The effect of probe type on cognitive interview results: A signal detection analysis

Rachel Levenstein[1], Frederick Conrad[2], Johnny Blair[3], Roger Tourangeau[4], Aaron Maitland[4]
[1]Institute for Social Research, University of Michigan, 426 Thompson Street, Suite 4050, Ann Arbor, MI, 48104
[2]Institute for Social Research, University of Michigan
[3]Abt Associates
[4]Joint Program in Survey Methodology, University of Maryland

1. Introduction

Before survey production begins, several methods are used to pretest the questionnaire in an effort to reduce measurement error. One such method is the cognitive interview. It is seen as a quick, cost-effective way to find major errors in questionnaire design (Willis, 2005; Presser and Blair, 1994). The technique typically involves asking a small number of test respondents to think aloud and respond to probes while answering questions in a draft questionnaire. The think aloud protocols are examined for evidence of possible problems for test respondents which might also be problems for respondents in production interviews.

Few guidelines exist for this technique, and as a result, the procedure varies widely across survey organizations. Blair and Presser (1993) found that, among organizations that use cognitive interviews to pretest survey questionnaires, cognitive interviewers are frequently highly experienced with advanced degrees and understanding of questionnaire design issues. As a result, these highly experienced interviewers are given flexibility about when to probe and what probes to use.

Although cognitive interviewing has become a best practice in many survey organizations, relatively little is known about its effectiveness. An important gap in our knowledge concerns the degree to which the results of cognitive interviews depend on who conducts the interviews. If different interviewers systematically identify different problems or different numbers of problems from one another this creates a kind of interviewer effect. In practice, a cognitive interview pretest is usually conducted by a single interviewer so interviewer effects will not be discernable. However, the possibility that the choice of interviewer may affect what problems are identified could undermine the effectiveness of this method.

In production survey interviews, observable interviewer attributes have been shown to affect the direction of answers when the attribute is related to the topic of the questions. For example, female interviewers elicit more liberal responses on questions about feminism than do male interviewers (Kane & Macaulay, 1993). Similarly, in cognitive interviews, an interviewer who appears to be knowledgeable and authoritative about questionnaire design and inquires about possible problems may be more likely to elicit confirmations from respondents that they have experienced those problems than if the interviewer appears to be less qualified to detect problems.

In addition to their observable characteristics, production interviewers have been shown to affect the data they collect through their behavior. Specifically, items that require more probing have been shown to increase interviewer-related response variance (Mangione, Fowler & Louis, 1992), presumably because probing is the least scripted part of the interaction, and different interviewers carry it out differently. In order to reduce exactly this type of interviewer effect, standardized interviewing is advocated (e.g. Fowler & Mangione, 1990). To reduce the impact of any one interviewer on overall results, the recommendation is to increase the number of interviewers and decrease the workload of each interviewer (e.g. Mangione, et al., 1992). In contrast, the practice of cognitive interviewing involves very few interviewers with a great deal of freedom to probe (Conrad & Blair, 2004; Blair & Presser, 1993).

When production interviewers ask a question, they follow a script, which, in theory, makes interviewers interchangeable. In contrast, cognitive interviewers are allowed the freedom to probe as they choose. They decide when to probe and what words to use, increasing the chances that they will do this differently. The findings of Mangione, Fowler, and Louis (1992) may extend to cognitive interviews; it is possible that how interviewers probe will affect respondents' verbal reports and whether the respondents indicate there is a problem with the question.

Our research tested the effect of both observable characteristics and interviewer behavior in cognitive interviews on the sensitivity of the method, i.e. how well it detects actual problems and the degree to which

it "detects" spurious problems. Specifically, we examined the effect of the cognitive interviewers' apparent experience and their probing behavior on the identification of problems.

2. Method

2.1 Study Design Overview

In this study, eight interviewers completed 60 cognitive interviews (each interviewer conducted 6 to 8 interviews) under several conditions. We manipulated the perceived experience of the interviewer; interviewers were either introduced as having a great deal of experience (high expertise) or as having just been trained (low expertise). The high expertise interviewers also wore white laboratory coats, which we hoped would increase their perceived authority. In addition, we manipulated the probe type used by any one interviewer. Half of the interviewers were trained to use *directive probes* that presupposed the presence of a problem; the other half were trained to use *generic probes* that did not (see below for a more complete description). Finally, we created two sets of questionnaire items: one set consisted of items borrowed from production questionnaires which had all been pretested. We believed them to be relatively free of problems. For the other set, we "damaged" these items, i.e. changed the wording so that the question would likely create a particular problem for respondents. A mixture of damaged and undamaged items was counterbalanced between two questionnaire versions (Version A and Version B). See Table 1 for a summary of the design.

