Public Opinion Quarterly, Vol. 75, No. 4, Winter 2011, pp. 636–658

SAMPLE SIZE FOR COGNITIVE INTERVIEW PRETESTING

JOHNNY BLAIR*
FREDERICK G. CONRAD

Abstract Every cognitive interview pretest designer must decide how many interviews need to be conducted. With little theory or empirical research to guide the choice of sample size, practitioners generally rely on the examples of other studies and their own experience or preferences. We investigated pretest sample size both theoretically and empirically. Using a model of the relationship of sample size to question problem prevalence, detection power of the cognitive interview technique, and probability of observing a problem, we computed the sample size necessary, under varying conditions, to detect problems. Under a range of plausible values for the model parameters, we found that additional problems continued to be detected as sample size increased. We also report on an empirical study that simulated the number of problems detected at different sample sizes. Multiple outcome measures showed a strong positive relationship between sample size and problem detection; serious problems that were not detected in small samples were consistently observed in larger samples. We discuss the implications of these findings for practice and for additional research.

JOHNNY BLAIR is a Principal Scientist and Director of the Cognitive Testing Laboratory at Abt Associates Inc., Bethesda, MD, USA. FREDERICK G. CONRAD is Research Professor at the Institute for Social Research, University of Michigan, Ann Arbor, MI, USA. We wish to thank a number of people who were essential to implementing this study: Allison Ackermann, Paul Beatty, Laura Burns, Rachel Casper, Greg Claxton, and Gordon Willis. We especially thank Edward Blair and Clyde Tucker who met with us and provided important insights on the development of the mathematical model, the replicate study design, and data analysis. Financial and institutional support were provided by Abt Associates Inc., the Bureau of Labor Statistics, the Michigan Center for Excellence in Health Statistics [UR6/CCU517481, James Lepkowski, Principal Investigator], and the Survey Research Center—the latter two at the University of Michigan, where the cognitive interviews were conducted. The opinions expressed in this article are solely those of the authors and do not reflect the positions of any of the supporting organizations. *Address correspondence to Johnny Blair, Abt Associates Inc., 4550 Montgomery Ave., Bethesda, MD 20814, USA; e-mail: Johnny_Blair@abtassoc.com.

doi: 10.1093/poq/nfr035

Introduction

The survey literature on cognitive interview pretesting typically considers the choice of sample size (number of interviews) only in passing, most often simply to note how many interviews are commonly conducted. This lack of attention has persisted even though, after more than two decades of cognitive interviewing, practitioners have little theoretical or empirical guidance to determine pretest sample size. That has left decisions about sample size to considerations of resource constraints, researcher's judgment, and the face validity of prior pretests that employed particular sample sizes—approaches that, while relevant, have not produced cumulative, generalizable knowledge. These practices have tended to result more often in smaller rather than larger samples, a tendency recognized by Willis (2005, p. 7), who lists "modest sample sizes" as one of the defining features of the cognitive interview pretest.[1]

Small samples do identify real problems with draft survey questions. At least three studies have found that question problems identified in small cognitive interview pretests will, if left uncorrected, occur in subsequent survey interviews (Willis and Schechter 1997; Fowler 2004; Blair et al. 2007). The face validity of numerous case studies supports the view that small cognitive interview pretests contribute to improved questionnaires. These kinds of results cannot speak to the possibility of additional problems those pretests may *not* have detected.

One of the few explicit discussions of sample size largely supports small samples. While granting that "to an extent, the more interviews we can do, the better," Willis (2005, pp. 226–28) holds that small cognitive interview samples are often sufficient for the purposes at hand. He notes, first, that the purpose of cognitive interviews is not statistical estimation: ". . .we do not evaluate survey questions simply by counting the number of interviews in which a problem occurs" (p. 227). Second, he argues that laboratory interviews are qualitative. The methods of problem identification are such that "a finding [might] be based on one interview" (p. 227). Third, judicious selection of laboratory subjects can compensate for their small numbers.

While each of these points is well taken, each one can also, in a particular application, suggest the need for a larger rather than smaller sample. First, although a cognitive interview does not evaluate a question by counting how often a particular problem occurs (as does, for example, behavior coding), problem prevalence is relevant to pretest sample size. Prevalence is an indication of both the potential effect on the survey of the flaw going undetected and the pretest sample size needed to find it.

1. In an early survey of cognitive interviewing practices at academic and government research organizations, Blair and Presser (1993) found that samples at that time were typically small—nearly always under 30 in academia, although larger (yet rarely exceeding 60) in federal agencies. Researchers guided by past practices will find that, historically, pretests have used small samples.

Second, although sometimes a single interview may prove sufficient to identify a problem, one cannot know whether that definitive case will occur on the fifth interview or the fiftieth. Some important problems may be neither ubiquitous nor easily uncovered.

Third, although sometimes one can purposively include respondents with some particular characteristic to detect problems in questions meant for them, this criterion begs the question of how many such respondents are sufficient.

These counterpoints are not to argue for larger samples *per se*, but simply to note that appropriate sample size cannot be settled on the basis of these kinds of considerations alone.

One question necessary to each cognitive interview pretest design is: How large a sample is needed to find all the problems *that warrant question revision*? Factors such as coverage of questionnaire paths or respondent subgroups that need to be covered can help inform the sample size decision. But we need theory and empirical studies to connect sample size directly to question problem identification. Without such a connection, the choice of adequate sample size is likely to be insufficiently informed.

Given that there has been little reported use of large samples, we don't know if more interviews will generally continue to find new problems, possibly including some serious ones. There is also very little research on whether a different sample of the same size would produce the same findings. Whether judged by comparing independent samples (Presser and Blair 1994), by independent coding of the same sample (Conrad and Blair 2004), or by independent teams testing the same questionnaire (DeMaio and Landreth 2004), reliability is not especially high. Some factors affecting reliability are likely to depend on features of the particular cognitive interview protocol, but sample size may also be an influence. If, for a fixed sample size, different results are found in different samples, one would expect that for larger samples, there would be less variation in problem identification. As Krosnick (1999) points out in a comparative review of pretest methods, "Cognitive interviews. . .tend to exhibit low reliability across trials. . .[which] might reflect the capacity of. . .[the]. . .method to continue to reveal additional, equally valid problems across pretesting iterations, a point that future research must address" (p. 542). This clearly implies that reliability may increase with additional interviews.

