Efficient Respondent Selection for Cognitive Interviewing

Allison Castellano Ackermann and Johnny Blair
Abt Associates, Inc.

Abstract

This paper explores a seldom considered area of cognitive interview pretesting best practices: sample selection. The questions we consider are: Do different types of respondents produce different results? What criteria should guide the selection of cognitive interview respondents? Often a convenience sample is selected for cognitive interviewing, possibly with some attention to the "variety" of the sample. Due to resource constraints, the number of cognitive interviews is generally small, making it impossible to cover the full range of types of respondents. This paper explores how to most efficiently stratify a quota sample based on respondent characteristics. To explore this idea, we analyzed data from 90 cognitive interviews. The implications of these results for practice are discussed.

Keywords: cognitive interview, sample selection, respondent characteristics, pretest

1. Introduction

Although cognitive interviewing has been accepted as a valuable pretesting technique by researchers from academia and government to marketing and medicine, there is still a lot of debate on what should be considered "best practices" (Presser et al., 2004). For example, should cognitive interviewing primarily aim to confirm the results of an expert review of the questionnaire, or should it aim to discover new problems (Willis, 1999)? Is it better to use scripted probes, which keep cognitive interviews somewhat standardized, or to rely more on think alouds and generic probing (Conrad and Blair, 2001)? Is it best to engage the respondent in discussion as he or she answers the questions or should the cognitive interviewer wait until the end of the interview to debrief the respondent (Redline et al., 1998)? In addition to alternative ways to conduct cognitive interviews, there are other aspects of the pretesting process that can interact with these protocol factors to affect the method's efficiency and results. Such additional aspects include interviewer selection and training, sample size, analysis or review procedures and respondent selection criteria.

Cognitive interviewing is based on the production of verbal reports about the response process. Inability to perform any of the response tasks can result in answers that are inaccurate, sometimes in a minor way, sometimes seriously. The amount written on cognitive interviewing topics has grown steadily in the past two decades. Ericsson and Simon's 1984 book, *Protocol Analysis: Verbal Reports as Data,* discusses verbal reports, which are sometimes used in psychology to understand high level mental processes. More recently *Methods for Testing and Evaluating Survey Questionnaires* (Presser et al. 2004), included a large section devoted to cognitive interviewing. The topics covered range from interviewing techniques (DeMaio and Landreth 2004) and interviewer effects (Beatty 2004) to the reliability of cognitive interview results (Conrad and Blair 2004). Finally, Gordon Willis (2005) published a book on cognitive interviewing. One goal of these works is to explore a variety of cognitive interviewing practices.

1.1 Respondent Recruitment for Cognitive Interviewing

We explore an important but previously unanswered question about cognitive interviewing best practices: Do different types of respondents produce different results? In addition, we explore a related issue--what mix of respondents should be enlisted for participating in cognitive interviews? One way to maximize the value of small sample sizes typically used in cognitive interviewing may be by stratifying a quota sample based on easily-determined respondent characteristics.

For surveys with a special, non-general target population, there may be little or no latitude about who should be used for questionnaire

testing. If a survey of elementary school teachers is to be done, then cognitive interviews must be performed with such teachers. Though even in this case, factors such as teachers' backgrounds and numbers of years of experience may be useful to take into account.

However, general population surveys contain a variety of respondent demographic types. Due to budget and time constraints, the number of cognitive interviews performed for any given project is generally somewhat small. Additionally, interviews conducted under government contracts prior to OMB clearance are subject to the "rule of nine.[1]" Under such severe constraints, efficient guidelines for choice of respondents can have important consequences.

Usually a convenience or quota sample is selected for cognitive interviewing, with some attention to the "variety" of the sample. It seems logical to expect that respondents more knowledgeable about the survey topic may uncover problems that less knowledgeable respondents do not notice; conversely, less knowledgeable respondents may have difficulty with some questions that are easy for the more expert respondents. In addition, education level may function in a similar manner, with education level being related to sensitivity to alternative question wording or ability to deal with complex sentence structure, and to articulate potential problems with either or both. But is there empirical support for these conjectures? And, if so, are there efficient ways to decide on the "mix" of respondent types?

