practices are a threat to the generalizations one can confidently draw. Another methodological issue concerned the lack of consensus about what current practice involves. We tried to overcome this by allowing the traditional interviewers to follow their ordinary practice, but it would have been preferable to know ahead of time that they were following a single, representative approach.

The first substantive conclusion is that the overall low agreement scores suggest that verbal reports in cognitive interviews (even when interviewers are constrained in their probing) often lend themselves to different interpretations. This is potentially of great concern when we consider that, based on cognitive interviews, designers change and decide not to change the content of questionnaires in major surveys that produce high profile statistics. If the information on which those decisions are based is inherently ambiguous, the decisions will be compromised, no matter how thoughtfully considered.

Similarly, the number of problems that are reported in a particular application of cognitive interviewing is greatly reduced by requiring identification by more than one analyst. This suggests that the number of problems produced in particular pretests can lead to different conclusions depending on what threshold is used.

Finally, conventional cognitive interviewers report more problems than conditional probe interviewers but they agree less often with coders about the presence of problems than do conditional probe interviewers. Conventional cognitive interviewers may be erring on the side of including questionable problems as actual problems, a strategy which may reduce the risk of missing actual problems but may also introduce new problems by leading to changes in questions that are not actually problematic.

XX.4 Future Work

In cognitive interviewing, a deceptively simple set of procedures – asking respondents to report what they are thinking while answering survey questions – sets in motion a complex series of mental and social processes that have gone largely unstudied. Yet the effectiveness of the method is certain to rest on the nature of these underlying processes. We have proposed an agenda for research that compares these processes between techniques. Our case study begins to address the processes involved in producing and interpreting verbal reports. Subsequent research on this topic might include reliability – in both senses mentioned above –across different types of questions and different types of probes. In addition, validity of problems found in cognitive interviews has received very little attention, in part because its measurement is elusive. We have suggested several possible measures though none is perfect. If it were known to what degree potential problems identified in cognitive interviews really are problems for respondents, this would enable practitioners to use cognitive interview results more wisely in revising questionnaires.

Similarly, little is known about the degree to which revising questions in response to cognitive interview results actually prevents the problems from recurring. This needs to be evaluated in laboratory as well as field settings. But more fundamentally, the revision process is a largely creative enterprise that may vary widely depending on who is involved and in what organization they work. By better examining how designers use the information from cognitive interviews to reword questions and redesign questionnaires, it becomes more feasible to codify their practices and disseminate this to students.

Finally, the kind of research we are proposing would be facilitated by certain procedural changes by practitioners. Most significant is greater vigilance by practitioners in defining their cognitive interviewing techniques. While a definition does not guarantee that the technique is actually used in a particular way, it provides a starting point for evaluation research. Techniques that are clearly defined make it possible to identify and then evaluate their key aspects and this increases the chances that the results are relevant and useful to practitioners.

XX.5 Acknowledgements

We thank Ed Blair, David Cantor, and Greg Claxton for their advice and assistance.5.

References

Blair, J. and Presser, S. 1993, "Survey procedures for conducting cognitive interviews to pretest questionnaires: A review of theory and practice." In *Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association*. Alexandria, VA: American Statistical Association.

Conrad, F. G., Brown, N. R. and Cashman, E. R. 1998, "Strategies for estimating behavioural frequency in survey interviews." *Memory,* 6: 339-366.

Conrad, F., Blair, J. and Tracy, E. 1999, "Verbal reports are data! A theoretical approach to cognitive interviews." In *Proceedings of the Federal Committee on Statistical Methodology Research Conference, Tuesday B Sessions*. Arlington, VA, pp. 11-20.

Dippo, C. S. 1997, "Survey measurement and process improvement: Concepts and integration." In Lyberg, L., Biemer, P., Collins, M., deLeeuw, E., Dippo, C., Schwarz, N., Trewin, D.

(eds.). *Survey Measurement and Process Quality.* New York: John Wiley & Sons, Inc., pp. 457-474.

Esposito, J. L. and Rothgeb, J. M. 1997. "Evaluating survey data: Making the transition from pretesting to quality assessment." In Lyberg, L., Biemer, P., Collins, M., deLeeuw, E., Dippo, C., Schwarz, N., Trewin, D. (eds.),. *Survey Measurement and Process Quality.* New York: John Wiley & Sons, Inc., pp. 541-572.

Ericsson, A, and Simon, H. 1993, *Protocol Analysis: Verbal Reports as Data* (2nd edition). Cambridge, MA: MIT Press.

