INTERPRETING VERBAL REPORTS IN COGNITIVE INTERVIEWS: PROBES MATTER [1]

Frederick G. Conrad, Bureau of Labor Statistics
Johnny Blair, University of Maryland
Frederick G. Conrad, Bureau of Labor Statistics, Room 1950
2 Massachusetts Ave. NE; Washington, DC 20212

Key Words: cognitive interviewing; think aloud methods; evaluation of survey pretesting methods

INTRODUCTION

Collecting high quality survey data frequently requires refining the survey questionnaire on the basis of laboratory pretests. The pretesting method that seems to be most widely used is cognitive interviewing. By requiring respondents to think aloud as they answer (or just after they answer) draft questions, cognitive interviewing is intended to uncover problems with the questions that may compromise the quality of responses. Cognitive interviews are based on Ericsson and Simon’s theory of verbal protocols (e.g. Ericsson & Simon, 1993), though the application of these ideas to surveys is not straightforward; this may compromise the quality of cognitive interview results.

Ericsson and Simon (1993) developed verbal protocol techniques and related theory for purposes and situations that differ from those associated with cognitive interviewing (see Table 1). For example, Ericsson and Simon were primarily interested in how people solve problems. Problem solving tasks generally involve multiple discrete mental steps each of which people can report on. The survey response task that cognitive interviews are concerned with often involves a single mental step such as retrieving a fact or opinion. One consequence is that, in some cases, there is little for respondents to report on.

In the administration of Ericsson and Simon’s method, the experimenter plays a passive role, primarily prompting the participant to keep talking, i.e. to report on his or her thinking; otherwise the experimenter tends to remain silent. In cognitive interviews, the interviewer generally plays an active role, probing as needed to expose problems.

In Ericsson and Simon’s method, two or more independent coders routinely classify verbal reports. This not only reduces the volume of data from lots of words to a small number of codes but also makes it possible to check that the reports are reliably interpreted, i.e. that there is agreement. In cognitive interviews, the interviewer typically interprets the verbal reports, listing the problems that are evident in the reports in a written narrative.

Ericsson and Simon’s theory of verbal reports has been extensively evaluated in the mainstream psychology literature (e.g. Ericsson, 1975; Flaherty, 1974; Newell & Simon, 1972; Nisbett & Wilson, 1977; Russo, Johnson & Stephens, 1989; Schooler, Ohlsssson. & Brooks, 1993; Wilson & Schooler, 1991). Cognitive interviews have received relatively little formal evaluation (see Presser & Blair, 1994 as well as Rothgeb, Willis & Forsyth, 2001 for examples of evaluation studies). Because the results of evaluating Ericsson & Simon’s method are mixed and also because many psychologists are generally suspicious of introspection, the Ericsson & Simon method is controversial in psychology (e.g. Payne, 1994). In contrast, cognitive interviews are widely used and accepted among survey researchers (e.g. Demaio & Rothgeb, 1996; though see Wilson, LaFleur & Anderson, 1996, for a more skeptical appraisal).

Ericsson & Simon	*Cognitive Interviews*
Tasks involve multiple discrete steps	Response often involves one step
Experimenter is passive; prompts to keep talking	Interviewer plays active role; probes to expose problems
Verbal reports coded; reliability can be assessed	Interviewer typically lists problems in written narrative
Extensively evaluated	Not extensively evaluated
Controversial	Widely used and accepted

Table 1. Comparison of think aloud methods as used by Ericsson & Simon and in cognitive interviews.

[1] We thank Elizabeth Brooks, Greg Claxton, Cathryn Dippo, Jane Joseph, Nileeni Meegama, Michael Schober, Elena Tracy, Clyde Tucker and Sara Woldehanna. The material is based upon work supported, in part, by the Bureau of Labor Statistics. The opinions expressed are those of the authors and not the Bureau of Labor Statistics.

