Verbal Reports Are Data!
A Theoretical Approach to Cognitive Interviews

Frederick Conrad, *Bureau of Labor Statistics*
Johnny Blair, *University of Maryland*
Elena Tracy, *University of Maryland*

A survey question is kind of instruction. Respondents are, in effect, instructed to carry out a task – to do whatever mental work is necessary to provide an answer. By this view, they can get into trouble in several ways. For example, respondents can misunderstand what they have been asked to do and, as a result, carry out the wrong task. Or they can understand the instruction just fine but find themselves unable to carry out the task they have been assigned. Or they can successfully carry out the assigned task but find themselves unable to fit their answer into the options provided. Each of these difficulties demands a different sort of solution, for example, making question wording clearer, simplifying the task, or better matching the response options to the way people think about the topic of the question.

Such problems could well go undetected once a survey is in the field. This would surely compromise the quality of the data that the survey produces. Laboratory pretesting, in particular the method known as cognitive interviewing, has been developed to identify such problems. While the method has come to include a variety of tools, almost all versions include the use of think aloud techniques. Respondents are asked to give a verbal account of their thinking – a "verbal protocol" – as they answer (concurrent) or immediately after they answer (retrospective) a draft survey question. The advantage of verbal reports is that they can provide a rich and continuous account of the underlying thinking, as opposed to the discrete data points produced by measures like response time. In cognitive interviews, respondents' verbal reports are examined for evidence of the kinds of problems listed above. Based on evidence of problems, the question author can rework the problematic questions so that they are less likely to cause these difficulties when asked in actual interviews. Judging by the ubiquity of cognitive interviews, they are considered to be effective in routing out problems prior to live data collection.

In this paper, we will first discuss the theory of verbal protocols and its relation to cognitive interviews as they are used in practice. Our main point is that, on theoretical grounds, the method can be effective in detecting certain kinds of problems, but not all problems. In the section that follows, we will report on our attempts to strengthen the method by standardizing both the collection and analysis of think aloud data. In particular we have developed a type of probing that we believe is consistent with the ideas behind the think aloud method yet should be able to extend the range of survey response problems that can accurately be detected in cognitive interviews. In addition we have developed a technique for classifying the problems uncovered in cognitive interviews that is designed to produce reliable interpretations of the protocols.

The theory of verbal protocols and its relation to cognitive interviews

The theory of verbal protocols is the backbone of the cognitive interviewing technique, at least as it was originally conceived (e.g. Lessler, Tourangeau and Salter, 1989). However, the method has come to include more than just think aloud procedures. For example, it seems to be common practice for cognitive interviewers to probe for additional evidence beyond what respondents report spontaneously (e.g. Willis, DeMaio and Harris-Kojetin, 1999). This represents a significant departure from pure think aloud methods in which the experimenter's job is to present the problem and remind the respondent to keep talking when they fall silent. The experimenter does not provide additional evaluation and tries to be unobtrusive. For the moment we will focus on pure think aloud methods and the degree to which this fits the task of identifying respondents' problems with draft questions. In the next section we will consider the pros and cons of probing.

We believe that think aloud methods are a valuable tool for pretesting and debugging questions

before they are used in the field. We use them regularly. Our purpose in the current section is to explore the conditions under which they provide the most accurate information about respondent problems with questions and the conditions under which they are less accurate in detecting problems The general point about think aloud methods in cognitive interviews is that they need to be used with consideration. While they may be appropriate in the majority of situations they are used, there are certain circumstances in which think aloud methods may fail to uncover problems and others in which they may indicate the presence of a problem when there actually is none.

People can only report what they are aware of. In order for people to articulate their thinking, they must be aware of it. People are typically aware of and can report about the components of high level mental processes, like the sequence of steps that leads to the solution of a problem. According to the theory of verbal protocols (Ericsson and Simon, 1980, 1993), people temporarily store these steps in their working memory and they are generally aware of what they are holding in working memory.

You could probably report your thinking as you solve a navigational problem like the following one. Suppose you find yourself in Paris and you do not know your way around. You are trying to get from your current (unknown) location to the Louvre. You might see the Eiffel Tower nearby and locate it on your map. Then, having gotten oriented, you might locate the Louvre on the map. The next step might be to plan a route between the Eiffel Tower and the Louvre. And then to follow that route. While you could probably report on this type of thinking, you probably could not report on more immediate processes like how you recognized the Eiffel Tower or how you know that Paris is the capitol of France. The point is that verbal protocol methods are designed to tap into certain types of thinking but not all.

In many cases, answering survey questions involves reportable thinking. For example when respondents answer questions about the frequency of events by retrieving and counting up all the events they can remember, they can report on the individual retrieval operations for each remembered event (e.g. Bickart & Felcher, 1996; Blair, Menon & Bickart, 1991; Conrad, Brown & Cashman, 1998; Menon, 1993). More importantly for pretesting, when there are problems with the question, such as a recall task that spans too broad a time period, respondents become aware of the difficulty and can report it much as one could report on the solution to the navigational task above. This is where cognitive interviews are most valuable.

In other cases, respondents are not aware of the information that figures into their answers. In these situations, verbal reports are less valuable and may even be misleading. If respondents have no awareness of how they are answering a question, then they cannot report anything at all; if the only information that comes to mind is not actually germane to the response process, their reports will be of questionable validity.

Suppose a respondent in a cognitive interview is asked “Have you ever had an allergic reaction to any medication?” Suppose, further, that she has specifically stored the fact that she has never had any such reactions. She can automatically (*e.g.* Shiffrin and Schneider, 1977) recall this fact rather than deliberately searching her autobiographical memory for allergy episodes. Yet, precisely because she does not deliberately control the retrieval process, she cannot report much about it (Ericsson & Simon, 1993). The rub for cognitive interviewing is 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.

For other question answering tasks, reportable information comes to mind but is not actually the basis of the response. Wilson, LaFleur and Anderson (1996) discuss this problem in the context of “Why” questions (“Is there anything about Mr. Reagan that might make you want to vote for him?”). In experimental studies that did not use think aloud methods, voters’ preferences for political candidates were shown to depend largely on factors like (1) the amount of exposure to the candidate and (2) the party affiliation of one’s parents (p. 96). Wilson et al. argue that most respondents are not aware of these influences on their preferences and would be unlikely to report them if asked to think aloud. Yet they are aware of and capable of reporting other, less

valid, reasons for their opinion. Wilson, et al. are concerned that a researcher who relies on verbal reports in such situations will come to a distorted understanding of the underlying processes.

Does thinking aloud affect ones thinking? It is conceivable that the act of thinking aloud alters the thinking being reported, presumably degrading performance of the primary task. This would be dangerous in cognitive interviews because the protocols could indicate there are problems where there are actually none. Ericsson and Simon (1993) argue that this is not the case, that verbal protocols are not "reactive." They review a long list of studies showing that people perform the assigned, primary task equally well whether or not they provide verbal reports. Although thinking aloud may slow down the primary task it should not change it fundamentally – assuming the method is used under certain conditions: think aloud methods will not react with the primary task if people are asked to report only on the content of working memory, if they are asked to report and not explain or evaluate their thinking, and if the primary task is primarily verbal. If the task involves recoding information from another format – say visual – into words, then the verbal demands of thinking aloud could react with the task. For pretesting surveys, this would argue against collecting verbal protocols when respondents are asked to interpret any kind of map or diagram.

Despite reassurances by Ericsson and Simon, there is some discouraging evidence to the contrary. Russo, Johnson and Stephens (1989) showed that thinking aloud reduced the accuracy of mental arithmetic (a working memory-intensive task) and actually increased the accuracy of a gambling task, relative to performance on the same kind of problems without thinking aloud. Accuracy was not affected for all of the tasks that Russo et al. tested. Two other tasks (including one requiring visual recoding) were performed with equal precision whether or not participants reported on their thinking.

The reactivity of mental arithmetic is a concern for cognitive interviews because it is required by many types of survey questions, for example, those involving expenditures, income, and frequency. Pretesting such questions with think aloud methods could potentially distort the findings. On the other hand, people use techniques like counting on their fingers to reduce the burden on working memory; such an approach by respondents would likely reduce the reactivity of the task. The more general point is that tasks placing severe demands on working memory could react with think aloud techniques; it is important to consider this when choosing a pretesting method or evaluating the reports collected in a cognitive interview.

The reactivity observed by Russo et al. (1989) was not limited to concurrent think alouds. Retrospective think alouds also degraded accuracy for mental arithmetic and increased accuracy for the gambling task. The authors attribute this to a variety of factors. The one that is most deleterious to performance is a competition between the primary and think aloud tasks for limited mental resources. The factors that most help performance are taking extra time to reflect on the solution and adopting less error prone strategies in the presence of an experimenter.

Other researchers (Schooler, Ohlssson, & Brooks, 1993) have shown that thinking aloud can degrade performance for so called insight problem solving. Schooler, et al. (1993) asked people to solve both conventional problems, whose solution involves a reportable sequence of steps, and insight problems, whose solution comes to mind, often after a period of silence, even though the process is hard to verbalize[1]. Half the participants were asked to think aloud and the other half were not. The researchers found that people's ability to solve insight problems was compromised by thinking aloud although their ability to solve conventional problems was not.

---

[1] An example of an insight problem used by Schooler et al. (1993) is "A prisoner is attempting to escape from a tower. He found in his cell a rope that was half long enough to permit him to reach the ground safely. He divided the rope in half, tied the two parts together and escaped. How could he have done that? *Solution*: He unraveled the rope and tied the two pieces together.

Schooler et al. (1993) suggest that thinking aloud focuses people on verbally relevant information and "overshadows" the non-verbal, non-reportable processes required to achieve the insight. In particular, they speculate that verbalization interferes with the search for information that might contribute to a solution. Such searches are believed to involve the passive spread of activation throughout memory, much like the automatic fact retrieval discussed in the earlier example about medical allergies, which is not under people's conscious control.

Although surveys probably never ask people to solve insight problems, certain questions may require thought that is unencumbered by verbal demands. For example, formulating an opinion may be based, at least in part, on information that is hard to verbalize, particularly if it involves emotions or preferences. Asking respondents to think aloud in a cognitive interview while formulating such opinions may interfere with their ability to develop a coherent opinion and indicate, incorrectly, that the question poses an unreasonable task.

One piece of evidence that verbalization can affect the formulation of preferences comes from a study by Wilson and Schooler (1991). They asked participants to evaluate different brands of strawberry jam. Half of the people wrote down the reasons for their preferences before rating the jams and the other half simply rated the jams. The group who only rated the jams agreed closely with experts while the group who first wrote down their reasons showed little agreement with the experts. Although verbalizing their reasons did not seem to make the task harder for these participants – it just degraded the quality of their responses – it certainly could hinder the process when respondents provide concurrent think alouds.

The lesson for our purposes is that thinking aloud can, under some circumstances, interact with the process being reported. If this is the case when experimenters do their best to remain detached, and just remind participants to keep talking, it suggests that the more active type of probing common in cognitive interviews stands a good chance of influencing what people report. In the next section we turn to an experimental technique designed to balance the need to collect information that respondents might not spontaneously report with the need to minimize the researchers' influence.

Increasing the objectivity and consistency of cognitive interviews

Over the past several years we have developed a version of cognitive interviewing designed to address a number of our concerns about the method as it is currently practiced. Among our concerns are the method's lack of theoretical grounding, wide variation in its administration, and impressionistic interpretation of the protocols.

We have developed separate methods for collecting protocols and for analyzing them. Our collection method – the interview process – adheres to pure think aloud procedures as much as possible but includes a set of generic probes. The point of these probes is to allow the interviewer to explore beyond what the respondent says or does not say without increasing the chances of invalid or reactive reports.

Our analysis procedure involves assigning segments of the protocols to a standard set of problem categories. The point is to encourage the analyst to exhaustively and repeatedly consider a broad set of criteria about possible problems.

The reason for temporally separating protocol collection from protocol analysis is to allow the interviewer to elicit protocols without interpreting them and to enable the analyst (coder) to devote full attention to the content of the reports, free from the demands of conducting interviews.

Collecting cognitive interview data. The practice of cognitive interviewing has come to involve a combination of think aloud procedures and other forms of self-report. The most common addition to pure think alouds is "verbal probes" (e.g. Willis et al, 1999). The use of probes seems to vary widely. Some probes are scripted prior the pretest session, some are chosen from a stock set when the interviewer deems them appropriate, and some are invented by

the interviewer as needed. Many probes are concerned with respondent understanding, e.g. “Can you repeat the question in your own words?” while others explore the respondent’s judgments about the response process, such as her confidence in the answer she just provided (Willis, et al., 1999). Some probes are general, designed to broadly elicit additional verbalizations (“Is there anything else?”) and some are sharply focused, designed to test particular hypotheses about problems, e.g. “Did you find it difficult to come up with an answer while keeping all of the response options in mind?”. And so on.

While the use of probes in cognitive interviewing is widespread it is also controversial. Probes can focus respondents on aspects of the response process they might not have otherwise considered. As Gerber and Wellens (1996, p. 20) point out, this is seen as a source of bias by some (e.g. Forsyth and Lessler, 1991) and an efficient way to reduce irrelevant verbalization by others (Willis, 1994). Our view is that certain types of probes can increase the chances of invalid and reactive protocols but others can accurately elicit information that respondents are aware of but, for some reason, do not report

Probes may lead to invalid or reactive reports for the same reasons that pure think aloud techniques may produce similarly flawed data. Probes may ask respondents for information to which they do not have access (e.g. asking “What were you thinking?” when the response was fast and automatic) and they may affect the response process by essentially adding a third task: in addition to answering the survey question and to thinking aloud, the respondent must now also answer (or prepare to answer) a probe question.

Having expressed these reservations about probes, we also believe that they can be useful under certain circumstances. In our experience, pure think aloud techniques sometimes produce a hint of a problem but not enough information to be adequately diagnosed. Thus we advocate a kind of probing in which interviewers ask respondents to amplify or clarify certain types of verbalizations in their protocols. For example, if the respondent’s protocol for a particular question includes behaviors that may signal uncertainty, e,g, *uhms* and *ahs* or particularly long silences, it may be fruitful to ask the respondent if something is confusing or difficult. It seems likely that a respondent could retrospectively articulate the source of a behavior he has just exhibited; the information to which he was reacting – of which he may well be aware – should still be present in working memory. The respondent might not spontaneously articulate a problem because solving it is too demanding to also verbalize it (e.g. Russo et al. 1989). But probing working memory after the fact should reduce reactivity and minimize the demand of the additional task.

In listening to a number of taped cognitive interviews we observed several classes of respondent verbal behavior that occurred repeatedly and seemed to signal potential problems. These seemed to be exactly the kind of situation for which probing might uncover additional information without threatening validity. For each class we developed a few example probes. These are listed in Table 1. We refer to these as *conditional probes* because they are contingent on the occurrence of certain classes of verbal behavior. Because the number of probing conditions is relatively small we believe interviewers can remember when to probe.

We recently completed 20 cognitive interviews using a combination of pure think aloud and conditional probing techniques. Our expectation is that this will lead to more complete and reliable identification of problems than will less theoretically grounded techniques. In order to asses the completeness and reliability of problem reports – to analyze them – we classify them into a taxonomy of problems. We now turn to this analysis approach.

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

Table 1. Conditional probes for cognitive interviews.

Analyzing cognitive interview data. Typically, the researcher who conducts a set of cognitive interviews writes a report summarizing the problems that seem evident in respondents' verbal reports. The reports tend to concentrate on the substance of the problem, e.g. "Respondents had difficulty reporting both frequency of consumption and typical serving size for the list of foods they were read." There is no doubt that this sort of description is extremely valuable for evaluating a draft questionnaire and is unlikely to be yielded by more conventional pretesting methods. We think that the analysis of cognitive interviews can be made even better.

Cognitive interviews primarily uncover problems with understanding – with question meaning (e.g. Gerber and Wellens, 1996). But certainly respondents find other aspects of answering questions to be problematic. We have developed a taxonomy of questionnaire problems that includes problems with question meaning – we call them *lexical problems* – but also includes *logical*, *temporal* and *computational* problems (see Conrad and Blair, 1996, for more details[2]). By encouraging coders to consider all of these problem classes when inspecting the protocols for potential problems, they should uncover more potential problems than if they concentrate on problems with meaning alone.

When a researcher identifies a problem in a cognitive interview, this does not guarantee she will identify a solution. In some cases the solutions are obvious – for example if some respondents don't know what a word means, it should be defined in the question or replaced by a more common synonym. But in other cases, it is less clear how to change the question to eliminate the problem. One reason for this is that problems may originate at one point in the response process but become manifest at another. If the origin of the problem, rather than the symptoms, can be identified, then the general type of solution becomes clearer. Consider a respondent who misunderstands what he is being asked to do and, as a result, ends up carrying out a task that is

[2] The version of the taxonomy presented in Conrad and Blair (1996) includes a fifth problem type, *inclusion/exclusion* problems. In the interest of simplicity the number of problem types was reduced to four and some definitions expanded to include *inclusion/exclusion* problems.

unreasonably difficult. The literal content of his protocol may indicate difficulty performing the task but the problem originated in the earlier stage of understanding the question. The solutions should involve the way the task is described in the question rather than simplifying the task. To enable this kind of analysis we include three general stages of responding[3] in the problem taxonomy. Each problem type, therefore, can occur at each stage in the response process.

The typical problem reports written by one researcher may be hard to compare to reports written by another. This is due, in part, to the fact that different researchers analyze the protocols at different levels and, also, it is due to the fact that the descriptions are textual and qualitative. By assigning segments of protocols to problem categories (problem types at particular stages in the response process) it becomes possible to compute the agreement between multiple coders. Because we separate the collection and analysis of cognitive interview data, it is possible for multiple coders to listen to the same taped interviews for problems; they can then assign problems to categories in the same taxonomy and it is possible to compute intercoder reliability.

It is usually the case that the researchers who conduct and interpret cognitive interviews are graduate level social scientists. The problem taxonomy, however, can be successfully used by more junior level staff. In a study that we recently conducted with junior research staff serving as coders, intercoder reliability, measured by Cohen's kappa, was in the .7 range. This is not perfect – conceivably because the cognitive interviews that were coded had been conducted without the benefit of the conditional probes described above and there were many protocol segments that were uncodable – but it suggests that the approach is usable by staff who are easier to find and less expensive than typical practitioners of cognitive interviews.

Conclusion

The use of verbal reports to pretest questionnaires represents the most tangible outcome of the dialogue between survey methods research and cognitive psychology. Yet somehow the rigor and theoretical basis of the think aloud methods and related types of self report, seem to have gotten lost along the way. We have tried to point out some of the dangers of forsaking what is known about thinking aloud. We have also presented an approach to collecting and analyzing cognitive interview data designed to reliably identify problems in answering survey questions. We are not necessarily advocating that the research community adopt this particular version of the technique. It first needs to be thoroughly evaluated and other versions may well be more precise in their diagnoses or easier to use. Instead we are encouraging practitioners to follow this general approach in order to help move questionnaire design from an art to something like a science.

References

Bickart, B. & Felcher, M. (1996). Expanding and enhancing the use of verbal protocols in survey research. In Schwarz, N. and Sudman, S. (Eds.) *Answering Questions: Methodology for Determining the Cognitive and Communicative Processes in Survey Research*. San Francisco: Jossey-Bass, 115-142.

Blair, J., Menon, G., & Bickart, B. (1991). Measurement effects in self vs. proxy responses: An information-processing perspective. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (eds), *Measurement errors in surveys*. New York: Wiley, pp. 145-166.

[3] These stages are adapted from the widely accepted framework used by cognitively oriented survey researchers to think about the response process (e.g. Sudman, Bradburn and Schwarz, 1996). Usually retrieval and judgment formation stages follow comprehension and precede response formatting. Because respondents may well be unaware of retrieval and judgment processes we do not expect to be able to distinguish between them when assigning segments of protocols to a category in the taxonomy. As a result, we have combined them into a single *task performance* stage

Conrad, F. and 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.

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

Ericsson, A. & Simon, H. (1980). Verbal reports as data. *Psychological Review, 8,* 215-251.

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

Forsyth, B. & Lessler, J. (1991). Cognitive laboratory methods: A taxonomy. In Biemer, P., Groves, R. Lyberg, L., Mathiowetz, N. and Sudman, S. *Measurement Errors in Surveys.* New York: Wiley, pp. 393-418.

Gerber, E. & Wellens, T. (1997). Perspectives on pretesting: "Cognition" in the cognitive interview? *Bulletin de Methodologique Sociologique,* 55, 18-39.

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

Menon, G. (1993). The effects of accessibility of information in memory on judgments of behavioral frequencies. *Journal of Consumer Research,* 20, 431-440.

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.*

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

Sudman, S., Bradburn, N. & Schwarz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Willis, G. (1994). Cognitive Interviewing and Questionnaire Design: A Training Manual. National Center for Health Statistics, Cognitive Methods Staff, Working PaperNo. 7.

Willis, G., DeMaio, T. & 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.

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: Jossey-Bass, pp. 91-114.

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.