Table 1. Design Summary.

	Generic Probes (4 interviewers)	**Directive Probes (4 interviewers)**
Low Expertise	Version A (iwers 1 and 2)	Version A (iwers 5 and 6)
	Version B (iwers 3 and 4)	Version B (iwers 7 and 8)
High Expertise	Version A (iwers 1 and 2)	Version A (iwers 5 and 6)
	Version B (iwers 3 and 4)	Version B (iwers 7 and 8)

We would expect that high expertise would produce more reports of problems, regardless of whether a "true" problem was present. We would also expect a main effect of probe type in that respondents who heard a directive probe would be more likely to have a problem answering the question, even if that question was considered by the researchers to be problem-free. Finally, we expected an interaction of probe type and expertise; that is, a respondent might be particularly prone to have a problem answering a question when an interviewer who appeared to be experienced asked a directive probe.

2.2 Respondents

60 respondents were recruited through the Craigs List websites for Ann Arbor and Detroit, MI. 39 were women, and 21 were men. The respondents were, on the whole[1], highly educated. Two had only completed high school or a GED, three reported some college or an associate's degree, 19 held bachelor's degrees, and 9 held post graduate degrees. Many of the respondents were also students; eight were full-time undergraduates and 12 were either part-time undergraduates or full- or part-time graduate students. The respondents were also relatively young (mean=31.5, SD=11). 50 of the respondents were white (non-Hispanic), 3 of the respondents were black (non-Hispanic), 3 were Asian, and 4 were Hispanic.

2.3 Interviewers

Eight experienced production interviewers participated in a four-hour training program on cognitive interviewing. There were four male and four female interviewers, all 40 years old or older. The interviewers had varying levels of conventional production interviewing experience, ranging from one to 15 years. All of the interviewers had completed at least some college, with five of them having a college degree.

2.4 Introductions

Interviewers were introduced as having more or less expertise in questionnaire design. Before meeting the interviewer, a staff member read some information to the participant about the interviewer. In the low expertise condition, the interviewer was referred to by his or her first name only and was described as having recently completed a training program and as still learning about survey questionnaires: "(interviewer's first name) will be interviewing you today. She recently took part in a training program on interviewing skills and techniques. She's still

[1] We inadvertently omitted the question about the highest level of education attained for 27 of the respondents. Because the assignment of respondents to experimental condition was arbitrary, we believe this educational level was roughly balanced.

learning about developing survey questionnaires." In the high expertise condition, the interviewer was described with a title (Mr. or Ms.), and the experience of the interviewer was emphasized: "Mr. /Ms. (interviewer's last name) will be interviewing you today. She has a great deal of experience with survey research. She has been extensively trained in interviewing skills and techniques. As a certified interviewer, her skills in this kind of interviewing have been extremely valuable for the Survey Research Center. She is highly valued for her knowledge of good questionnaire design and provides her expertise to lead researchers."

2.5 Questionnaires

23 questionnaire items were taken from existing surveys. As mentioned above, each item had been tested in the field and has been in use at a survey organization. We believed these to be relatively problem-free. We also deliberately damaged each item, creating a set in which we believed there to be at least one problem per item. An expert panel reviewed the damaged questions and the undamaged questions and concurred with our assessment of which items did and did not have a problem. We created separate versions of the questionnaire that counterbalanced the damaged questions so that half of the questions were damaged in each version. An example of an undamaged question and its damaged counterpart is:

Undamaged: *Are you optimistic or pessimistic that the next generation will live in a better world than we do now in terms of the environment?*

Damaged: *Do you think that the next generation will live in a better world than we do now in terms of the environment, including air and water quality, biodiversity, nonrenewable resources, and global warming related to hydrocarbons?*

Here, experts identified the multi-barreled nature of the damaged question as well as the prevalence of technical terms.

Each version of the questionnaire also included one of two types of scripted probes, creating a total of four versions. The probes were either 1) directive, i.e. they presupposed the existence of a problem in the question, or 2) generic, i.e. they merely asked about the cognitive process of answering the question. For example, if the above item was asked in either the damaged or undamaged wording, the following probes might be used:

GENERIC: How did you come up with your answer?
DIRECTIVE: What words in the question didn't you understand?

There were actually two types of directive probes: "specific" and "general." A specific-directive probe asked about a particular problem as in the example above. A general-directive probe asked about a problem without naming it, but the wording implied a problem was believed to exist: "What parts of the question were hard to understand?" About half of the directive probes were generic and half specific; type of directive probe constructed for a particular item was randomly determined. As it turned out, the distinction between general-directive and specific-directive probes did not affect the results, and we do not discuss it further.