Sample Size in Usability Testing: Implications for Survey Pretesting

As a problem discovery enterprise, usability testing has some informative similarities with cognitive interview pretesting. Specifically, usability test sample sizes are typically small, yet are given considerable weight. There have been recent reconsiderations of sample size determination in the usability literature. Given the paucity of relevant survey research literature, we reviewed

a set of recent usability testing articles on sample size (see the online appendix, section A).

Our review found both theoretical and empirical research on sample size for usability testing. The theoretical studies have not been definitive, due partly to disagreement about appropriate models and values for model parameters. Both usability testing and cognitive interviewing have produced many examples of testing achieving "useful" results with few cases. Similarly, in both fields, research studies that used small sample sizes have usually found low reliability of each method (e.g., Presser and Blair 1994; Conrad and Blair 2004; DeMaio and Landreth 2004 regarding cognitive interviewing; and a usability literature review by Turner, Lewis, and Nielsen 2006). The studies in both fields have typically, to our knowledge, been based on a single, fixed sample size. An exception in usability testing is Faulkner (2003), who ran simulations based on a pool of 60 subjects. Her findings indicate that small samples are sometimes effective, though often not. More importantly, even modest increases in sample size—e.g., from 10 to 15—produced substantial improvement in problem detection and reliability.

The usability research findings can usefully inform cognitive interview research. However, we must also be cognizant of differences between the fields that make direct application of the usability findings inadvisable. The problems with many computer interfaces may be relatively universal; i.e., they are likely to occur across different types of users. In contrast, a given survey instrument problem may be experienced only by respondents with particular characteristics such as demographics, education, experiences, or behaviors that are not known *a priori*. Additionally, in the case of computerized survey instruments such as Web-based questionnaires, a thorough pretest would address both usability issues (such as difficulty entering answers, changing answers, or navigating through the instrument) and problems of survey response.

Modeling Cognitive Interview Pretest Sample Size

The priority of most pretests is to identify problems that, left uncorrected, would most increase measurement error. Each problem's impact on measurement error derives from (a) its "prevalence," the percentage of interviews in which the problem occurs; and (b) its "severity," the problem's effect, each time it occurs, on the discrepancy between the measured and true values.[2] These two factors are conceptually different. On the one hand, prevalence is simply the probability that the problem will occur each time the question is asked. On the other hand, severity is the effect of the problem on each measurement. Severity depends primarily on the nature of the problem, but may also be affected by other factors such as

2. A problem may have consequences other than response inaccuracy, such as requiring respondents to exert more effort to answer accurately. In practice, one would also want to eliminate that kind of flaw, but the effect on measurement is our present concern.

respondent characteristics or the type of measurement. For example, assume that the author of the question "Do you own a car?" intended the word "car" to include small trucks and SUVs. If that is not how some respondents interpret the word "car," the subset of those respondents who own only a truck or an SUV will answer incorrectly. The effect on answers is high when (a) this misunderstanding occurs; and (b) the respondent owns only a truck or an SUV. How frequently a problem occurs and the problem's effect on each answer are different factors with different implications for measurement error.[3]

Both the question response format and the statistic to be estimated can affect the severity of a problem. Consider the question "How many books have you purchased this year?" Assume, for example, that some respondents understand the question to mean only books bought for themselves when, in fact, the item intent is to include books bought as gifts. The form of response could be open, in order to estimate average book purchases; or the response format could require choosing a category to represent the number of purchases, to estimate the proportion of people in each category.

In the open response case, problem severity (magnitude of the underreporting) depends on how many respondents misunderstand the question and the number of their omitted gift purchases. Using the wrong number of purchases will always contribute error to the estimated mean.

For the categorical response, the intent is to estimate the proportion of respondents in each category. A respondent can be wrong about the number of purchases, but that number may still correspond to the correct category.[4] The problem severity (number of miscategorizations) depends on how many respondents misunderstand the question *and* how often this leads to selecting the wrong category.

In our view, to gauge the impact of a problem and therefore the importance of observing it with a given number of pretest interviews, one should consider both its prevalence and its severity. From the perspective of sample size, a problem's prevalence affects the number of pretest interviews needed to identify it. For example, if we conduct a specified number of cognitive interviews (n) and a particular problem (f) occurs with prevalence (π), what is the probability (P_f) that it will be observed at least once by the n^{th} *pretest* interview, i.e., at some point in a sample of size n? The probability of observing a problem in the pretest sample depends on two factors: how often the problem occurs (π)

3. The type of measure affects the severity. If the question asked "How many cars do you own?", then respondents who misunderstand the question and happen to own both a small truck and a conventional car will answer 1 rather than 2, the correct response. But the effect on the survey estimate will be different than in the first example, both because the statistic is a mean rather than a proportion and because the proportion of the sample that can potentially make the response error is different.

4. The same issue can arise at the analysis stage. If the researcher collapses respondents' open numerical answers into numeric ranges, the misconception would have to push a response into the wrong range.

and how likely it is to be detected when it does occur (d).[5] We distinguish the factors d and π, in part, because they are theoretically separable, and that separation clarifies how the magnitude of measurement error is affected. A problem with a survey question (e.g., a key word the respondent does not sufficiently understand, or an ambiguity in what the respondent is being asked to do) can occur and affect answers, yet not be detected by the respondent, the interviewer, or an independent observer.

A theoretical model of survey pretest sample size needs to allow for problems that occur in some interviews but are not always identified. From the practical perspective of pretest design, it is useful to consider these factors separately because while π is primarily a feature of the problem itself, d is at least partly determined by decisions within the practitioner's control. In cognitive pretesting, each time a question is administered, an observation is made by an interviewer, a coder, or someone else as to whether or not a particular problem, f, occurred. Assuming that the problem f did in fact occur, its probability of detection, d, depends on the features of the problem and the detection mechanism.

Features of the problem can affect d if they affect the amount of overt evidence respondents report as a result of the problem. For example:

- Whether or not the problem prevents the respondent from providing an answer; e.g., *not understanding* is likely to place the respondent at an impasse, thus producing more evidence, rather than *misunderstanding*, in which the respondent answers the question believing he or she has understood it as intended.
- The degree to which the problem involves processes about which respondents can report verbally; e.g., known autobiographical facts come to mind, bypassing working memory, and leave little to report when thinking aloud; similarly, a previously stored judgment such as "I do not favor mandatory sentencing" is not accessible to introspection (see, for example, Ericsson and Simon 1993) and leaves little for respondents to verbally report beyond statements such as "I just know that."

