Inside the Black Box:
Analysis of Interviewer-Respondent Interactions in Cognitive Interviews[1]

Nadra Garas
Development Associates, Inc; 1730 North Lynn Street; Arlington, VA 22209
Johnny Blair
Abt Associates, Inc.; 1110 Vermont Ave.; Suite 610; Washington, DC 20005; Johnny_Blair@abtassoc.com
Frederick Conrad
University of Michigan; 426 Thompson Street; Ann Arbor, MI 48106; fconrad@isr.umich.edu

Conversational Interaction in Cognitive Interviews

Survey methodologists borrowed the think aloud method from cognitive psychology in order to make respondents' thinking explicit in a variety of tasks. The method has become a central feature of cognitive interviewing to pretest questionnaires. In adapting the method for pretesting, survey researchers changed it in important ways that may have implications for data quality, either for better or worse. The procedures described by Ericsson and Simon (e.g. 1993) for eliciting verbal reports limit the interviewer/researcher's role to providing initial instructions to the subject and, during the laboratory session, reminding the subject to keep thinking aloud. Under this approach, the content of the verbal reports should be largely unaffected by the interviewer/researcher's actions. In cognitive interviewing, the interviewer's role has generally been more active and more interactive. The interviewers query respondents' understanding of survey questions, request they perform additional tasks beyond thinking aloud, e.g. paraphrasing the question, and interact with respondents, for example, conversing about what has already been said (e.g. U.S. Bureau of the Census, 1998; Willis, DeMaio & Harris-Kojetin, 1999).

The current paper explores the impact of different kinds of probes on the information that is subsequently produced by respondents in cognitive interviews. Our concern is that some probes, particularly those that are not motivated by what the respondent has already said, may lead to the spurious detection of problems. We worry that as a result designers may often modify questions that really do not need to be changed. This can be costly. It consumes designers' time, may introduce new problems and, if the survey is on-going, unnecessarily complicates comparisons between data collected with the old and new versions. It is possible that question designers weight the reported problems on the basis of their plausibility and only revise those to which they give the most weight. But this seems unlikely given that the written summaries prepared by interviewers rarely describe the conversational context – respondents' utterances, interviewer's probes, respondents' reactions to probes, etc. – in which the listed problems were observed. Without such context, designers cannot really assess whether or not a probe has genuinely uncovered a problem. The default process may be to give equal weight to all listed problems leading to the kinds of costs just mentioned.

With one class of probes, interviewers ask respondents to elaborate on possible indications of problems they have spontaneously provided. Interviewers may administer such probes to follow up on a comment or question by the respondent or an inadequate answer. We describe probing under such circumstances as "evidence-based." Evidence-based probes are motivated by some indication of a problem or difficulty. The probes themselves may be generic (e.g. "Can you tell me more about that?") or question-specific (e.g. "Why do you think your answer could be either 'somewhat urgent' or 'extremely urgent'"?).

Interviewers administer a different class of probes in the absence of any suggestion that the respondent has experienced problems. For example, the data we present below indicate that interviewers often ask respondents what a particular word means or to paraphrase the survey question and they usually do this without any indication that respondents are misinterpreting the question or are uncertain about its meaning. Such "context-free" probes can be motivated by the interviewer's judgment about how a question may be flawed, or asked because a previous respondent had a problem with the question, etc.

[1] We thank Greg Claxton and Sarah Woldehanna for coding and analyzing interactions. We thank Michael Schober for comments on earlier draft of this paper. We thank the Bureau of Labor Statistics for supporting much of this work. The content of the paper reflects the opinion of the authors and not that of the Bureau of Labor Statistics.

A major part of developing cognitive interview procedures is deciding what sorts of probes to use, as well as when to use them, whether to write them beforehand or improvise them during the interview, whether to ask for respondents' interpretations of questions, and so forth. It seems reasonable that these interviewer behaviors can make the reports richer and more informative. But it is also possible they will have no effect, or worse, that they will have a detrimental effect on the quality of the verbal report. We have little basis, beyond the anecdotal, to say how often particular interviewer actions produce useful information about the performance of survey questions, generate empty verbiage, or result in misleading verbal reports. Our conjecture is that probes are highly relevant to the quality of the information they elicit because, in part, they can affect respondent's thinking, possibly suggesting problems that respondents had not actually experienced (see Russo, Johnson & Stephens, 1989 for a discussion of "reactive" effects in verbal reports).