Some recent project experiences suggest that respondent differences may matter a great deal in cognitive interviewing. While cognitively testing health-related surveys, we considered two types of respondents. One type was contacted because of his or her association with a particular special health condition support group. Other respondents were from the general population with no special experiences or expertise. The idea behind recruiting both types was that, between the two, a wide range of potential problems with the questionnaire should be identified. When using these two types of respondents to test health-related questions intended for a general population, we found that they did, indeed, have different types of problems and insights while responding to the same questions.

Yet another example of respondent differences was found in the results of testing a questionnaire to be administered to a company's employees to rate the leadership skills and other attributes of peers, supervisors, or subordinates. In this case, the divide between respondent subgroups was junior-level staffers and higher-level supervisors and directors. The respondents were given a list of employee qualities, such as "responds non-defensively to feedback." The cognitive interviewing indicated that junior staffers had a lot of difficulties with the jargon that was more familiar to senior-level staff managers who also had more experience with managerial concepts and understood the vague wording of, for example, the leadership qualities asked about in the instrument.

Based on this experience, we concluded that there may be other ways in which respondents might differ that would also affect pretest results. There are several dimensions along which cognitive interview respondents differ that may affect performance. For example, cognitive abilities (narrowly defined), such as when the elderly have less short-term memory capacity, may lead to different respondent performance than with younger respondents. Longer and more varied life experiences (e.g., the elderly have longer health histories) may lead to those respondents knowing more about some health issues seldom experienced by younger respondents. Conversely, this same factor may also lead to more difficulty recalling personal health experiences that span many years. A different type of respondent characteristic might be cognitive tendencies due to gender or cultural factors (e.g., some men generally, or Hispanic men in particular, being less willing to rate their health as "less than average" or to interpret such a question differently from women or people from other ethnic backgrounds).

Basic demographics seemed a reasonable basis for testing the idea that respondent differences could lead to different pretest results. The value of this approach would be that a sample efficiently stratified on demographic characteristics might provide much better coverage of question problems for little

[1]Under OMB guidelines, no more than nine interviews can be conducted using a single version of an instrument without OMB approval.
.

additional cost. The particular demographics may differ from one study to another. One would not expect sex to be a factor in most instances, e.g. differences in cognitive abilities like recall or short-term memory. However, age might be such a factor for the reasons just mentioned. It is more likely, though, that there are cognitive differences among people with varying education levels. Higher education is generally associated with better reading comprehension, larger vocabulary, and better abstract analytic skills. For this reason, one might expect less educated respondents to have more problems with comprehension, reasoning, mental arithmetic and other kinds of estimation, for example. But these are empirical conjectures.

2. Study Design

To test whether there are significant differences in the cognitive interviewing results produced from different respondent types, we looked at data from a study that involved 90 cognitive interviews. The original purpose of this study was to test the effects of sample sizes on cognitive interviewing results (Blair, Conrad, Ackermann & Claxton, 2006). The questionnaire covered many topics (e.g., attitudes toward the environment, reading behavior, internet use, and health), and all respondents were recruited by e-mail invitation. The background data on each of the 90 respondents included demographic information on age, sex, and education levels. These categories will be the basis for comparison.

Each of the 60 questions in the questionnaire had at least one embedded problem. The 90 cognitive interviews yielded a total of 210 unique problems. Problem coding was done on each interview to see what types of problems, if any, each respondent had with each of the 60 questions. The number of true problems found will be an indicator of the level of productivity of the interview. For the purposes of the present study, the measure of cognitive interview productivity is the number of problems correctly identified in the questionnaire.

2.1 The Questionnaire

The questionnaire's 60 items were compiled from ten sources, designed to contain a wide range of question types, varying both in content (behavior versus attitude) and in format (yes/no versus agree/disagree questions). Questions were borrowed from government surveys, university studies, as well as public opinion polls. Items on employment status were taken from the Current Population Survey (CPS); items on the internet and computers were taken from the CPS Computer Use Supplement; items on health were taken from the Behavioral Risk Factor Surveillance Survey (BRFSS); items on the respondents' opinions of their neighborhoods were taken from the National Survey on Drug Use & Health; items on the economy were taken from the University of Michigan's Institute for Social Research (ISR) Survey of Consumers; and finally items on a variety of public opinion topics were taken from Harris, Gallup, Pew, The New York Times, and CBS.