Features of the detection mechanism that can affect d include:

- The effectiveness of the protocol design in identifying a problem f—such as the amount and type of probing and investigation interviewers are licensed to conduct.

5. Our notion of detection concerns "hits," i.e., the identification of actual problems when they occur, as opposed to "false alarms," that is, the "detection" of spurious problems (see Conrad and Blair 2009). "Misses" (failing to detect true problems when they occur) are accounted for by the prevalence indicator. While *correct rejections*—no problem detection when no actual problem has occurred—are important for pretesting, they are not at issue for assessing the effect of sample size. However, a larger number of interviews would give one more faith that the decision to reject a problem is correct.

- The process of determining when verbal report evidence is sufficient to conclude that a problem has occurred.[6]

The disaggregation of the factors that affect problem detection shows how problems with the same prevalence may vary in the likelihood of being observed in a pretest. Acknowledging that there is variability in "problem discoverability" implies that there may not be a single, fixed value for the factors d and π.

If the probability that a problem occurs in any given interview is π, then the probability of the problem not occurring by n interviews is conditional on its having not occurred in any interview prior to n , that is, $(1 - \pi)^n$. If the problem f is detected d_f percent of the times it occurs, the probability P_f of observing it at least once in a pretest of size n is

$$P_f = 1 - \left(1 - d_f.\pi_f\right)^n. \qquad \text{Eq. 1}$$

This expression shows that the probability of observing a problem in a given sample depends on its prevalence, the effectiveness of the detection method, and the sample size. There are different ways that one might make use of the relationship between these parameters. For example, practitioners designing a pretest may want a specific level of confidence that problems at or above a particular prevalence will be observed in the sample; i.e., they may require a particular value for P_f. This might vary for different problems, but one would assume it is high and, if cognitive interviewing is the sole pretest method, may be close to 1 for severe problems or the value of the pretest would be seriously reduced. Given a required value for P_f and an assumption about the efficiency of the detection procedure, what sample size would be required?

To address this, we examine the effect of these factors on sample size by solving the above expression for n,

$$n = \frac{\log(1 - P_f)}{\log\left(1 - d_f{*}\pi_f\right)}. \qquad \text{Eq. 2}$$

Although prevalence and detection are theoretically distinct, they cannot be separately measured. Thus, we cannot set these values empirically. However, we can illustrate the effects of these parameters on sample size using a set of parameter values (figure 1) that we consider to be both plausible and of practical importance. Each panel in the figure corresponds to a different hypothetical problem that occurs with a particular prevalence π: .05, .07, .08, and .10. We chose this range of prevalence values because, in our judgment, most researchers would be concerned about a problem that had a severe effect on answers even if the problem occurred in only 5 percent of interviews, and would be similarly concerned if a problem with only a moderate impact on answers occurred in 10 percent of interviews. For each of the four hypothetical

6. Other factors that influence detection can include the quality of the interviewers' and coders' training, and the amount of the interviewers' and coders' experience.

Figure 1. Effect of Parameter Values on Sample Size.

problems, i.e., each prevalence level, we consider four cumulative probabilities of observing the problem P_f: .5, .8, .9, and .95. We selected these values to span the range from a barely acceptable (50 percent) probability of observing a problem in a pretest to the ideal near certainty of observation (95 percent).

Recall that the probability of observing a problem in a given interview is a function of how frequently it occurs (prevalence) and the probability of detecting it when it does. Thus, we calculate the effect of $(\mathbf{d}^*\pi)$ on sample size for each level of P. The range of values for d contributing to each curve runs from .4 to .9. While d is simply a feature of the detection mechanism, rather than a selected value, we chose to use values that span a range with a lower bound at what we think some practitioners would consider a marginally adequate probability of detecting a problem (e.g., if cognitive interviewing was not the only pretest method being used) when it occurs (.4). Most practitioners would accept that perfect detection is probably not attainable, so we have chosen a value (.9) as close to perfect detection as one might reasonably assume.

The sample sizes necessary to observe a problem, given these parameter values, are shown in figure 1. Each panel shows a different problem prevalence. The products of detection and prevalence are displayed along the x-axis. The probability of detection increases from left to right, beginning with the smallest specified value, .4, and increasing to perfect detection (marked by the vertical line that intersects the x-axis) when $d = 1$ and $(\mathbf{d}^*\pi) = \pi$. The required sample sizes, shown on the y-axis, decrease as the probability of detection becomes larger.

Each curve shows the range of sample sizes for a specified probability (P) of observation (a hit); the sample sizes for highest and lowest detection probabilities are circled on the end of each curve.

For example, consider the problem with 5-percent prevalence (figure 1) and perfect detection, i.e., $d = 1$. Perfect detection means that, for the 5 percent of interviews in which the problem occurs, it would always be detected by the cognitive interviews. In that case, with a sample of 14 interviews, there would be a 50-percent chance ($P = .5$) of observing that problem. With 31 interviews, the probability of observing it increases to .8; it rises to .9 with 48 interviews and then to .95 with 58 interviews—again, assuming that whenever the problem occurs it is never missed.

However, it would be unrealistic for a practitioner to expect perfect detection. With any method, a problem may occur and, for various reasons, go unnoticed. With imperfect detection, the sample sizes necessary to observe the problem, with probability P, increase as shown in the four panels, with one curve for each value of P (figure 1).

The relationships in figure 1 demonstrate that if one wants a high probability (e.g., $P = .9$ or $P = .95$) of observing a relatively rare problem ($\pi = \mathbf{.05}$) in a cognitive interview pretest, the required sample sizes are going to be relatively large (from 50 to well over 100), whether detection efficiency is poor (.4, in which case $n = 148$) or very good (.9, in which case $n = 65$). Even if one is willing to accept a more modest probability of observation (e.g., $P = .8$), the required samples range from $n = 35$ to $n = 80$. It is only at the coin-toss probability of observation (.5) that

sample sizes drop to what is probably at the upper end of samples typically used in current practice, $n = 15$ to 35 per round (Willis 2005, p. 7).