In the current paper we explore how some probes can affect respondents' thinking (apparent in their verbal reports) by examining the patterns of interaction in which the probes occur. In reaction to the instructions and interviewer behaviors in cognitive interviews, respondents can do more than just think aloud and respond to probes. Respondents can ask questions of the interviewer, express frustrations, make comments about the questions and engage in other speech acts beyond simply reporting their thinking while answering a survey question. Taken together, the sequences of interviewer and respondent behaviors across conversational turns produce patterns of interaction that serve as the data in the current study.

Because the potential number of interaction patterns is quite large (in theory, unlimited), examining all such interactions in a set of cognitive interviews is unlikely to be fruitful. However, there are good reasons to focus on certain interaction types above others. We focus on those interactions that contain either evidence-based or context-free probes. We believe that evidence-based probes will lead to respondent descriptions of problems that are more compelling and unequivocal than are those elicited by context-free probes. In the first case, respondents elaborate on evidence they have spontaneously produced whereas, in the second, they describe a problem only in response to a probe. We pay special attention to interactions that include interviewer instructions to paraphrase. This is standard practice (e.g. U.S. Bureau of the Census, 1998; Willis, et al., 1999) but we don't know how effectively it finds problems.

Classifying Interaction

Some interviewer-respondent interaction patterns can be represented generically, i.e. independent of any particular question administration. Such patterns can be treated as units of analysis. We can then count the frequency of different types of interactions, and otherwise analyze them statistically.

The following generic interaction describes an evidence-based probe and an elaboration on the possible problem: {I asks the question; instead of answering, R mentions some problem with the question; I asks R to say more about the perceived problem; R describes the problem alluded to in more detail.}. A specific example of this pattern might be: {I asks "Do you think the police and courts are doing a good job?" Instead of answering, R says, "I don't know how to answer that"; I says, "Can you tell me more?" R responds by saying "I think the cops do a good job but not the courts. So I can't give a single answer"}. In this case, the respondent initially suggested that there might be a problem, but not enough to be sure there was a problem or, if there was, to learn what might be its nature. This kind of hint of a possible problem is what we refer to as "evidence." By itself, the comment "I don't know how to answer that" could have meant that the respondent feels she or he isn't knowledgeable enough about the police or courts to answer, which would not, in our opinion, provide sufficient grounds on which to determine that the question has a problem. However, when the interviewer probed, the respondent clearly described the problem from the double-barreled question. The response to the evidence-based probe, thus, clearly revealed a problem. Had the respondent's initial statement more explicitly described a problem, e.g. "You're asking me about two different things at the same time," the probe would have added little.

A context-free probe could occur when: {I asks question; R answers the question satisfactorily, with no added comments; I probes about some aspect of the question.} The probe might, for example, be about the meaning of the question as a whole, the meaning of a particular word, or something else other than meaning. The point is that the probe is not motivated by any indication of a problem.

Study Design and Analysis

Four experienced cognitive interviewers conducted and audio-taped a total of 20 interviews, using a questionnaire comprised of questions from draft instruments at the University of Maryland Survey Research Center. Each of the interviewers had five or more years of experience; three of the four had PhD's in psychology. The interviewers were asked to use whatever

methods they normally employ to prepare for and conduct cognitive interviews. The only study constraint was the format of their reports; the reports had to be uniformly structured in a way that would permit the planned analysis. The interviewers wrote reports of each interview, listing and describing problems they had uncovered for each question. The interviews were transcribed and two coders jointly labeled each conversational turn taken by the interviewer and by the respondent in all of the interviews. The interviewer codes most relevant to current purposes are: generic prompts to keep talking; instructions to paraphrase or define terms; probes about a previous respondent utterance or behavior such as "you paused for a while there; can you tell me why?" or "why can't you answer the question?" The respondent codes most relevant to current purposes are: adequate answers; inability to answer; long pauses; indications of uncertainty; requests for clarification; request for a question to be reread; explicit descriptions of problems; and confirmation of a problem described by the interviewer. The two coders reached agreement on all assigned codes.

Note that this type of coding is finer grained than standard "behavior coding" often used in pretesting (e.g. Fowler & Cannell, 1996). In behavior coding, it is the presence of particular behaviors (such as interviewers reading the question not as worded) that are noted for a particular administration of the question. For interaction coding of the sort we carried out, it is not the behavior associated with the administration of the question, but the sequence of behaviors at each conversational turn that is of interest. (See various chapters in Maynard, Houtkoup-Steenstra, Schaeffer & van der Zouwen, 2002, for examples of this approach).

In addition, four independent judges evaluated each question-administration, i.e. a particular question asked in a particular interview, for the existence of problems and, when problems were detected, they classified them using a problem taxonomy (Conrad & Blair, 1996). This allows us to see if particular interactions led the judges to believe those interactions contained problems. Finally, so that we could compare the judges' evaluations to those of the interviewers, two research assistants working together mapped the interviewers' written reports to the problem categories in the same problem taxonomy.

Results

Our first research question concerns how often interviewers' probes produce respondent statements that indicate the presence of a problem. Does it take many probes to uncover a small number of problems or are the problems detected whenever interviewers probe? In the following interaction, the interviewer prompts the respondent to verbally report on her thinking (line 5) after which the respondent describes a potential problem (lines 6-8, 10, 12 and 14). The gist of the problem is that the question, which concerns racial attitudes, uses the phrase "in general" and then focuses specifically on White people's attitudes toward Black people. On methodological grounds, this an ideal interaction in that, after minimal, non-substantive prompting by the interviewer, the respondent articulates not only that the question has a problem but also exactly what the problem is.

(1) Prompt to think aloud followed by description of potential problem[2]

01 **I:** In general, how do you think people in the United States feel about people of other races? Do you think
02 only a few white people dislike blacks; many dislike blacks or almost all white people dislike
03 blacks?
04 **R**: I'd say only a few.
05 **I:** Okay, and what do you think about when you answer that question?
06 **R**: Um. Well I wonder why they, they're using "In general". I think that it's very much in general. How do
07 people, in the United States feel about people of other races. And then we go to a specific. White
08 people.
09 **I:** Mmhm.
10 **R**: and how they feel about blacks.
11 **I:** Okay, okay.
12 **R:** It just seems that um . It just seems again so overly simplified.
13 **I:** Mmhm. Okay.
14 **R:** I find it, I just, I'm in, I'm creeping out (inaudible) is, who wrote the question. Why they choose this
15 specific?

[2] In these examples, a period indicates a long pause, a question mark indicates rising intonation, a colon indicates a drawn out vowel, and capitalization indicates increased speaking volume. In the interest of clarity and brevity we have removed small portions of the interaction and indicated this "(portion omitted)." In all cases these passages were not relevant to the example or they repeated previous material. Full transcripts are available from the authors.

Our second question concerns the degree to which judges are persuaded that interactions that contain respondent descriptions of problems actually contain problems. We compare the judges' decisions for two types of interactions that include such descriptions but which vary in what preceded those descriptions. In one type of interaction the preceding exchange included some evidence of a problem by the respondent and a probe about that evidence, i.e. an evidence-based probe. In the other, the preceding exchange included an adequate answer by the respondent followed by a context-free probe. Our expectation was that problem descriptions preceded by evidence-based probes would be more convincing to independent judges than those preceded by context-free probes. This is because the latter may have the character of a hypothetical confirmation of a hypothetical problem, rather than something the respondent actually experienced while answering that particular administration of the question. If this is the case, fewer judges will classify questions as containing problems when the probes are context-free than evidence-based.

The following is an example of an interaction in which the respondent provides evidence of a problem (line 9), the interviewer probes about the evidence (line 10) and the respondent describes a possible problem (line 11). The respondent's problem seems to be that her lack of knowledge about drug use in the neighborhood could indicate a lack of urgency or just a lack of knowledge. The point is that the probe concerns explicit evidence of a problem, i.e. her inability to answer

(2) Evidence-based probe followed by description of potential problem

I: For each problem listed below, please indicate to me whether it is a very urgent problem, a somewhat urgent problem, a small problem, or not a problem at all in your neighborhood. Drug abuse.
R: . Mm.
I: Tell me what you're thinking.
R: Well, I haven't, heard of, any, well. I haven't heard of any, drug use in my neighborhood
I: Okay.
R: Um. So, um.
(portion omitted)
I: Okay.
R: In my neighborhood. Um, so, um. . I don't know how to answer that. I don't, know.
I: Why don't you know how to answer it?
R: Just, because since, I:, don't, hear of ANY, I don't know like, what answer you're looking for, because.
I: What, whatever it means to you. Again, the question, for each problem listed below, please indicate to me whether it is a very urgent problem, a somewhat urgent problem, a small problem, or not a problem at in your neighborhood and the first one
(portion omitted)
R: Oh, it's a very urgent problem.
I: Okay.