After compiling the questions, each one was then "damaged" in some way to embed a problem in it. An effort was made to see if cognitive interviewing had already been conducted on any of them. In some cases (e.g., CPS and CPS Computer Use), we were able to find the original questions before they were "fixed" after cognitive interviewing. In these cases, we used the earlier versions of the questions. In other cases (e.g., BRFSS), we found the results of cognitive interviewing, but the version used in the final draft was the same as the original one tested. Sometimes, due to constraints such as budget or comparability to past surveys, questions that are known to be flawed are still used in surveys, especially if the flaws are judged to be somewhat minor. For the rest of the items we purposely implanted problems, so that in the end we had a questionnaire in which every question had some type of problem. We varied both the problem types and their severity, which is their likely affect on answers. For example, we damaged questions by removing the time frame reference, replacing a simple word with a more complicated one, or creating a double-barreled item. Each item had at least one "problem"; some items had two or three such problems.

2.2 The Interviewers and the Interviews

Ten cognitive interviewers from ISR's Survey Research Center conducted the 90 interviews (with nine interviews per interviewer). Ten interviewers were used to lessen the potential impact of the individual interviewers. Interviewers were chosen based on criteria typically used in recruiting production interviewers (i.e., interpersonal skills, speaking,

writing, listening, interest in subject, background of social science, ability to interact with persons of a wide range of demographic backgrounds, etc). However, these ten interviewers had little or no experience with questionnaire design or cognitive interviewing. They were required to complete a training workshop conducted by senior research staff from Abt Associates and ISR. Training involved learning about cognitive interviewing in general, as well as practice exercises with the specific questions used for this study. Since the interviews were to be coded by trained staff, the focus of the training was on correct administration of the cognitive interview protocol rather than on problem identification.

The interview protocol combined think alouds with scripted probes. After the first half of the interviews were completed by each interviewer (i.e., four or five interviews each), the interviewers were asked to examine their protocol probes and to make changes based on what had been discovered in the first batch of interviews. This was done to provide a reasonable parallel to common practice, in which what is learned from some interviews can influence what is focused on in subsequent interviews. The question-specific probes aimed at checking for expected problems; think-alouds left room for new problems to be discovered. The interviews were conducted in November and December of 2004 at the University of Michigan Survey Research Center in Ann Arbor, Michigan.

2.3 The Respondents

Respondents were recruited via an e-mail invitation. 20,000 e-mail addresses were purchased from Genesys, and the invitations were sent out in four weekly waves of 5,000 each. We purposely avoided recruiting participants only from the University of Michigan community to obtain a greater variety of respondents. Potential respondents provided their age, sex, and education level via e-mail screener questions, allowing us to control the distribution of respondents based on these characteristics. The sampling frame may have overrepresented highly educated participants, making it difficult to find respondents with high school only education levels. Our recruitment efforts yielded a wide range of ages and education levels and a mix by gender.

2.4 Coding

The analyses are based on problem coding of the 90 cognitive interviews. Two trained coders, who were experienced with cognitive interviewing and questionnaire design, listened to recordings of each interview, marking each instance where respondents' verbal reports indicated a problem with the question. Both coders independently listened to all 90 interviews and coded them. The coders then compared results and reconciled any differences. We opted for consensus coding to remove inter-coder reliability as a factor and to avoid multiple code s for each question administration, both of which would complicate the analysis. Interviews were also by respondent characteristics. The number of problems found per interview was then used to compare the different respondent types.

A problem found in a cognitive interview was not necessarily a problem that the respondent reported. For example, one question asked, "During the past year and a half, how many books did you read?" One respondent answered, "I'm thinking I read 2 books a month, so let's say 24." The respondent in this case believes he has provided an appropriate answer to the question, but he has answered based on a time frame of one year, rather than a year and a half, as the question asks. Therefore, a problem was coded in this instance. The problem counts include both a respondent-identified and a coder-identified problem.

In addition to coding whether or not a problem occurred, the coders also assigned a problem "type" to each problem. The coding scheme for problem types was taken from Presser and Blair (1994), and included 29 possible codes. The problem codes consisted of two main types: problems with how readily the question is understood (semantics) and problems with retrieving information or formulating an answer (response task). Each problem was also associated with a question-type: behavioral versus attitudinal. There was a relatively even mix of problem types and question types

One final problem characteristic that was coded is problem severity, or its expected impact on measurement error. This variable, which is more subjective than the previous ones, was created from the judgment of three survey instrument design experts. Each expert was asked to rate the severity of each problem identified in the 90

cognitive interviews using a scale of 1 to 10, with 1 meaning "not too severe" and 10 meaning "very severe." The expert ratings of each question were averaged to produce a single score for each question problem. These multiple problem characteristics were used as variables in our analyses.