Consider the problem in figure 1d that is twice as prevalent ($\pi = .\mathbf{10}$) as the one in figure 1a. For a moderate ($P = .8$) probability of observing the problem, the sample sizes go from 17 to 39 as d ranges from .9 to .4. Only at a level where one is just as likely to miss the problem as to identify it ($P = .5$) do the samples all fall below $n = 20$.

In most practical situations, we cannot know the values of all the parameters affecting sample size. However, these calculations demonstrate that, for a wide range of assumptions about those parameters, the likelihood of observing a problem with 10-percent or lower prevalence requires substantially more interviews than are typical in current practice.

From another perspective, figure 1 shows that—again, for a fairly wide range of assumptions—small samples (under 25 or so) will suffice only if one is willing to settle for about a 50-50 likelihood ($P = .5$) of observing relatively frequent problems ($\pi \leq .10$). If a problem has a higher prevalence than in these illustrations, of course, the situation changes. For example, if a problem has 20-percent prevalence, the likelihood of observing it is fairly high with sample sizes of around 30, even with a poorly performing detection mechanism ($d = .4$ or .5). Samples of 20 or fewer would suffice to find problems with a prevalence of one in three interviews, a value that strikes us as rather high.

The sample size required to achieve a particular probability of observation (P) increases rapidly as the efficiency of the detection mechanism decreases, especially if the problem is relatively rare. This means that a strong detection method is particularly important to keep the sample size down if the problem does not occur very frequently.[7]

7. This examination shows that computing sample size for cognitive interview pretests would be a relatively straightforward matter if one knew the relevant parameter values. There are, however, some caveats about the computations specifying sample sizes just presented. First, the mathematical relationship of prevalence to probability of observation, for a given sample size, assumes both a simple random sample of respondents and independence of the observations. If the observations are to some extent correlated (as could happen, should interviewer or coder expectations—perhaps colored by their personal judgments of which questions will perform poorly—affect either their behaviors or judgments over the course of the set of interviews), then that correlation will reduce the incremental value of each observation and hence require a larger sample size to achieve the same cumulative probability of identification.

Second, samples for cognitive interview pretests may be selected randomly. But they may also be selected by some nonrandom method in order to ensure representation of population subgroups expected to have more problems than others (e.g., those with lower education), or who may be better able to articulate their thoughts or do other cognitive interview tasks. For example, Ackermann and Blair (2006) found a greater yield of useful verbal reports from more educated respondents. Thus, for higher-educated respondents, d seems likely to be higher, lowering the necessary sample size. Nonrandom samples may also be employed to ensure coverage of all the paths through the instrument. Again, the exact consequences of manipulating sample composition are of interest here only to note that they may affect the accuracy of the mathematical relationship.

The nature of a problem may sometimes affect the ease of detecting it. For example, problems that the respondent becomes aware of will generally have more possibilities of being detected than those of which the respondent is unaware, if only because the respondent's explicit descriptions of problems are possible. Similarly, prevalence may be related to ease of detection, if only because as prevalence increases, there are more opportunities for identification. Even a weak detection technique may be adequate for uncovering pervasive problems. Problems that are both prevalent and severe should be detected early and easily. But the more general point is that particular protocols may be more effective for uncovering some types of question flaws than others. For example, think-aloud methods may be better at uncovering recall difficulties, while question-specific probes may more easily find out when words are not understood as the question intends.

Given the range of possible detection mechanisms, the variations in questionnaires and populations, along with the inherent probabilistic nature of the relevant variables, an empirical investigation in which we control as many factors as possible allows us to more concretely test the relationship of sample size to problem identification. We designed a study that allows for the simulation of alternative sample sizes and examination of the number and nature of observed problems.

Empirical Study

The study was conducted to compare problem identification across different sample sizes. We created a pool of cognitive interviews that we used to simulate a range of sample sizes. Cognitive interviewers used a questionnaire we constructed—containing embedded question flaws unknown to the interviewers—to conduct the interviews that made up the pool. Each cognitive interview was coded for problem identification. Independent random samples of different sizes were selected from the pool. For each set of samples of a given size, we computed several measures of problem identification.

QUESTIONNAIRE CONSTRUCTION

We constructed a questionnaire to represent a range of question types and response tasks. Sixty pretested questions were selected from major government, academic, and commercial surveys.[8] Thirty-four of the questions were behavioral, and twenty-six were attitudinal. In order to be sure the questionnaire contained a sufficient number of problems and that these problems were known to us in advance, each

8. 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 and Health; items on the economy were taken from the University of Michigan's Survey of Consumers; and finally, items on a variety of public opinion topics were taken from Harris, Gallup, Pew, *New York Times*, and CBS polls. The questionnaire and the embedded problems are given in the supplementary online appendix, section C.

question was "damaged"; i.e., its wording was modified so that the question was expected to cause at least one problem for at least some respondents. The types of problems and their expected severity were based on the authors' judgments.

PROBLEM RATING

Problems can vary in their impact on survey measurement error, from relatively minor to extremely serious. A study of problem detection is more informative if it includes some type of evaluation of each problem's potential impact. The analysis can then be used to assess a method's capacity to detect problems with higher than lower impact. We used experts' judgments as an indicator of each problem's potential impact.

Three questionnaire design experts independently rated each experimenter-embedded problem on two dimensions: first, in what percentage of interviews they thought each problem would occur in an actual survey; and second, when the problem occurred how severe the effect on the measurement would be, where "severe" was defined as the degree, rated on a scale of 1 to 10, to which they thought the problem would distort the answer's accuracy. The correlations between each expert's severity and frequency ratings were high, with a mean correlation of .76 ($p < .01$).

We multiplied each question's severity and frequency rating to create a *problem impact* score.[9] The three experts' impact scores were averaged to produce a single impact score for each problem.[10] The experts also identified some additional question problems, which they rated in the same way as they had rated the embedded problems. Finally, additional new problems beyond these two sources were detected in the cognitive interviews and were also rated. The mean of the pairwise correlations between the experts' impact scores was only .23 (p significant at either .05 or .01).[11] Only those problems that

9. Combining the severity and frequency scales by multiplying them causes some loss of information because the same impact score (product) can result from many different combinations of severity and frequency values. For example, a 5 on frequency and a 1 on severity produce the same impact score as would a 1 on frequency and a 5 on severity. These identical impact scores do not provide a sense of which dimension contributed most, or whether both were about equal. There may be surveys for which the designer is more concerned about one or the other dimension. However, for our study, we needed only to distinguish problems with greater impact from those with less. The origin of the impact is of less interest.