Considering the widespread use of cognitive interviewing, it is surprising that the method has received so little evaluation. There is virtually no published information about how reliably problems are detected by the method, that is, the degree to which different analysts agree that particular verbal reports indicate there is a problem. And there is virtually no published information about the validity of problems detected by the method, that is, the degree to which problems identified by cognitive interviews actually occur in field administration of the test questionnaire[2].

A concern for any evaluation of cognitive interviews is the prevalence of "false alarms" and "misses." In the first of these, the interviewer believes there is a problem when, in fact, there is not. This can occur when the interviewer presses the respondent to say something but there is little that the respondent can faithfully report. According to the theory of verbal protocols, a respondent would be particularly at risk for this when directly retrieving permanently stored information because such retrieval processes are not available to introspection (e.g. Ericsson & Simon, 1993, pp. 133-134). False alarms can also occur because the act of thinking aloud may interfere with answering the question, introducing problems that would not exist when answering silently. Thinking aloud has been observed to degrade performance of several tasks such as mental addition (Russo, Johnson & Stephens, 1989) and formulating preferences (Schooler & Wilson, 1991). False alarms may be a concern whenever the interviewer goes beyond the think aloud data. For example, if the interviewer reviewed the questionnaire prior to conducting the think aloud sessions, he or she might have become convinced that certain problems will occur with particular questions. If the interviewer lists these problems even though there is no explicit verbal report evidence, this could well qualify as a false alarm.

In the case of misses, actual problems are not detected for several reasons. First, respondents may not be able to articulate those aspects of their thinking that involve problems. For example, if they retrieve inaccurate information but believe it is accurate, they are unlikely to provide any verbal evidence of the error. In addition, thinking aloud can actually improve performance on the primary task resolving problems that occur when the task is silently performed; these problems are, thus, not detected (i.e. missed) in the cognitive interview: Russo et al. (1989) observed that when participants were asked to think aloud they were more accurate in a mental multiplication task than when they performed the task silently. Finally, interviewers may simply overlook evidence of problems that is present in respondents' verbal reports; interpreting verbal reports – like answering survey questions – is prone to human error.

In the current paper we explore an alternative approach to collecting verbal reports in cognitive interviews that is based on Ericsson and Simon's ideas about what kinds of information respondents can and cannot report. The general principle underscores the primacy of respondents' verbal reports as the evidence that questions are problematic rather than interviewers' intuitions about the presence of problems. However, interviewers can probe about the content of respondents' verbal reports. This should help clarify inconclusive reports without encouraging respondents to say more than they can. In particular, the number of false alarms should be reduced by requiring interviewers to justify problem identification on the basis of respondents' verbal reports, in effect, preventing them from going beyond the think aloud data; the number of misses should also be reduced by allowing interviewers to probe for more information about what respondents have said or done, in effect licensing the interviewers to explore reports that could reflect a problem that might otherwise be overlooked. We refer to this general approach as the conditional probe method because the interviewer can probe for more information when the respondent's behavior meets certain generic conditions, but the interviewer otherwise plays a background role (similar to the experimenter in the Ericsson & Simon method).

The basic approach is for the interviewer to provide the respondent with ordinary think aloud instructions, for example "report what comes into your head without explaining or justifying your thinking." The interviewer reads each question and, if the respondent lapses into silence, prompts the respondent to keep speaking. When the respondents' verbal reports indicate possible problems but are not definitive, for example, the respondent answers after a long period of silence or changes an answer, the interviewer should probe for additional evidence of a problem. This type of probe is triggered by what the respondent has said or done, not by the interviewer's prior beliefs about possible problems with the question; in this sense the probes are conditional on the respondent's behavior.

The interviewers are presented with a set of generic conditions that could indicate a problem and, for each, there is an example probe. The probes are intended to prompt additional thinking aloud by the respondent, not to test the interviewer's hunch about a particular problem (see Table 2 for the set of conditions and associated probes used in the current study.) Note that

[2] Willis and Schechter, 1997 report a validation study, though it is limited to five questions.

the interviewer must recognize the presence of such a condition and use her own words to probe for additional evidence so, while the conditions of probing are restricted, the interviewer must quickly assess the need for more information and formulate a probe.