3. Results

Table 1 shows the mean number of identified problems for each of the demographic subgroups. The only demographic characteristic that had a significant effect on the mean number of problems identified in a cognitive interview was education ($F(2, 87) = 6.497$, $p < 0.05$). The higher a person's education, the higher the mean number of problems he or she identified in a cognitive interview.

Table 1. Mean Number of Problems Found by Sex, Age, and Education.

Characteristics		N	Mean # Problems Found Per Interview
Sex	Male	36	13.7
	Female	54	13.6
Age	18-35	33	14.8
	36-50	34	12.2
	50+	23	13.4
Education*	High School Only	12	9.3
	Some College	35	12.7
	College Grad or Higher	43	13.4
Total		90	13.4

*$F(2, 87) = 6.497$, $p < 0.01$

Education was categorized as: a) high school only, b) some college, or c) college graduate or more. Table 2 shows the absolute difference between education subgroups. The only significant difference exists between high school only and college graduates. When respondents had a college degree, on average, nearly six more problems per interview occurred than for respondents with only a high school education.

Further analyses were conducted to see if problem type (i.e., semantic versus response task) was related to education. For example, did respondents with less education have more difficulty with formulating responses or with comprehension than respondents with more education? However, there was not a significant correlation (Pearson's r = 0.04) between education level and problem type. This indicates that all respondents had a relatively even mix of semantic and response task problems.

Table 2. Absolute values of the difference in mean numbers of problem across education subgroups.

Group Comparison	Difference in means (absolute value)	P-value
High School Only vs. Some College	3.41	.165
Some College vs.College Graduate	2.55	.114
High School Only vs. College Graduate*	5.96	.004*

*$F(2, 87) = 6.497$, $p < 0.05$

The questionnaire contained 34 behavior questions and 26 attitude questions. We found there was no significant correlation between education level and problems with different question types (Pearson's r = 0.01). Respondents had a relatively even mix of problems between the behavior and attitude questions.Overall, the relationship between education level and frequency of problems identified in cognitive interviewing was found to be significant ($F(2, 1203) = 6.504$, $p < 0.01$). Problem frequency was calculated as the number of interviews in which a problem occurred divided by the total number of interviews. Problem frequency is of interest because the less frequent a problem is, the more interviews that are required to identify it. The frequency of a problem is unrelated to its impact in each instance on measurement error. Therefore, it is possible to have a low frequency problem that causes a high level of measurement error. The average frequency of problems found by respondents with a high school education was

0.36, compared with 0.27 for respondents with some college and 0.26 for college graduates. This indicates that higher educated people were more likely to find the low frequency problems. Figure 1 further illustrates this point by showing the distribution of infrequent, frequent, and very frequent problems across education subgroups.

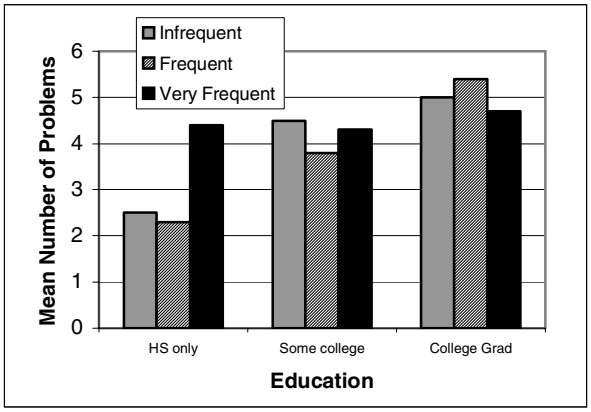

Figure 1: Mean distribution of low, medium, and high frequency problems by education subgroups.

Another interesting finding was the relationship between the respondents' education levels and the average severity scores of their interviews. Each problem had been given a severity score by the expert judges. Each interview, therefore, had a severity score which was the mean severity of the problems it yielded. The less educated respondents had an average severity score of 6.23, compared with 6.06 for respondents with some college and 5.8 for college graduates ($F(2, 1177) = 3.414$, $p<0.05$). The severity scores were grouped into tertiles (high, medium, and low) in order to examine the distribution of problems by severity within education subgroups. Figure 2 shows that less educated respondents yield a higher proportion of high severity problems than low severity problems. College graduates yield a smaller proportion of high severity problems than low severity problems. However, college graduates yield higher actual numbers, on average, of high severity problems. That is, more educated respondents identify the high severity problems and also find more problems at the lower severity levels as well.