Interviewers were permitted to ask respondents to "Tell me more" or "Tell me what you're thinking" if they felt the respondent had not provided codable information up to that point.

2.6 Procedure

Upon arriving at the Survey Research Center laboratory, respondents were taken to one of two rooms depending on the experience manipulation level of the interview. The room used for the high expertise interview was larger and had more comfortable furniture; the room also had a computer. The room used for the low expertise condition was smaller with more utilitarian furniture and no computer. We hoped to make the high expertise interviewer seem of higher stature because of the greater space and the equipment, thus strengthening the manipulation.

A research staff member read the appropriate experience manipulation introduction to the respondent and left the respondent in the room alone for a moment to invite the interviewer into the laboratory. The staff member introduced the interviewer per the expertise manipulation described above. The staff member then left the room.

All interviews were recorded with an audio Sony IC recorder. The interviewer began by providing examples of thinking out loud and asking the respondent to practice thinking out loud. The interviews lasted an average of 23 minutes. One interview had to be discarded due to technical difficulties; the final number of codable interviews was 59.

After each interview, the respondent was asked to rate the interviewer in a self-administered questionnaire on the dimensions of knowledge, experience, professionalism, and level of education.

2.7 Problem Coding

Two coders independently listened to each interview and coded whether a problem existed or not in each question for each interview. Presser and Blair (1994) created a coding scheme to classify problems in pretest interviews. This scheme allows the coder to identify problems specific to the respondent or the interviewer. For this analysis, we are interested in the presence or absence of a problem for the respondent and whether the observed problem matched the one we had introduced; therefore, all codes were reduced to "problem" or "no problem" for a particular administration of a question.

Agreement between the coders was quite high in determining whether there was a problem with the question for a given respondent (proportion agreement=0.85, κ=0.66) and in determining whether the problem matched the damage inflicted (agreement=0.62, κ =0.63). Before proceeding with the analysis, the coders reconciled any differences. All analyses are based on the reconciled codes.

3. Results

3.1 Do Directive Probes Increase The Odds Of Identifying A Problem?

In order to understand what might predict the odds of a problem being identified, a logistic regression was run regressing the odds of a problem being identified on probe type, expertise, whether the item was damaged, an interaction between probe type and expertise, and an interaction between probe type and damage. We included the final term in the model to see if the odds of having a problem for undamaged versus damaged questions was different with directive than generic probes. Our thinking was one kind of probe might be better at detecting actual (with damage) than spurious (without damage) problems. The model indicated that probe type (OR=4.35, p<.000) and damage (OR=3.42, p<.000) both significantly predicted the odds of a problem being identified. Contrary to our expectations, expertise did not predict the odds of a problem being identified (OR=1.26, p=n.s.), nor did the interaction of expertise and probe type (OR=.913, n.s.). We had expected these factors to interact because a respondent might be particularly prone to acknowledge the presence of a problem when an experienced interviewer asked a directive probe. Finally, the interaction of probe type and damage was marginally significant (OR=.63, p=.07). That is, the difference in odds between directive and generic probes is smaller for true problems than in false problems.

3.2 Sensitivity of Cognitive Interviewing

It is possible that there is an upside to the increase in false alarms from directive probes. If directive probes generally increase the likelihood that respondents will produce evidence of a problem, then when a problem is actually present the hit rate should increase. It is of course possible that directive probes have their impact on only the false rate and do not increase hits. Signal detection theory can allow us to evaluate such differences.

Often used in engineering, signal detection theory was developed to evaluate, for example, a technician's ability to detect a true signal on a radar screen or a radiologist's ability to accurately make a diagnosis based on an x-ray (e.g., Green & Swets, 1966).

Every time a respondent answered a given question, four outcomes could occur--a *hit*, *false alarm, miss*, or *correct rejection* (MacMillan, & Creelman, 2005) . A hit occurred when the question was damaged *and* the respondent had a problem that reflected that damage. When the question was undamaged *and* the respondent had a problem, regardless of whether the problem matched the damage, then we counted it as a false alarm. A miss occurred when the question was damaged and the respondent *either* had some other problem unrelated to that damage *or* had no problem at all in answering the question. A correct rejection occurred when the respondent had no problem *and* the question was undamaged.

Based on our calculations of hits and false alarms, we can determine how well cognitive interviews can discriminate between actual problems and nonproblems, i.e., respondents' experience with questions that are assumed to be problem-free. Table 2 displays the false alarm rates, and Table 3 displays the hit rates.