Everitt, B. S. and Haye, D. F. 1992, "Talking About Statistics: A Psychologist's Guide to Data Analysis." New York: Halsted Press.

Forsyth, B. H., Lessler, J. T., and Hubbard, M. L. 1992, "Cognitive evaluation of the questionnaire." In Turner, C. F., Lessler, J. T., and Gfroerer, J. C. (eds.), *Survey Measurement of Drug Use: Methodological Studies*. Rockville, MD: U.S. Department of Health and Human Services, pp. 13-52.

Fowler, F. and Roman, A. 1992, "A study of approaches to survey question evaluation." Final report for CSMR, SRD division of the US Bureau of the Census, Washington, DC.

Gray, W. D. and Salzman, M. C. 1998, "Damaged merchandise? A review of experiments that compare usability evaluation methods." *Human Computer Interaction,* 13: 203-361.

Lessler, J. T., Tourangeau, R. and Salter, W. 1989, "Questionnaire design in the cognitive research laboratory." *Vital Health Statistics,* Series 6, No. 1.

Presser, S. and Blair, J. 1994, "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology*. In Marsden, P.V. (Ed.), *Sociological Methodology, 24.* Beverly Hills, CA: SAGE, pp. 73-104.

Rothgeb, J., Willis, G. and Forsyth, B. 2001, "Questionnaire pretesting techniques: Do different methods and different organizations produce different results?" In *Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association*. Alexandria, VA: American Statistical Association..

Russo, J., Johnson, E. and Stephens, D. 1989, "The validity of verbal protocols." *Memory and Cognition*, 17: 759-769.

Schooler, J. W., Ohlsson, S. and Brooks, K. 1993, Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122: *166-183.*

Shiffrin, R. and Schneider, W. 1997, "Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory." *Psychological Review*, 84: 127-190.

Snijkers, G. 2002, *Cognitive Laboratory Experienced on Pre-Testing Computerized Questionnaires and Data Quality*. Heerlen, Netherlands: Central Bureau of Statistics.

Tourangeau, R., Rips, L. and Rasinski, K. 2000, "An Introduction and a Point of View," *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Willis, G., DeMaio, T. and Harris-Kojetin 1999, "Is the bandwagon headed to the methodological promised land? Evaluating the validity of cognitive interviewing techniques." In Sirken, M, Herrmann, D., Schechter, S., Schwarz, N., Tanur, J., and Tourangeau, R. (eds.) *Cognition and Survey Research*. New York: Wiley, pp 133-153.

Willis. G. B. and Schechter, S. 1997, "Evaluation of cognitive interviewing techniques: Do the results generalize to the field?" *Bulletin de Methodologie Sociologique,* 55 (June): 40-66.

Wilson, T. and Schooler, J. 1991, "Thinking too much: Introspection can reduce the quality of preferences and decisions." *Journal of Personality and Social Psychology, 60*, 181-192.

Wilson, T., LaFleur, S. and Anderson, D. 1996, "The validity and consequences of verbal reports about attitudes." In Schwarz, N. and Sudman, S. (eds.) *Answering Questions: Methodology for Determining the Cognitive and Communicative Processes in Survey Research.* San Francisco: Jossey-Bass, pp. 91-114.

Table 1. Average kappa values for interviewer-coder pairs

	Conventional Cognitive Interviews	Conditional Probe Interviews	Difference
Is there a problem?	.24	.38	p = .001
If so, what type?	.39	.47	Not signif.

Table 2. Average kappa values for coder-coder pairs

	Conventional Cognitive Interviews	Conditional Probe Interviews	Difference
Is there a problem?	.27	.36	p = .077
If so, what type?	.36	.43	Not signif.

I Methodology for comparing techniques

- define each technique, including its components
  - e.g. instructions, respondent tasks, probing
- specify each technique’s objectives
  - problem detection
  - problem repair
- Design and conduct stand-alone experiments that compare techniques
  - Compare existing techniques
  - Create alternative techniques by varying one or more components.

II Data preparation:

- code verbal reports
- transcribe and code interviewer-respondent interactions

III Criteria for comparing techniques

- Problem detection
  - number of problems
  - types of problems
  - quality of verbal report data
    - reliability
    - validity
  - thresholds for problem acceptance
- Question repair
  - Reason for problem
  - Situations when the problem occurs
  - Effectiveness of question revision

Figure 1: Research agenda for the evaluation of cognitive interview techniques