4. Discussion

Based on these results, it seems that the most productive (i.e. leading to the most problems identified) cognitive interviewing, overall, is that conducted with respondents having above average education. No differences in interview findings were noted for sex or age. Why did education level have such an effect? One would expect that the respondents who misinterpreted the questions most often would be those with a lower reading level and fewer analytic skills--that is, those with less education. However, our findings show the opposite. Those respondents with higher educations yielded higher numbers of problems per interview. This may simply indicate that lower education respondents were less adept at recognizing problems rather than they actually experienced fewer problems. To determine this we would need to analyze the proportion of respondent -versus coder-identified problems.

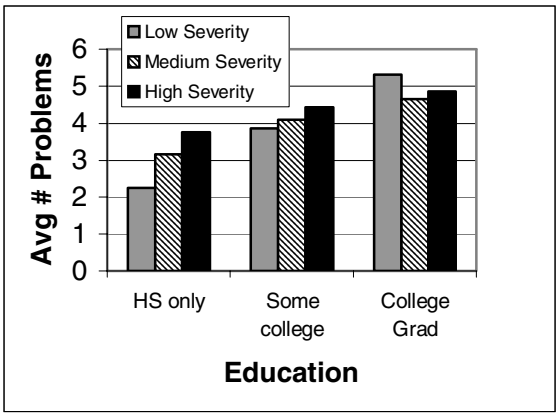

Figure 2: Average distribution of low, medium, and high severity problems by education subgroups.

Another explanation might be that the more educated respondents simply spent more time on the question-answer process. One indication of this might be interview length. If more educated respondents spent more time talking about the questions, then it is assumed that their interviews lasted longer. We found this to be true. The correlation between interview length and education was significantly positive[2] (Pearson's r

[2]To ensure that interview length was not due to particular interviewers, we also looked at the correlation between interviewer and interview

= 0.23, p<0.05). The average interview length for respondents with high school-only educations was 33.32 minutes, compared with 38.81 minutes for respondents with some college and 41.64 minutes on average for respondents with college degrees (p=0.09). These findings imply that the more educated respondents spend more time either thinking about or discussing each question. Finally, more educated respondents may identify potential problems that they do not actually experience. For example, a respondent may point out that a word used in a question is unnecessarily complex or that a question's syntax is difficult to follow, even though that respondent was able to overcome the question flaws. These would still count as identified problems. If true, this would lead to the expectation that interviews with higher educated respondents will have a higher rate of respondent-identified to coder-identified problems.

5. Implications for practice

We want to stress that this is very preliminary research. Still, it suggests the potential usefulness of considering at least one demographic characteristic when recruiting for cognitive interviews. For general population surveys that have a large and varied target population and limited resources to test a questionnaire, it is important to be as efficient as possible. One way efficiency is measured is by interview productivity, or the number of problems found in a cognitive interview. However, problem counts are not the only important efficiency measure. We also examined the types of problems identified by respondents from differing educational subgroups. For example, we found that highly educated respondents yielded a higher count of low frequency problems than less educated respondents. "Low frequency" problems may be described as "subtle" or "hard to find" problems. If a researcher's goal is to identify as many problems as possible with as few interviews as possible, then it appears that using more highly educated respondents may increase the efficiency of this process.

It is also important to identify problems judged to have a high impact on measurement error. In our study we found that highly educated respondents had higher counts of high severity problems than less educated respondents. If a researcher's goal is to identify as many high severity problems as possible with as few interviews as possible, then it appears that oversampling more highly educated respondents may increase pretest efficiency.

The purpose of this research was to better understand which respondents are the most productive in regards to cognitive interviewing. The results suggest that well-educated respondents may be more efficient for detecting problems of all types. Additionally, one might say that if highly educated respondents are having problems with a questionnaire, one can safely assume that less educated respondents will also have at least some of those problems. The converse of this idea is not a safe assumption, though. It may be that more highly educated respondents can better cope with difficult questions, for example those with complex syntax, which will cause difficulties for other, less-educated respondents. If this is the case, then limiting the sample of respondents to only highly educated respondent may result in important problems being overlooked. While respondents with more education identify more problems overall, it is not to say that they will not miss some problems less educated respondents will experience.