Table 2: False alarm rates (frequency).

	Generic Probes	*Directive Probes*	*Overall*
Low Experience	0.159 (27)	0.440 (74)	0.299 (101)
High Experience	0.169 (27)	0.468 (80)	0.323 (107)
Overall	0.164 (54)	0.454 (154)	0.311 (208)

Table 3: Hit rate (frequency).

	Generic Probes	*Directive Probes*	*Overall*
Low Experience	0.274 (48)	0.411 (72)	0.343 (120)
High Experience	0.333 (49)	0.461 (76)	0.401 (125)
Overall	0.301 (97)	0.435 (148)	0.37 (245)

Sensitivity (*d'*) is defined as the difference between the normalized hit rate and the normalized false-alarm rate (e.g., Macmillan & Creelman, 2005). To determine if *d'* is significantly different from zero, confidence intervals were constructed using Gourevitch and Galanter's (1967) formula for the variance of *d'*. Sensitivity for the entire sample and for subgroups was effectively zero; that is, the hit rate was equal to the false alarm rate. Table 4 gives the values of *d'* and its variance.

Table 4: Sensitivity (variance).

	Generic Probes	*Directive Probes*	*Overall*
Low Experience	0.4 (0.66)	-0.06 (0.10)	0.13 (0.10)
High Experience	0.59 (0.59)	0.04 (0.08)	0.27 (0.08)
Overall	0.49 (0.31)	-0.01 (0.05)	0.16 (0.04)

3.3 Post-Session Questionnaire

The post-session questionnaire was used to determine the efficacy of the expertise manipulation. Four regression models show that the expertise manipulation was not associated with knowledge ($F(1,55)<1$, *n.s.*), perceived experience ($F(1,55) < 1$, *n.s.)*.), professionalism ($F(1,58)<1$, *n.s.)*. , or perceived level of education ($F(1,58)<1$, *n.s.)*. That is, both low and high expertise interviewers were rated as being equally knowledgeable, experienced, professional, and educated. Thus it may well be that the absence of any effects of apparent expertise were actually do to not adequately manipulating this factor.

4. Discussion

This study shows that the behavior of cognitive interviewers can increase the odds of the respondents being classified as having a problem with a particular question. Specifically, interviewers whose probes presupposed problems were more likely to elicit respondent reports of problems than were interviewers whose probes merely asked about thought processes. We find this troubling from a data quality perspective because directive probes are among the tools used by cognitive interviewers in current practice. While it is possible that respondents might reject the interviewers' implication that they have experienced a problem, in this study respondents were more likely to accept it than to produce similar evidence of a problem when asked generically about problems. This has much of the character of acquiescence reported in the measure of opinions (e.g., Schuman & Presser, 1981).

The expertise manipulation did not affect whether respondents were classified as having a problem with the question or not. This may be because the manipulation was not strong enough. Perhaps respondents believed that all of the interviewers were highly experienced by virtue of their being employed at a university facility. In this study, we were only concerned with reactions to *perceived* expertise; perhaps respondents may also be differentially receptive to interviewers who are actually more and less experienced.

Another worrisome finding from the current study is that our cognitive interviews, irrespective of the type of probe, were unable to discriminate actual problems from non-problems. To the extent that this extends beyond the current study—and we believe that it may well do so—this suggests that cognitive interviewing may trigger modifications to questions that really do not warrant revision. If cognitive interviewers do not probe about specific problems, the risk of overlooking actual problems increases because criterion for what counts as a problem is raised. There may well be an optimal point at which the criterion can be set balancing detection and rejection of respondents' evidence of problems—perhaps this is the case for skilled cognitive interviewers—and that was simply not represented in the current experiment. We see the tuning of this criterion as the next step in research on cognitive interviewing.

References

Blair, J., & Presser, S. (1993). Survey procedures for conducting cognitive interviews to pretest questionnaires: A review of theory and practice. Proceedings of the American Statistical Association, Survey Research Methods Section.

Conrad, F.G. & Blair, J. (2004). 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, pp. 67-88.

Gourevitch, V, & Galanter, E. (1967). A significance test for one-parameter isosensitivity functions. *Psychometrika, 32,* 25-33.

Kane, E.W., & Macaulay, L.J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly, 57,* pp. 1-28.

MacMillan, N.A., & Creelman, C.D. (2005). *Detection theory: A user's guide.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Mangione, T.W., Fowler, F.J., & Louis, T.A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics, 8*(3), pp. 293-307.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology, 24*, pp. 73-104.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context.* New York: Academic Press.

Willis, G.B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks: Sage Publications.