10. This produces a conservative assessment of impact because a high impact score requires relatively higher agreement between experts.

11. There are few studies that assess the validity of experts' judgments of question problems. However, a recent study of questionnaire evaluation by expert reviewers found that "despite the lack of reliability [across six experts], the average expert ratings successfully identi[fied] questions that had. . .higher levels of inaccurate reporting" (Olson 2010, p. 295). Although that study's focus was on questions, rather than individual problems, the finding of low between-expert agreement is consistent with the low correlation between the experts' judgments of impact in the present study. Olson's finding that average expert ratings of questions were good indicators of reporting error suggests that mean problem impact scores should be valid indicators of measurement error.

were actually detected in the cognitive interviews were used in the analysis. All of the embedded problems actually occurred in the interviews.

SAMPLING AND RESPONDENTS

A general population sample of 90 respondents was recruited from a commercial e-mail list. We purchased 20,000 e-mail addresses from Genesys that were linked to physical addresses from the Ann Arbor, MI, area. We purposely avoided recruiting participants only from the University of Michigan community to obtain a greater diversity of respondents. Sample members were recruited via an e-mail invitation, sent out in four weekly waves of 5,000 each. Potential respondents provided their age, sex, and education level in response to e-mail screener questions, allowing us to control the distribution of respondents based on these characteristics. Quota sampling was used to ensure variation of respondent age, sex, and education (see the online appendix, section B, for details).

The characteristics of respondents were roughly balanced across the cognitive interviewers. The interviews were conducted in November and December 2004 at the University of Michigan Survey Research Center in Ann Arbor, MI. Respondents received $35 upon completing the cognitive interview.

COGNITIVE INTERVIEWING

Ten interviewers, whom we trained for this study, each conducted nine cognitive interviews to create a pool of 90 interviews. The 10 interviewers had experience conducting standardized production interviews, but most had little prior experience with cognitive interviewing. For the current study, they were trained to use a cognitive interview technique that combined think-aloud instructions and scripted probes that the interviewers devised after studying the questionnaire. After asking a question and allowing the respondent to think aloud, the interviewer administered any probes written for that question.

Each interviewer conducted nine interviews in two iterations, the first consisting of five and the second of four interviews.[12] Prior to the first iteration, the interviewers identified potential problems and crafted scripted probes designed to determine whether those problems occurred. After the first iteration, the interviewers were instructed to review the protocol and change the scripted probes based on what they had found to that point. The revised protocols were used to conduct the second iteration. In this way, we hoped to approximate the cognitive interviewing practice in which interviewers are free to modify their procedures based on what they have learned to date. The same questionnaire was used in both iterations.

12. The level of problem identification did not differ significantly between iterations. Nor did position in the questionnaire (early vs. late items) affect the relationship between sample size and the number of problems identified. These detailed analyses are provided in the online appendix, section B.

PROBLEM CODING

After data collection, for each time a question was asked, the verbal interaction between the interviewer and respondents was coded for problem occurrence by two coders working in collaboration.[13] They used the coding frame from Presser and Blair (1994), in which a problem is either (a) a difference between question intent and question interpretation (called "semantic" by Presser and Blair); or (b) some other aspect of the response task judged likely to result in response error ("task").[14]

Simulating different sample sizes: Treating the pool as a universe of interviews, we simulated different sample sizes by selecting different sets of interviews from this pool. For each sample of a given size n, repeated random samples (replicates) were selected, with replacement, from the pool. We did this for samples whose size was increased by intervals of five interviews: $n = 5, 10, 15, \ldots, 85, 90$ interviews, which created 17 sample sizes. For each sample size, we selected 90 sample replicates by simple random sampling with replacement.

Averaging across replicates produces a more stable estimate of problem identification (for a given sample size) than would a single sample. The number and nature of identified problems in each of the 90 replicates was determined, allowing estimates of mean and total number of problems for each sample size and comparisons across sample sizes.

Results

MEAN NUMBER OF IDENTIFIED PROBLEMS

The first research question concerns the mean number of unique problems identified at each sample size. Across the 90 interviews, a total of 210 unique problems were identified. For each sample size, the mean number of unique problems per interview (irrespective of how many instances of each problem were observed) was calculated as the total number of unique problems in all 90 replicate samples divided by 90, along with the standard deviation for each mean.[15] The mean for each sample size is displayed in figure 2.

Although a tenet of current practice is that small cognitive interview pretests are sufficient to identify most questionnaire problems, these data do not support that position. At sample size 5, the mean number of unique problems identified is

13. The coders collaborated rather than working independently, so that we did not have to consider inter-coder reliability as an error source, one that was not relevant to the study objectives.
14. See the online appendix, section B, for further details of study procedures and the coding frame.
15. The mean number of problems per interview was about 13. Mean interview length was 33.7 minutes.

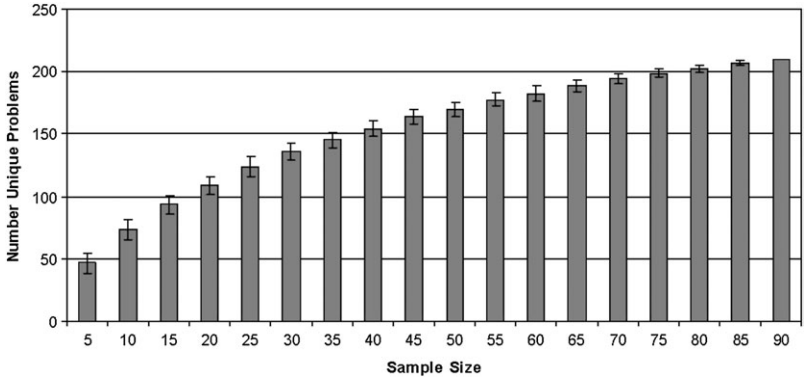

Figure 2. Mean Number of Problems per Sample Size.

46.69, less than a quarter of the 210 problems known to be in the instrument. While modest sample size increases do produce substantial gains, this, alone, can be misleading. When those gains are considered in the context of findings from larger sample sizes, we see that they account for only a fraction of all problems in the corpus of 90 interviews. The number of identified problems doubles from sample size 5 (mean = 46.7) to sample size 15 (mean = 93.5). Considering that the questionnaire consists of 60 items, the identification of about 47 to 94 problems appears quite productive. However, additional problems continue to be found with more interviews. Only about a third of all 210 problems are found with 10 interviews, half are found at $n = 20$, and a sample of 50 is necessary to reach the 80-percent-coverage mark.