C1	Respondent cannot answer or does not know the answer; does not provide a protocol.
P1	"What was going through your mind as you tried to answer the question?"
C2	Respondent answers after a period of silence.
P2	"You took a little while to answer that question. What were you thinking about?"
C3	Respondent answers with uncertainty: explicit statements of uncertainty or implicit markers such as frequent use of "um" and "ah," changing an answer, etc.
P3	"You seem to be somewhat uncertain. If so, can you tell me why?" "What caused you change your answer?"
C4	Answer contingent on certain conditions being met, e.g. "I'd say about 25 times if you don't need a super precise answer."
P4	"You seem a little unsure. If so, can you tell me why?"
C5	Erroneous answer; verbal report implies misconception or inappropriate response process
P5	Clarify respondent's understanding of particular term or the process respondent uses. Suppose the respondent's report suggests she misunderstood the word "manage". Probe this term. "So you don't manage any staff?"
C6	Respondent requests information initially instead of providing an answer
P6	"If I weren't available or able to answer, what would you decide it means?" "Are there different things you think it might means?" If yes: "What sorts of things?"

Table 2. Six generic conditions (C) possibly indicating problems and six example probes (P).

EXPERIMENT

We compared the performance of the conditional probe method to "traditional cognitive interviewing" by asking 4 interviewers trained in each method (8 total) to evaluate a draft questionnaire. The traditional cognitive interviewers were veteran practitioners, having conducted cognitive interviews for between five and ten years. We asked them to conduct their interviews using whatever technique they ordinarily use, so we had relatively little control over their actual conduct. The conditional probe interviewers were new to cognitive interviewing in general and were trained for one day in this particular version of the method.

Each interviewer conducted and tape-recorded five cognitive interviews (total of 20 interviews with each technique). Then, each interviewer produced a problem report for each of his or her five interviews. The two groups differed in how they reported problems: the traditional cognitive interviewers wrote a narrative summary of the problems they identified in each interview and the conditional probe interviewers assigned each problem they identified in each interview to one of 12 problem categories adapted from Conrad & Blair (1996). The categories were derived by crossing three response stages (comprehension, task performance, response mapping) with four problem types (lexical, logical, temporal and computational). If the interviewer did not detect a problem in a particular verbal report, the trial was classified as having "no problem." To make the two types of problem reports comparable, each narrative description of a problem was assigned to one of the problem categories by two coders working together; if no narrative was provided for a question, the coders mapped this to the no problem category. Finally, four additional coders independently listened to all 40 of the interviews and assigned the problems they detected to the same set of problem categories. The point of coding problems (by interviewers or coders) was to make it easier to assess agreement in interpreting verbal reports; the point was not to assess these particular problem categories.

The questionnaire that was evaluated by both sets of interviewers was compiled from preliminary drafts of several questionnaires brought by clients to University of Maryland Survey Research Center. In the authors' judgment, the questions had numerous problems, though these were not independently validated prior to the interviews. The questionnaire consisted of 49 substantive questions, about half of which concerned facts and half of which concerned opinions. Questions were grouped into sections on nutrition, health care, AIDS, general social issues and computer use. Respondents were recruited at random from the local telephone book and paid $25 to participate in a face-to-face cognitive interview at the Survey Research Center.

Results

Traditional cognitive interviewers identified 1.5 times more potential problems than conditional probe interviewers, .36 versus .24 problems per question for traditional interviews and conditional probe interviews, respectively (F[1,22]=6.12, p=.022). While this is a substantial difference it does not necessarily mean that the conditional probe interviews failed to detect large numbers of problems that the traditional cognitive interviews picked up. Instead, the difference could reflect a large number of false alarms by the traditional interviews or a moderate number of false alarms by the traditional interviews and a moderate number of misses by the conditional probe method. We really cannot resolve this without somehow validating that the potential problems identified in the cognitive interviews are actually experienced by respondents under field administration conditions. While the current study does not provide direct validation information, it does enable us to assess the degree to which there is agreement about the presence of a problem.