In interpreting these results, it is important to keep in mind the cognitive interview protocol and the protocol's effect on results. The cognitive interview protocol used for this study contained scripted probes as well as generic probes and think alouds. Problems were then identified in three ways: 1) the respondent had a problem, was aware of it and verbalized it; 2) the respondent had a problem, was not aware of it, but verbalized it; 3) the respondent did not have a problem, but was aware that a potential problem exists and verbalized it[3]. An implication of our findings may be that less educated respondents need more probing while more highly educated respondents do well with just think alouds. One way to examine this theory would be to code whether or not identified problems were preceded by scripted

length, but this was not significant.

[3] It is also possible that a respondent could: 4) have a problem, but neither verbalize it nor be aware of it. This, however, would not lead to an identified problem.

probes or think alouds. One could also recode the interviews in order to distinguish between problems that the respondent discovered him or herself, and those unnoticed by respondents, but that the researcher discovered based on verbal reports. We did not have resources to code the interactions in this way, but recommend this level of analysis. Without further analysis we were not able to address the issue of to what extent more educated respondents engaged in different cognitive interviewing behaviors that produced more problem identification.

In conclusion, most general population surveys are administered to all respondent types and one cannot be sure that age, sex and education, or any other demographic variables, will not be factors in the response process. While the questionnaire used in this study covered a wide variety of topics and question types, there still might be other topics and question types that would be affected differently. In the end, recruiting a variety of respondents is useful and beneficial to the pretest process if the goal is to exhaust all possible respondent effects. However, oversampling better-educated respondents may be particularly cost effective. Finally, it is important to note this research was not designed primarily to examine the effect of respondent characteristics on problem identification. Further respondent-factor research, with larger sample sizes and wider ranges of respondent characteristics is needed.

References

Beatty, Paul, "The Dynamics of Cognitive Interviewing," Chapter 3 in Presser et al (eds.), Methods for Testing and Evaluating Survey Questionnaires, New Jersey: Wiley and Sons, 2004.

Blair, J., Conrad, F., Ackermann, A., and Claxton, G. "The Effect of Sample Size on Cognitive Interview Findings" in *Proceedings of the American Statistical Association,* 2006.

Conrad, F. and Blair, J., "Data Quality in Cognitive Interviews: The Case of Verbal Reports," Chapter 4 in Presser et al (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, New Jersey: Wiley and Sons, 2004.

Conrad, F. and Blair, J., "Interpreting Verbal Reports in Cognitive Interviews: Probes Matter," in *Proceedings of the American Statistical Association*, Section on Survey Methods Research. Alexandria, VA: American Statistical Association, 2001.

DeMaio, T. and Landreth, A., "Do Different Cognitive Interview Techniques Produce Different Results?" Chapter 5 in Presser et al (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, New Jersey: Wiley and Sons, 2004.

Ericsson, K.A. and Simon, H., *Protocol Analysis, Verbal Reports as Data,* Cambridge: MIT, 1984.

Garas, N., Blair, J., and Conrad, F., "Inside the Black Box: Analysis of Interviewer-Respondent Interactions in Cognitive Interviews." Presented at the Federal Committee on Statistical Methodology Research Conference, 2003.

Presser, S., Couper, M., Lessler, J., Martin, E., Martin, J., Rothgeb, J., and Singer, E., "Methods for Testing and Evaluating Survey Questions," Chapter 1 in Presser et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, New Jersey: Wiley and Sons, 2004.

Presser, S. and Blair, J., "Survey Pretesting: Do Different Methods Produce different Results?" Chapter 2 in P.V. Marsden (ed.), *Sociological Methodology, 24,* Beverly Hills, CA: Sage, 1994.

Redline, C., Smiley, R., Lee, M., and DeMaio, T., "Beyond Concurrent Interviews: An Evaluation of Cognitive Interviewing Techniques for Self-Administered Questionnaires," in *Proceedings of the American Statistical Association,* 1998.

Willis, Gordon B. *Cognitive Interviewing*. Thousand Oaks, California: Sage, 2005.

Willis, G. B., DeMaio, T.J., Harris-Kojetin, B., "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, 1999.