In contrast, in the following interaction the respondent provides an acceptable answer (line 3) and then the interviewer administers a context-free probe (line 4); the respondent next describes a possible problem (lines 5-7 and 16). The probe asks the respondent to paraphrase the question; the problem that he subsequently articulates is that the answer – the degree of threat to ones health posed by AIDS – could depend on the person's size and health. The fact that the respondent has already provided an acceptable answer underscores the lack of evidence that he had experienced a problem prior to the probe. Of course one can experience a problem without being aware of it or manifesting it behaviorally but our concern is with explicit indications of problems.

(3) Context-free probe followed by description of potential problem

I: Okay, would you say that getting AIDS is an extremely serious threat to a person's health, a very serious threat, somewhat serious or not too serious?
R: Um, extremely serious
I: Okay and what were you, what do you think that question was asking?
R: Um, I don't know, I'm kind of confused maybe, extremely or not extremely. Like, depending on like, the size of the person, or like the health condition they're in. Like how healthy they are, like their age
(Portion omitted)
I: Okay, So you're saying that, to you, this question was asking sort of like, um, what's the chances, or how, how fast is it going to progress or what are the chances that the person will die once they got it, or something like that?
R: Mmhm.
I: Okay, so I'll just put here, what are the chances of dying and or how fast and like, depend on,

13 you know size or health. So, you said, it's kind of confusing how you would answer that.
14 **R:** Mmhm
15 **I:**Because it depends.
16 **R:**There's like two answers.

To address our first point we will concern ourselves with the most frequent type of probe, namely those about the meaning of particular terms or phrases. These accounted for 41% (405 out of 987) of all probes. We examined 257 interactions that included a probe about meaning followed by a respondent utterance relevant to the status of a problem. In particular, this respondent utterance either (a) specifically described a problem, (b) gave no explicit indication of a problem, or (c) explicitly indicated there was not a problem, e.g. the respondent's paraphrase was consistent with the meaning that seemed to have been intended by the author. In 89.9% (231 out of 257) of these interactions, the respondent's subsequent utterances were of types (b) and (c), i.e. either gave no explicit indication of a problem or explicitly indicated there was not a problem. In only 10.1%, was the probe followed by a description of a specific problem, i.e. utterance type (a). Thus, the interactions that included the most frequent type of probe (requests for paraphrases and other inquiries about meaning) did not often turn up respondent descriptions of problems. The issue then is to what degree these descriptions reflect actual problems. On the one hand, the relatively low ratio of specific problem descriptions to probes could indicate that frequent probes about meaning serve as a kind of filter through which only unequivocal problems pass. On the other hand, it could be that such probes have an arbitrary character, which could, in turn, elicit speculative descriptions of problems.

Our second set of analyses begins to address this issue. Based on the possibility that problem statements – utterance type (a) – might be more convincing to judges when preceded by evidence-based rather than context-free probes (example interactions 2 and 3 respectively), we compare the number of these interactions that are classified as containing problems by independent judges. For current purposes, we treat the judge's decisions as authoritative assessments of problems though there are certainly other ways to assess the validity of problems (see Conrad & Blair, in press).

The most liberal criterion for considering a question-administration to contain a problem is if at least one judge decides that it does. By this measure, 79.3% (23 out of 29) of the question-administrations in which a respondent problem description was elicited by an evidence-based probe were judged to problematic. In contrast only 53.8% (14 out of 26) of the administrations in which a problem description is elicited by a context-free probe were judged to be problematic ($X^2[1]=5.91$, $p < .05$). Although the sample sizes are small, there is clear evidence of a difference in how persuasive problem descriptions are when they follow different types of probes. So even though, in the first analysis, only a small portion of context-free probes elicited problem descriptions, the interactions containing such descriptions, in the current analysis, are not judged to reflect the presence of problems as often as interactions that first include evidence of a problem.

Not only were the judges less often persuaded that problem descriptions after context-free probes actually indicate a problem but the interviewers themselves seemed to be sensitive to this distinction. Based on the final reports they wrote about their own interviews, interviewers judged there to be problems in 75.9% (22 out of 29) of the question-administrations involving problem statements after evidence-based probes. In contrast they judged there to be problems in only 61.5% (16 out of 26) of the question-administrations involving problem statements after context-free probes ($X^2[1]=4.31$, $p < .05$).

If we require that multiple judges concur about the presence of a problem, the percentages of question-administrations that are judged to reflect problems drop off precipitously. The scores in Table 1 decrease from left to right suggesting that the judges are not uniformly persuaded about the presence of problems. Moreover, when we examine those interactions that more than one judge considered indicative of a problem (columns 2-4), then the judges as a group are far less convinced about the presence of problems than are the interviewers who reported them. Recall that interviewers reported 75.9% and 61.5% of the interactions that correspond to rows 1 and 2 respectively to indicate problems with the question. The entries in columns 2-4 are smaller are smaller than the figures for interviewers. Apparently interviewers have a lower threshold for what they consider to indicate a problem than do independent judges whose function is not necessarily to find problems. Note also that even when we restrict our analysis to those relatively few interactions that multiple judges considered indicate problems, the counts are smaller when respondents do not provide evidence of problems (second row) than when they do (first row).

Problem statement after probe after	At least one judge	At least two judges	At least three judges	Four judges
Evidence of problem	79.3 (27)	55.2 (16)	27.6 (8)	3.4 (6)
No evidence of problem	53.8 (14)	26.9 (7)	11.5 (3)	0 (0)

Table 1. Percent interactions that judges considered indicated a problem (number of interactions in parentheses).

Discussion and Conclusion

These results indicate that the discussion in cognitive interviews about possible problems sometimes leads to verbal reports that indicate the presence of a problem. We propose that when such problem descriptions occur, they may reflect actual problems or they may reflect respondents' acquiescence to interviewer speculation about problems. The difficulty in distinguishing the former situation from the latter is potentially a serious consequence of context-free probing.

It seems possible that what respondents say is quite sensitive to what interviewers say and so respondents may well consent to what seems like a knowledgeable interviewer suggestion. Moreover, when they have already given what they perceive to be an adequate answer to the question, respondents might infer that the interviewer probe indicates that the interviewer believes there is a problem. After all, the respondent might reason, the interviewer would only probe after having been provided with a reasonable answer if there was something wrong with the answer. In ordinary conversation, speakers are obliged to contribute new and relevant information with each successive turn (e.g. Grice, 1975). Thus, to be cooperative (in Grice's sense) a respondent who has previously indicated no problem might well respond to an interviewer's probe by describing a potential problem. This could be the source of the judges' apparent skepticism that problems exist when their description is not first indicated in overt respondent behavior. It raises the possibility that interviewers may be "detecting" problems in interactions that judges with more distance from the exchange do not detect. If there is not agreement about the presence of a problem, revising the question wording may not be justified. As we have already suggested, unwarranted changes to question wording are costly and can themselves introduce problems. Thus questionnaire designers might more confidently revise questions for which interviewers have reported problems if those reports are based on explicit evidence from respondents or if multiple judges agree on the presence of a problem.

The current study indicates that that some patterns of interaction are more effective than others in uncovering compelling problems with survey questions. If this is supported by further studies, then examples of effective patterns, and how to foster them, could be helpful in training cognitive interviewers. Beyond suggesting ways in which cognitive interviewing techniques can be improved, the current study has begun to explore the processes that underlie respondents' reports of possible problems. This kind of effort should help guide the use and interpretation of the rich information produced by the technique.

References

Conrad, F.G. & Blair, J. (in press). Aspects of data quality in cognitive interviews: The case of verbal reports. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin & E. Singer (Eds.) *Questionnaire Development, Evaluation and Testing Methods*. New York: John Wiley and Sons.

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

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

Fowler Jr., F.J., & Cannell, C.F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp. 15-36). San Francisco, CA: Jossey-Bass.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.) *Syntax and Semantics: Volume 3, Speech Acts* (pp. 41-58). New York: Academic Press.

Maynard, D. W., Houtkoup-Steenstra, H., Schaeffer, N. C. & van der Zouwen (Eds.) (2002). *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*. New York: John Wiley and Sons.

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

U.S. Bureau of the Census (1998). *Pretesting Policy and Options: Demographic Surveys at the Census Bureau*. Washington, DC: US Department of the Census.

Willis, G. B., DeMaio, T. J., Harris-Kojetin, B. (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: John Wiley and Sons, pp.133-153.