Agreement measures are less definitive than validation measures but they provide valuable information about the quality of problem detection. In particular, it is possible for two judges to agree that there is a problem (or that there is not a problem) and still be wrong; however, if they *disagree* about the presence of a problem, we can be sure that one of the judges is wrong which clearly reflects lower quality than if they had agreed. Table 3 presents average kappa scores (which adjust for chance agreement) between all interviewer-coder pairs for the two types of interviews. The first row of the table lists average kappa scores for the overall decision of whether or not there is a problem with a particular question. Agreement indicates that either the interviewer and coder both judged there to be no problem or that both believed there to be a problem, though not necessarily the same problem. The second row lists average kappa scores for cases in which the interviewer and coder both identified a problem. Agreement here means they detected the same problem and assigned the case to the same problem category (out of 12); disagreement means they detected different problems (or at least assigned the evidence of a problem to different categories).

	Traditional Cognitive Interviews	Conditional Probe Interviews
Is there a problem?	.20	.35
If so, what type?	.39	.47

Table 3. Average kappa scores between all interviewer-coder pairs.

There is a striking, though unexpected, result in the table: all of the kappa scores are quite low, indicating that interviewers and coders rarely agreed about the presence of problems. This suggests that verbal reports may lack the resolution to definitively expose respondents' problems answering survey questions. This is disturbing because cognitive interviews serve as the basis of modifications to important surveys yet, based on these results, it seems that any two people would be not be very likely to interpret the same verbal report in the same way. Unfortunately, the low agreement rates cannot be attributed to a difficulty using the coding system. If that were the case, then agreement would be better for the overall problem-no problem decision (row 1) than the specific problem category assignment (row 2) because in the first case the judgment does not involve the coding system. However, the kappa scores for the overall problem decision (row 1) are lower, if anything, than the kappa score for the problem category decision (row 2).

Despite the low agreement rates, coders agreed more often with conditional probe interviewers than they did with traditional cognitive interviewers. The advantage for conditional probe interviews is reliable for both the overall problem-no problem decision (t[14]=3.01, p = .009) and the specific problem category decision (t[12]=3.35, p = .006). We interpret this as evidence that restricted interviewer interventions in which they request only that respondents elaborate what they have said or done, produces more definitive evidence of problems than when interviewers probe beyond the think aloud data.

Another view, however, is that the traditional cognitive interviewers (who were quite experienced at pretesting questionnaires) were simply more sensitive to evidence of problems than either the conditional probe interviewers or the coders (who were all relatively inexperienced). This could explain the lower agreement rates for the traditional version of the method: traditional cognitive interviewers recognized evidence of problems in respondents' verbal reports that coders could not detect; because conditional probe interviewers lack such extra sensitivity, their interpretations of the verbal reports were more similar to those of the coders. If this is so, then agreement rates among pairs of coders should be higher than among interviewer-coder pairs for the traditional cognitive interviews because all coders should lack such extra sensitivity.

	Traditional Cognitive Interviews	Conditional Probe Interviews
Is there a problem?	.19	.30
If so, what type?	.36	.43

Table 4. Average kappa scores between all coder pairs.

Intercoder agreement scores are presented in Table 4. The kappa scores are very similar to those in Table 3. The advantage for the conditional probe interviews is again evident for the problem-no-problem decision, t[5]=2.94, p<.05, though not reliably for the specific problem type decisions, t[5]=1.11, n.s. The important point is that intercoder agreement for the traditional cognitive interviews (Table 4, column 1) is not higher than is agreement between interviewer-coder pairs (Table 3, column 1); in fact, the intercoder scores are lower, if anything, than the interviewer-coder scores, though statistically, there is no difference: t[18]=0.352, n.s., for the overall problem decision and t[17]=0.922, n.s., for the specific problem type decision. So, apparently, the difference between the techniques in agreement rates has more to do with the quality of the evidence than with the interpretation skills of the practitioners.