The most substantial gains in problem coverage occur at the lower end of the sample size distribution, yet the mean number of identified problems increases approximately in proportion to the increase in sample size($R^2 = .92$). Even at samples of 70 or more, some of the problems were still not identified. In fact, the total pool of 90 interviews had to be examined before all 210 problems were identified. The most striking result illustrated in figure 2 is that when only a small number of interviews are conducted, many problems are not uncovered and become evident only with a larger number of interviews.

As expected, larger numbers of interviews produce more stable counts of the number of unique problems across individual samples than do smaller numbers of interviews; the standard deviation for average number of unique problems identified in samples of size 5 is 8.02 and decreases until samples of 85, when the standard deviation is 1.58.[16] The size of the standard deviation helps explain the

16. Of course, the greater amount of replacement (interviews that are reselected) across samples of size 85 than size 5 is partly responsible for the drop in standard deviation. At sample size 90, the entire study universe is included, so there is no standard deviation.

low reliability of small samples; a single sample of 5 can uncover a number of problems much larger or smaller than the mean (for many samples of size 5).

While the number of problems identified continues to grow as the sample size increases, the *rate* of identification per interview (as measured by mean number of problems divided by sample size) does slow down. At $n = 5$, the mean rate is 9.34 problems per interview; and at $n = 15$, the rate declines to 6.23. The rate drops to 3.40 by $n = 50$. The decreased rate of problem discovery with sample size is nearly linear ($R^2 = .81$). In other words, although many problems that are undiscovered at smaller sample sizes are identified with larger samples, the efficiency of subsequent interviews in finding new problems decreases as sample size grows larger.

IMPACT OF A PROBLEM ON MEASUREMENT ERROR

Not all problems are of equal concern; if coverage of the more serious problems is high, the practitioner may be willing to miss (or discover by other means, such as conventional field pretesting or behavior coding) some, or even a substantial number, of the low-impact flaws. The next question, then, is: If we imagine a total number of serious problems, how much of that total is identified at smaller than larger samples? How big does the sample need to be to cover most or all of the serious problems?

Although the number of unique problems increases with sample size, it is conceivable that the most serious problems are found with small numbers of interviews. We tested this possibility using our impact score, described above. To examine identification of problems with varying impact, we divided the problems into impact quartiles (first-quartile problems are lowest impact; fourth-quartile problems are highest impact). Figure 3 shows the proportion of all unique problems in each impact quartile identified at each sample size. Although a large proportion of the highest-impact (fourth-quartile) problems are uncovered at small sample sizes, additional high-impact problems continue to be uncovered as the sample sizes increase; even at sample size 85, some serious problems were found that had not been observed at samples of 80. The proportion of less serious problems in each of the other three quartiles also increases with sample size, yet about a quarter of them remain unidentified even at sample size 40.

Overall, these results show that small numbers of cognitive interviews do expose proportionally more high-impact than low- or intermediate-impact problems, but larger numbers of interviews expose substantially more problems at all four impact levels.

LIKELIHOOD OF OBSERVING A PROBLEM IN THE PRETEST

Thus far, our analyses have considered the total number of problems identified at each sample size. But what is the likelihood of any given problem being observed in *a single sample* of a particular size? Irrespective of impact, some problems may be sure to be found with a small sample, while uncovering other problems may require more interviews. To address the question of how probable it is that a particular

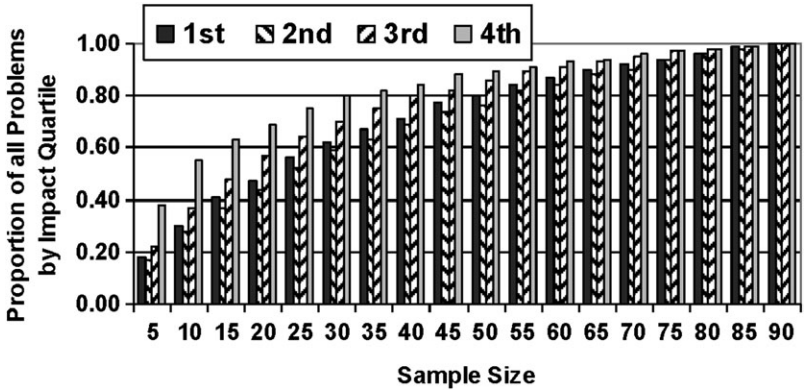

Figure 3. Proportion of Unique Problems per Impact Quartile and Sample Size.

problem will be observed in a given sample of size n, we constructed a likelihood statistic. The likelihood of a problem being discovered is simply the number of samples of size n in which the problem was identified in at least one interview, divided by 90 (the total number of replicates of that sample size). For example, if out of the 90 replicates of size 5, a particular problem was identified in 30 replicates, then the likelihood for that problem at that sample size is $30/90 = .33$.

This analysis shows, for each sample size, the number and proportion of problems that have P probability (from equation 1) of being observed.

Table 1 shows these statistics for five likelihood categories at different sample sizes.[17] The main point of the table is that most problems have a very low likelihood of discovery with a small sample size. At sample size 5, only a small percentage of problems have a 100-percent chance of being observed, while over 70 percent of all problems have a 25-percent or smaller chance of being observed. The percentage of problems in the low likelihood category decreases steadily as sample size increases, with that lowest category virtually disappearing by sample size 30. However, identification of most problems is far from certain even in larger samples. By sample size 20, less than half of the problems are observed frequently (51- to 100-percent chance of being observed at least once). Samples of about 55 are necessary before all problems have a better than 50-percent chance of being observed in a single sample.

If we focus only on high-impact problems (figure 4), we see that at $n = 25$, about one fifth of high-impact problems have less than a 50-percent chance of identification. If our main concern is to ensure that we discover nearly all of the

17. The five likelihood categories are 1–25%, 26–50%, 51–75%, 76–99%, and 100%. The sample sizes are 5, 10, 15, 20, 25, 30, 35, 40, 45, . . ., to 90.

Table 1. Likelihood for All Problems Identified at Each Sample Size

	Sample Size																	
Likelihood	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
1–25%	0.71	0.48	0.30	0.22	0.08	–	–	–	–	–	–	–	–	–	–	–	–	–
26–50%	0.17	0.30	0.35	0.34	0.39	0.32	0.30	0.24	0.15	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
51–75%	0.08	0.11	0.14	0.19	0.21	0.29	0.24	0.25	0.25	0.26	0.30	0.28	0.20	0.03	0.00	0.00	0.00	0.00
76–99%	0.03	0.09	0.16	0.19	0.24	0.28	0.32	0.35	0.40	0.45	0.44	0.41	0.44	0.58	0.52	0.44	0.30	0.00
100%	0.01	0.02	0.04	0.07	0.08	0.11	0.13	0.16	0.21	0.23	0.26	0.31	0.36	0.39	0.48	0.56	0.70	1.00

most serious problems, we see that not until $n = 75$ do all high-impact problems have a greater than 75-percent chance of being identified.

Discussion

Both the theoretical and empirical components of this research found a strong positive relationship between sample size and problem identification. The likelihood of observing a given problem in a set of cognitive interviews clearly increases as the size of the set grows. For a wide range of values for prevalence and detection, the theoretical model indicates that moderately large sample sizes are needed to uncover a high proportion of all problems.

Consistent with this result, in the empirical study, additional interviews continued to produce observations of new problems, although the *rate* of new problems per interview decreased. The benefits of increasing the sample size hold under different outcome criteria:

- The total number of problems identified (figure 2)
- The likelihood of problem discovery (table 1)
- The likelihood of high-impact problem discovery (figure 4)
- The proportion of all problems, regardless of impact level, that are identified (figure 3)

The results strongly indicate that small sample sizes may miss a substantial percentage of problems, even if concern is limited to those problems with a serious impact on measurement error. This is not to say that a small sample may not sometimes achieve pretest objectives, to the extent that they are well specified. As noted, small samples are effective when problems are very prevalent or

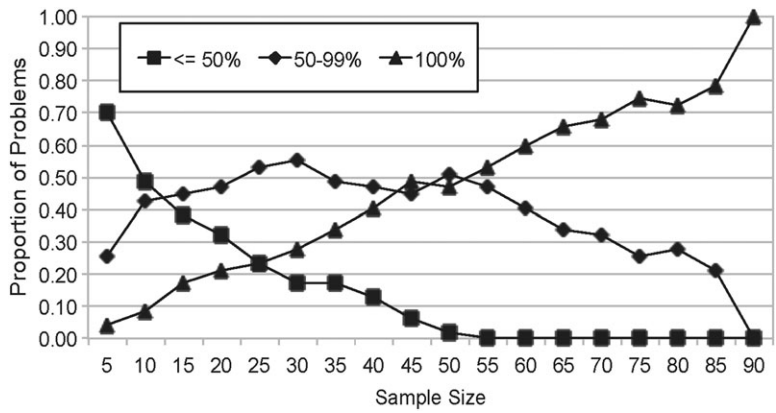

Figure 4. Proportion of High-Impact Problems in Each Likelihood Category.

the detection mechanism is very strong. However, it seems prudent to expect that (a) not all of the problems warranting revision will occur in the majority of interviews, but will *vary* in frequency; (b) a pretest will seldom be limited to identifying only the most common problems; and (c) serious problems will not always be so easy to detect as to consistently achieve values for $d > .9$. In most circumstances, if one depends on a small number of interviews, the risk of missing important problems is high—as shown in the likelihood analysis result for high-impact problems.

Before discussing implications of our findings, we should first note the decisions required by the design of this research. In the theoretical model, we chose parameters that we judged to represent a realistic range of problem frequency and detection effectiveness. But, just as we noted in reviewing the usability literature, experts' judgments may differ as to the appropriate parameter values.

The empirical study needed both to reflect cognitive interview practice and to simulate problem discovery across a wide range of sample sizes. This goal required a number of decisions about features of the design. Other researchers may have made different choices. By its nature, an empirical study cannot speak to the possible effects on its findings of a different set of design decisions.[18]

IMPLICATIONS FOR PRACTICE AND FUTURE RESEARCH

Although cognitive interviewing goals may vary from one survey to another, in our judgment, problem detection is the primary objective of most pretesting.[19] The main thrust of our findings applies whether the intent is to identify only the most serious problems or the broader goal of detecting all the problems that may warrant revising the question.[20]

Low levels of problem coverage would be of less concern if (a) practitioners had some method to estimate coverage, and/or (b) multiple pretest methods were routinely conducted. Periodically, there are calls to determine how best to combine multiple methods to optimize use of resources (see Esposito and Rothgeb 1997). If practitioners believe that pretest problem coverage is high, they will be more likely to rely on cognitive interviewing as their only pretest method. We suspect that coverage rates of only about 50 percent would raise doubts about depending solely on small cognitive interview pretests.

18. For a more detailed description of the design and a discussion of the study's key features, see the online appendix, section B.

19. When pretesting is concerned with issues other than problem identification—such as assessing item validity, hypothesis testing, or determining reasons particular problems occur, among others—the factors influencing sample size are likely to be much different and require separate study.

20. It must also be recognized that the needs of large, multipurpose surveys, surveys to support policy decisions, rapid information collections, and small studies for independent researchers vary enormously in their required levels of problem identification and question repair verification.

Even modest increases in sample size can uncover enough additional problems to have practical significance, and probably justify the marginal added costs. The use of substantially larger samples requires deciding how best to allocate the limited resources available in the fixed-cost context of most surveys. One needs to decide whether the expected payoff in problem detection justifies the additional resources. From another perspective, one might consider whether the cost of doing more cognitive interviews could be better used to support a second pretest method. These are important, complex issues that we need to know far more about.

There are at least two other broad areas that both pretest practice and future research should address. First, we need to investigate methods to improve the effectiveness of any given pretest sample. Such methods might include (a) alternative procedures for reviewing verbal reports to decrease the chances of missing problems that have occurred (e.g., employ more rigorous coding methods, or use a combination of independent judges and interviewers to identify problems); (b) experimental designs in which the performance of alternative versions of the survey questions are compared (see, e.g., Blair 2009); or (c) iterative methods, as advocated by Willis (2005).

In an iterative design, interviews would be conducted in a series of small, independent samples, permitting information gained from each sample to be used to improve problem identification in subsequent samples. For example, if an initial random sample of respondents found that question problems seemed to be more prevalent in a particular demographic group, the second-round sample might oversample that subgroup to improve the yield of identified problems.[21]

Second, we need to learn how to tailor pretest procedures and staffing (interview techniques, interviewer staff, sample selection method, etc.) to be most effective for the types of problems of greatest concern at a particular stage of questionnaire development. Consider just the composition of the interviewing staff. It may be that very experienced interviewers will uncover problems at higher rates (or particular types of problems); on the other hand, using interviewers who cost less and are easier to recruit, but have less experience, may permit larger sample sizes. Possibly there may be gains from adjusting the composition of the interviewing staff depending on, for example, the stage of instrument development. It may be that particular types of interviewer experience—conducting cognitive interviews, conducting survey interviews, designing questionnaires, interviewing particular populations, or particular

21. Also note that the question of appropriate sample size remains relevant in these alternative approaches to cognitive interviewing and to other pretest methods that may be used. Despite the agreed importance of pretesting, in general, pretest sample sizes seem to often be set rather arbitrarily, and the adequacy of a chosen sample seldom questioned (for an example, see Blair and Srinath 2008).

knowledge of the survey topic—may increase effectiveness depending on the types of survey questions, population, or specific measurement concerns.

Conclusions

In the course of designing and implementing this study, we have become sensitive to some conceptual points that have seldom been noted, some of which concern pretesting generally. First, the nature of the impact of problems on measurement error is more complex than has been generally acknowledged. Both the problem's prevalence and severity are relevant to measurement error.

Second, severity can be affected in complex ways, involving not only the survey question text and the response task, but also (as in the book-purchasing example) the type of statistic to be estimated, along with characteristics or experiences of the respondents.

Third, attributes of problems—e.g., problem prevalence and distribution across the population, problem type, and problem "discoverability"—may affect what strategies are most effective for detection. Finally, the available detection mechanisms may vary in complex ways and differ considerably in efficiency.

We also reconsidered some general issues about research on cognitive interviewing methodology. In practice, one can find variation in how almost every aspect of cognitive interviewing is implemented. When one adds to this the various types of survey questions (behavioral, autobiographical, attitudinal, etc.) and the problems they engender, it is clearly difficult for any single research study to speak confidently about its implications for the full range of practice. But that does not obviate the need for carefully designed studies that will cumulatively build knowledge. Both experience gained from practice and findings from research will continue to contribute to our understanding of the factors that influence pretest effectiveness.

Essential to research design is clarity about the dependent variables in the study and the factors expected to influence them. For example, if the researcher posits that more or fewer interviews will be *sufficient*, or that interview protocol A will be more *effective* than B, then it is obviously necessary to define "sufficient" or "effective." Such a definition need not encompass every aspect of a concept that is relevant to practice. Equally important to explanatory progress is that the researcher's expectation have a basis, either in theory or in a line of reasoning that predicts the anticipated outcomes. If research on cognitive interviewing and pretesting generally adhere to the practices of other areas of survey research, the future for methodological research on pretesting will be rich and the potential for important advances large.

Supplementary Data

Supplementary data are freely available online at http://poq.oxfordjournals.org/.

References

Ackermann, Allison, and Johnny Blair. 2006. "Efficient Respondent Selection for Cognitive Interviewing." Proceedings of the Survey Research Methods Section of the American Statistical Association, 3997–4004.

Blair, Johnny. 2009. "Experiments for Evaluating Survey Questions: Designs Using Cognitive Interviews." Paper presented at the NCHS Workshop on Question Evaluation Methods. Hyattsville, MD, USA. Reprinted in *Question Evaluation Methods*, edited by Jennifer Mafrans, Kristen Miller, Aaron Maitland and Gordon Willis. Hoboken, NJ: John Willis & Sons (2011).

Blair, Johnny, Allison Ackermann, Linda Piccinino, and Rachel Levenstein. 2007. "Using Behavior Coding to Validate Cognitive Interview Findings." Proceedings of the Survey Research Methods Section of the American Statistical Association, 3896–900.

Blair, Johnny, and Stanley Presser. 1993. "Survey Procedures for Conducting Cognitive Interviews to Pretest Questionnaires: A Review of Theory and Practice." Proceedings of the Survey Research Methods Section of the American Statistical Association, 370–75.

Blair, Johnny, and Kadaba P. Srinath. 2008. "A Note on Sample Size for Behavior Coding." *Field Methods* 20:85–95.

Conrad, Frederick, and Johnny Blair. 2004. "Data Quality in Cognitive Interviews: The Case of Verbal Reports" In *Methods for Testing and Evaluating Survey Questionnaires*, edited by Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

———. 2009. "Sources of Error in Cognitive Interviews." *Public Opinion Quarterly* 73(2): 32–55.

DeMaio, Theresa J., and Ashley Landreth. 2004. "Do Different Cognitive Interview Techniques Produce Different Results?" In *Methods for Testing and Evaluating Survey Questionnaires*, edited by Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

Ericsson, K. Aanders, and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data.* Rev. ed. Cambridge, MA: MIT Press.

Esposito, James L., and Jennifer M. Rothgeb. 1997. "Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment" In *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. Hoboken, NJ: John Wiley and Sons.

Faulkner, Laura. 2003. "Beyond the Five-User Assumption: Benefits of Increased Sample Sizes in Usability Testing." *Behavior Research Methods, Instruments, and Computers* 35:379–83.

Fowler, Floyd J. 2004. "The Case for More Split-Sample Experiments in Developing Survey Instruments." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: John Wiley and Sons.

Krosnick, Jon A. 1999. "Survey Research." *Annual Review of Psychology* 50:537–67.

Olson, Kristen. 2010. "An Examination of Questionnaire Evaluation by Expert Reviewers." *Field Methods* 22:295–318.

Presser, Stanley, and Johnny Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology* 24:73–104.

Turner, Carl W., James R. Lewis, and Jacob Nielsen. 2006. "Determining Usability Test Sample Size: How Many Users Is Enough?" *International Encyclopedia of Ergonomics and Human Factors.* Second ed., vol. 3, ed. Wardemar Karwowski. Boca Raton, FL: CRC Press.

Willis, Gordon B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* Beverly Hills, CA: Sage.

Willis, Gordon B., and Susan Schechter. 1997. "Evaluation of Cognitive Interviewing Techniques: Do the Results Generalize to the Field?" *Bulletin de Methodologie Sociologique* 11:40–66.