A closer look at the types of probes used by both groups of interviewers is consistent with the idea that the verbal reports differed in effectiveness between the groups. In particular, the traditional cognitive interviewers probed 4.2 times more often than did the conditional probe interviewers (987 versus 236 times across the 20 interviews for each group) yet detected only 1.5 times as many problems. More specifically, the conditional probe interviewers solicited information about respondents' utterances (i.e. administered conditional probes) in 61% of their probes (144 times); the traditional cognitive interviewers probed about potential evidence of problems in respondents' utterances in only 13% of their probes (128 times).

The balance of the probes administered by the traditional cognitive interviewers was largely comprised of requests for paraphrases and other inquiries about the meanings of terms in the question (41% of probes), but these were far less prevalent in the conditional probe interviews (14% of probes). These probes were preceded by evidence of some kind of problem – though not necessarily problems about question meaning – only about half of the time they were administered (47% in the traditional cognitive interviews and 58% in the conditional probe interviews) and they were only moderately effective at uncovering problems: at least one coder judged the questions in which these probes were administered to be problematic 55% of the time in traditional cognitive interviews and 53% of the time in conditional probe interviews. In contrast, when interviewers administered conditional probes, at least one coder judged those questions to have a problem 78% of the time for traditional cognitive interviews and 89% of the time for conditional probe interviews. This suggests that conditional probes are substantially more effective than probes about question meaning (which are often administered even though there is no evidence of problems with meaning).

However, even the most effective probes become less helpful in finding problems if we require that multiple coders concur about the presence of a problem – a reasonable requirement in light of the overall low agreement rates for both methods. If we impose a stricter criterion for what counts as a problem then interviewers identify many potential problems about which there is no agreement, regardless of the method. Figure 1 presents the number of problems detected by interviewers and the number of problems on which 1 or more coders agree. The clearest result in the figure is that two coders agree about the presence of very few potential problems and that as the criterion is increased to agreement by three and four coders the number of reliably detected problems drops off to almost zero. This reinforces the idea that interpreting verbal reports – at least these reports – is inherently subjective and variable.

CONCLUSIONS

We have focused on just two versions of cognitive interviewing, using one questionnaire and a small

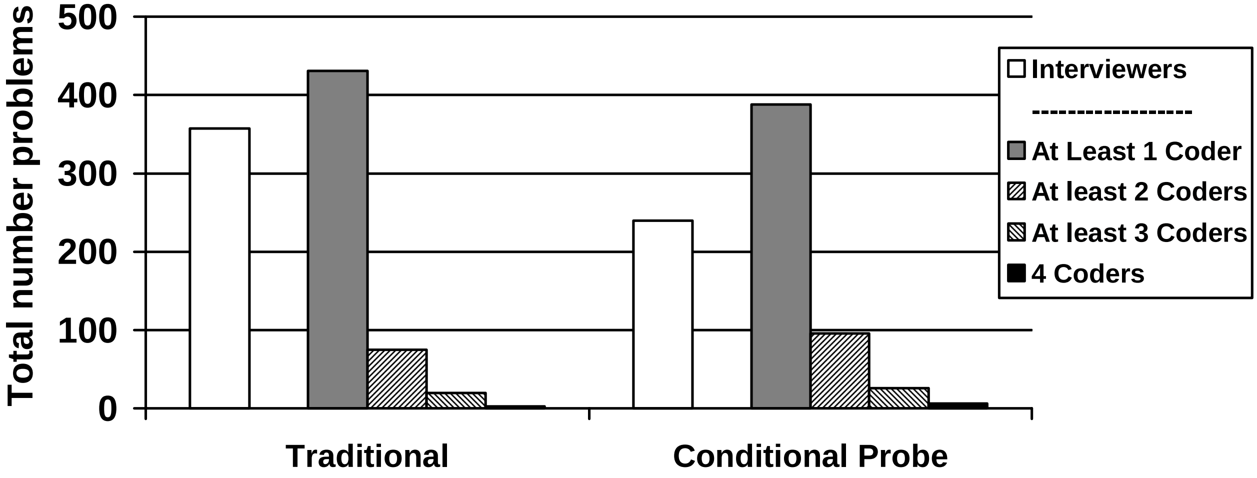

Figure 1. Number of problems found by interviewers and coders in traditional cognitive interviews and conditional probe interviews

number of interviewers each conducting a small number of interviews. It is entirely possible that any differences in these factors might produce quite different results. Nonetheless it is hard to escape the implication of the current results that data collected with cognitive interviews – verbal reports about the process of answering survey questions – are far less definitive than we have assumed them to be.

Low agreement could reflect genuine problems that are missed by an interviewer (or coder) or it could reflect spurious problems that are falsely reported by an interviewer (or coder). In either case, the result is not encouraging. If the technique is missing genuine problems, then data quality is not improved to the degree that is usually assumed by using the technique. If the technique is leading to false alarms, it may promote unnecessary changes to questionnaires, which can introduce more problems and compromise time series by altering questions between time periods. At the very least, false alarms squander resources by "fixing" questions that do not need to be fixed. Given the prevalence of probes about question meaning and the fact they are often administered when there is no evidence of problems with meaning, it seems that false alarms may be more responsible than misses for the low agreement. But it is difficult to confidently interpret the low agreement rates without some kind of evidence about respondents' experience answering the same questions in actual interviews.

On a more positive note, the current findings suggest that some components of cognitive interviews are relatively successful. In particular, conditional probes seem to lead to verbal reports that indicate a problem more often than do probes about question meaning. Presumably there are other relatively effective types of probes that we can identify with continued exploration. There is every reason to believe that as we learn more about cognitive interviews we can refine the technique and use it more effectively.

REFERENCES

Conrad, F. & Blair, J. (1996) From impressions to data: Increasing the objectivity of cognitive interviews. In *Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association*. Alexandria, VA: American Statistical Association., pp. 1-10.

DeMaio, T. J. & Rothgeb, J.M. (1996). Cognitive interviewing techniques in the lab and in the field. In Schwarz, N. & Sudman, S. (Eds.) *Answering Questions: Methodologies for Determining Cognitive and Communicative Processes in Survey Research.* San Francisco: Jossey-Bass, pp. 177-195.

Ericsson, K. A. (1975). Instruction to verbalize as a means to study problem solving with the eight puzzle: A preliminary study (No. 458). *Reports from the Department of Psychology*. Stockholm: University of Stockholm.

Ericsson, K.A. & Simon, H.A. (1993*). Protocol Analysis: Verbal Reports as Data, rev. ed.* Cambridge, MA: MIT Press.

Flaherty, E.G. (1974). The thinking aloud technique and problem solving ability. *Journal of Educational Research, 68,* 223-225.

Newell, A. & Simon, H.A. (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Payne, J. W. (1994). Thinking aloud: Insights into information processing. *Psychological Science, 5*, 241- 248.

Presser, S. & Blair, J. (1994). Survey pretesting: Do different methods produce different results? In Marsden, P.V. (Ed.) *Sociological Methodology, Vol 24.* Washington, DC: American Sociological Association, pp. 73-104.

Rothgeb, J., Willis, G. & Forsyth, B. (2001). Questionnaire pretesting methods: Do different techniques and different organizations produce similar results?. Paper presented at 56th Annual Conference of the American Association of Public Opinion Research, Montreal, CA.

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

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

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

Wilson, T. & 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. & 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: