# The Effect of Sample Size on Cognitive Interview Findings

Johnny Blair[1], Frederick Conrad[2], Allison Castellano Ackermann[1], and Greg Claxton[2]
Abt Associates, Inc[1]
Institute for Social Research, University of Michigan[2]

## Abstract

An often-mentioned strength of cognitive interview pretesting is its ability to identify most question problems using a few interviews. However, this is not based on empirical research. We report a study investigating how the number of cognitive interviews affects the number of problems identified by conducting a 90 interviews, drawing samples of size 5 through size 50 from the pool of 90 interviews, and comparing the number and impact of problems identified at each samples size. It is clear that small numbers of cognitive interviews, typical of most pretests, fail to detect many problems including some that are quite serious. Even in samples of size 50, some problems are not uncovered. We conclude that conducting more cognitive interviews than are typically carried out is probably a good investment.

**Keywords:** Cognitive interviewing, sample size, question wording problems

## 1. Background

Nearly twenty years of pretesting practice has produced a large body of experiential evidence that cognitive interviewing is effective in detecting flawed survey questions, identifying response difficulties and, to some extent, providing guidance for question repair. The face validity of these results has established cognitive interviewing, in its various forms, as an industry best practice. One feature of standard practice, indeed one of cognitive interviewing's selling points, is the possibility of using relatively small sample sizes. While there are certainly examples of studies employing large samples, some with over one hundred cognitive interviews (e.g. Davis et al. 2001), typical practice relies on far smaller numbers. Willis (2005) notes that samples of 5 to 15 per round of testing are common (p7), and goes on to note that while generally "...the more interviews we can do, the better," small samples are not a fatal drawback for at least three reasons. First, the purpose of cognitive interviews is not statistical estimation; instead the goal is to include "*a variety of individuals* who will be useful in informing our decisions [italics his]." Second, the nature of cognitive interviewing is qualitative; useful information can come from a few or even a single interview. Third, judicious selection of respondents can, even with a small sample, provide coverage of the items following a filter question or other rarely-used paths through the questionnaire.

While each of these points is well taken, each one can, in a particular survey application, suggest the need for a larger than smaller sample. What constitutes a sufficient variety of individuals will vary from one survey to another; and what number of respondents produces an adequate variety may be hard to specify beforehand. While a single interview may provide sufficient evidence of a problem, one cannot know whether the key interview will occur on the fifth case or the fiftieth. And although questionnaire coverage can, to some extent, be handled by careful choices of respondents, a satisfactory sample size could, say in the case of some of the long, complex government health surveys, be quite large because the more possible paths through the instrument the more respondents that are needed to insure that at least some of them encounter the potential problems with all the questions. These points are not to argue for large samples per se, but simply to suggest that the question of adequate sample size needs to be considered for each application and deserves empirical investigation.

The general research question is how might problem detection be affected by changes in sample size? Some problems may be easily identified, while others are harder to root out, perhaps because they are relatively subtle or affect only respondents with certain characteristics or experiences. Similarly, some response tasks may pose difficulties for most respondents, while others affect only a small proportion of respondents, but perhaps cause serious measurement error when they occur.

The core objective of cognitive pretesting is problem identification. Logically, with larger sample sizes, the number of identified problems should be expected to increase. This increase is, of course, limited by the number of problems that exist in the instrument. All unique problems will be identified at some point so the issue is whether this point is 5 or 50 or even more interviews. Assuming there are problems to be found, conducting more interviews allows additional opportunities for each unique problem with each question to be exposed. We distinguish between the

number of unique or different problems and the total number of instances of each unique problem, of which there are almost certainly more.

Cognitive interviewing shares some techniques, such as thinking aloud, and goals, such as problem detection, with the field of usability testing, in which the object is to empirically determine the kinds of errors that people make in using particular interfaces to software systems. The issue of sample size in usability testing has been examined from a few perspectives (e.g. product testing and user interface tests), but the results uniformly come down on the side of small samples being sufficient. Rubin (1994) suggests that if the goal is to "generalize to your specific target population," then a large enough sample of participants is needed to support the planned analyses. If, as is more common, "you are simply attempting to expose as many usability problems as possible in the shortest amount of time, then test at least four to five participants (p128)." Hix and Hartson (1993) suggest that "....three participants per user class is [often] the most cost-effective number (p291)."

Both of these recommendations come with caveats about careful selection of "representative users." Nielsen and Landauer (1993) take a more rigorous approach, in which they estimate the number of problems remaining to be detected based on the number detected in early test iterations. They also used judgments of expert evaluators to identify problems that might potentially be encountered by test users, analogous to expert panel assessment of survey instruments and actual pretest respondents. They recommended (based on a number of empirical criteria) testing between 7 (designated "Small") and 20 (designated "Very large") users. In more recent writing, Neilson has advocated testing 20 users when collecting quantitative usability data (2006).

Perhaps the usability methodology has lessons for survey instrument testing. However, there are enough differences between user interfaces and survey questionnaires and between users and respondents that these lessons may be of limited value. First, the range of intended users of an interface can certainly be wide, but is probably seldom as varied as sample for a national general population survey. To the extent that potential problems may differ by type of respondent, this is an important sample size concern. Second, the task of answering a survey question typically calls on different cognitive skills than interacting with a user interface. And third, an interviewer often administers survey questionnaires whereas computer users typically interact directly with the computer (although, of course, questionnaires can be self-administered and people can interact with computers through another person as when making travel reservations on the phone). So even if the prescriptions about number of test users in usability testing are correct, applying them directly to pretesting questionnaires could be a mistake. Survey researchers require guidance specifically about the number of cognitive interviews they should conduct in a pretest. We report a study here that examines the impact of the number of cognitive interviews conducted on the number and type of problems that are detected.

## 2. Methods

The research design simulates different sample sizes by selecting, with replacement, repeated samples of a given size from a pool of cognitive interviews. So, it is possible to select repeated samples of 5 interviews, 10 interviews, 15 interviews, etc. and examine the results. The design was implemented in three phases. First, cognitive interviews were conducted creating a pool from which samples could be drawn, and each administration of each question was coded for problem occurrence. Second, using this pool of interviews as a universe a set of samples of a given size was selected with replacement. Third, for each set of samples the number and nature of identified problems was determined, allowing comparison across sample sizes.

The questionnaire was designed to represent a range of types of questions and a range of response tasks. Sixty previously pretested questions were selected from major government, academic and commercial surveys[1]. About half the questions were behavioral and half attitudinal (34 and 26 respectively). In order to be sure the questionnaire contained a sufficient number of problems and that these problems were known to the researchers in advance, each question was "damaged," i.e. its wording was modified so that the question would be expected to cause at least one problem for respondents.[2] The types of problems were varied as was the expected impact of the problem on measurement.

The impact of measurement error on a sample estimate depends on the frequency and magnitude of the error. That is, for a particular survey question, how often are respondents' answers affected (frequency), and how much are the answers changed (magnitude of error)? While cognitive interview pretests do not produce assessments of impact, judgments about the seriousness of identified problems are essentially judgments about impact on sample estimates. Since problems can vary in impact from extremely serious to relatively minor, an investigation of problem detection is more meaningful if it includes some evaluation of impact.

A judgment of problem impact was provided by a three-expert panel. The experts independently rated each

experimenter-embedded problem on two dimensions: first, how often did they think, on a percentage basis, each problem would occur in actual data collection; and second, when it occurred how severe, on a scale of 1 to 10, would be the effect on the measurement (where "severe" was defined as the degree to which the problem would distort the answer). These two ratings were multiplied to create a *problem impact* score. The three experts' problem impact scores for each question were averaged to produce a single impact score for each embedded problem. The experts also noted when they thought a question would have any other problems, in addition to the embedded one, and rated these in the same way. Finally, any problems beyond these that were actually detected in the cognitive interviews were similarly rated. All of the problems, and only those problems, that were actually detected in the cognitive interviews were used in the analysis.

Ten interviewers were selected and trained in a cognitive interview protocol that combined think aloud instructions and scripted probes devised by the interviewers.[3] Each interviewer conducted nine interviews in two batches. After the first batch of five interviews, the interviewers were instructed to review the protocol and make any changes in the scripted probes based on what they had found to that point. The revised protocols were used to conduct the second batch of four interviews. In this way, we hoped to represent to some degree the flexibility of cognitive interviewing practice, in which interviewers are free to modify their procedures based on what they have learned to date.

A general population sample of respondents was recruited from a commercial email list. Quota sampling was used to produce a mix of ages, sex, and education. After data collection, two coders, using a problem-coding scheme from Presser and Blair (1994), worked together to reach consensus on coding the verbal reports about each answer in each interview. If interviews had been coded in the order they were conducted, it is possible coders might have expected to find already identified problems in subsequent interviews, which could affect their coding behavior. To avoid this possible confound, a randomization -scheme was used to determine the coding order (details are available from the authors). The basic problem types were semantic and task performance.

Starting with a sample size of 5, 90 replicates of size 5 were selected, with replacement, from the pool. The same procedure and number of replicates was used for samples of size 10, 15, 20, 25, 30, 35, 40, 45, and 50. So, 90 samples of size 5 were selected, then 90 samples of size 10, then 90 samples of size 15 and so on.

Several measures were computed for each sample size: the mean number of unique problems (irrespective of how many instances of each one was observed) per interview; the mean total number of problems per interview, i.e. all instances of all unique problems; the mean impact score. The replicate sampling design, in which 90 samples of each sample size are selected, produces a more stable estimate of these means than would a single sample of a given size, and also permits computation of the standard deviation for each mean.

## 3. Findings

### 3.1 Unique problems

Do small numbers of cognitive interviews uncover most of the problems in a questionnaire? The answer appears to be "no." At sample size 5, the mean number of unique problems is 46. This figure increases in proportional to the increase in sample size so that at sample size of 50 an average of 169 unique problems is identified (see Figure 1). In the 90 cognitive interviews that were conducted, a total of 210 unique problems were identified.

The increased yield of unique problems is most striking at the low end of the range of sample sizes, doubling from sample size 5 (problems=46) to sample size 15 (problems=93). The rate of gain then tapers off, but the number of problems steadily increases as a function of sample size. Clearly, when only a small number of interviews is conducted many problems are not uncovered that do become evident when a larger number of interviews is conducted.

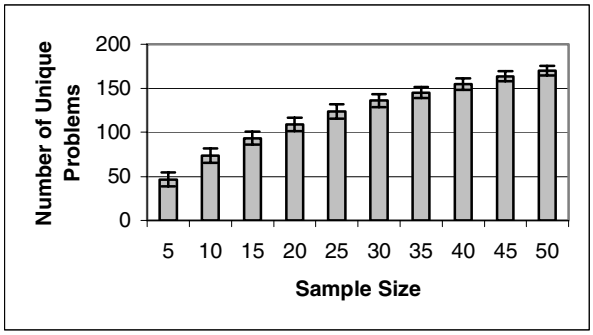


**Figure 1: Mean Number of Unique Problems by Sample Size**

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 at sample of size 5 is 8.02 and

decreases more or less monotonically across sample sizes until at sample of size 50 the standard deviation is 5.61 (see Figure 2). Of course we cannot rule out the possibility that the greater amount of replacement (interviews that are re-selected) across samples at samples of size 50 than size 5 is partly responsible for the drop in standard deviation. But it certainly is sensible that the larger the number of interviews the smaller the variance due to particular samples.

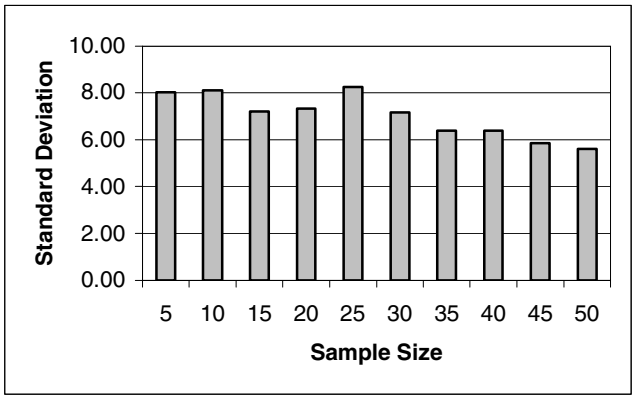


**Figure 2: Standard Deviation for Mean Unique Problems by Sample Size**

### 3.2 Unique versus total problems

Of course, many problems occur multiple times in a given sample size. The mean number of total problems (the count of all instances of all unique problems) ranged from 68 (s.d. 11.02) for a sample of 5 to 665 (s.d. 26.70) for a sample of 50. Although more unique problems are identified the larger the sample size, the rate of increase for identifying unique problem appears to slow down slightly starting at about sample size 20. In contrast, the rate of identifying total problems is consistent across sample sizes. The counts for unique and total problems at each sample size are displayed in Figure 3. The stable rate of increase for total problems is sensible considering that more interviews create more opportunities for instances of unique problems to be exhibited. This would be true even if unique problems did not increase with sample size. Thus we find the average number of unique problems to be a more diagnostic measure of how sample size affects cognitive interview results.

### 3.3 Impact

Although the number of unique problems increases with sample size it is possible that the most serious problems are detected with small numbers of interviews. We tested this possibility using our impact measure. Impact combines an expert judgment about how frequently a problem might occur with the corresponding judgment

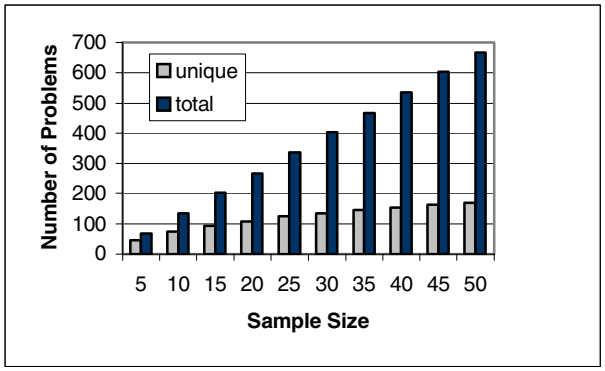


**Figure 3: Mean Number of Unique and Total Problems by Sample Size**

about its effect on measurement error when it does occur. A problem judged to occur seldom and to have a small effect on the measurement when it does would be at the low end of the impact continuum, while a problem expected to occur frequently and have a severe impact on measurement would be at the high end. High frequency and low severity or its converse would produce mid-range impact values.

To examine how impact affects problem detection, the problems were divided into impact quartiles (first quartile problems are lowest impact, fourth quartile are highest impact). Figure 4 shows the proportion of all unique problems from each impact quartile identified at each sample size. Although a large proportion of the highest impact (4th quartile) problems are detected even at the small sample sizes, additional high impact problems continue to be detected as the sample sizes increase. The proportion of less serious problems, in each of the other three quartiles, also increases. However, about a quarter of the less serious problems remain undetected even at sample size 50.

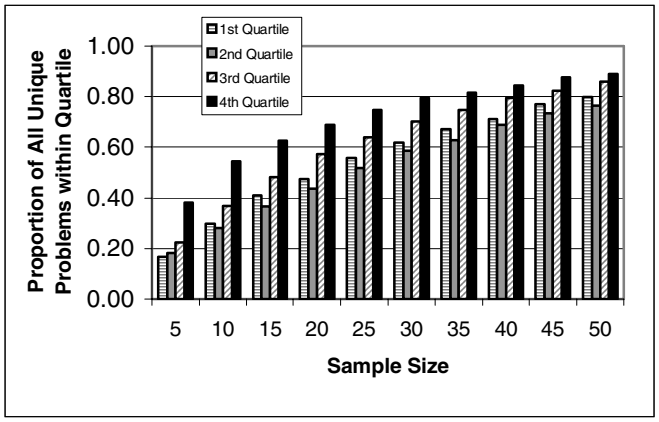


**Figure 4: Proportion of all unique problems from each impact quartile detected at each sample size.**

The proportion of problems from each impact quartile changes with sample size so that proportionally more high impact problems are detected at small sample sizes while at larger sample sizes high and low impact problems are equally likely to be detected (see Figure 5). Thus small numbers of cognitive interviews do expose proportionally more high impact problems but large numbers of interviews expose substantially more problems at all levels of impact.

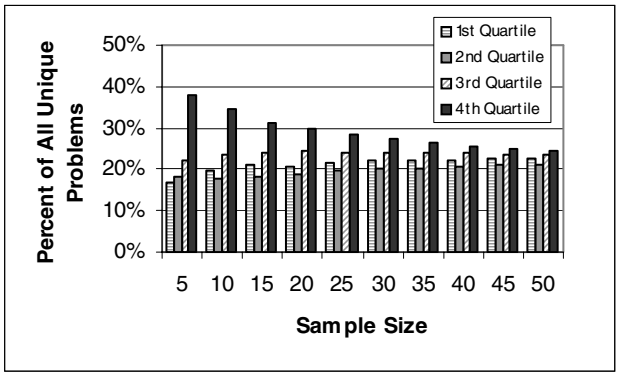


**Figure 5: Percent of unique problems detected at each sample size by impact quartile.**

### 3.4 Likelihood

Irrespective of impact, some problems may be sure to be detected at small sample sizes while, for other problems, certain detection may require larger numbers of interviews. To address the question of how probable it is that a particular problem will be detected with a given sample size, we constructed a likelihood estimate. The likelihood of a problem being detected is simply the number of samples of size n in which the problem was identified divided by the total number of samples of size n that were selected. For example, if 90 replicates of size five were selected and in 30 of those samples a particular problem was identified, then the likelihood of that problem is 30/90 = .33. Figure 6 shows the likelihood, in five categories, of problems being detected at different sample sizes[4]. At sample size 5, only a tiny percentage of problems have a 100% change of being detected while over 70% of all unique problems have a 25% or smaller chance of detection, i.e. they are rare. The percentage of rare problems decreases steadily as sample size increases, with that category virtually disappearing at sample size 30.

However, detection of most problems is far from certain in samples of 30 and larger. Even by sample size 50, only about half of the problems are detected frequently (76% to 99% chance of detection) and only 23% of problems are detected in 100% of the samples of size 50.

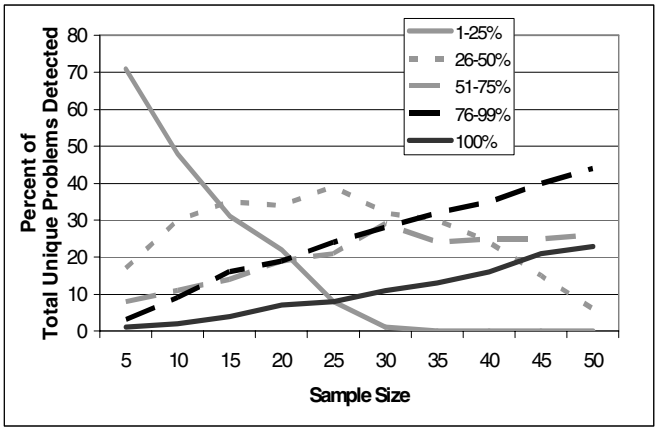


**Figure 6: Likelihood of Problem Detection at Each Sample Size**

## 4. Discussion

These findings show a strong relationship between sample size and problem detection. Whether examined from the perspective to total unique problems identified, or the likelihood of particular problems being detected, increasing the sample size increases the number of problems detected.

While methodological research on cognitive interviewing, such as the current study, cannot encompass the range of procedures used in practice, problem detection is certainly a central goal of cognitive interviewing. It is clear that for the range of questions and problems examined in this study, small sample sizes miss a large number of problems of all types, including many that are quite serious.

The major question raised by these findings is how many interviews are needed to be confident that a questionnaire is relatively problem-free. Based on just this study, it is too soon to say: only one set of questions was tested with only one cognitive interviewing protocol. Changes to these or other factors could affect the optimal number of interviews. Yet even at this early stage it seems safe to say that more cognitive interviews is probably a good investment.

## Acknowledgements

We thank Abt Associates, the Survey Research Center at the University of Michigan, and the Bureau of Labor Statistics for support. In addition, we thank Laura Burns, Stephanie Chardoul and Clyde Tucker for advice and assistance.

## References

Davis, D., J. Blair, E. Crawley, K. Craig, M. Rappoport, C. Baker and S. Hanson (2001) Census 2000 Quality Survey Instrument Pilot Test, Final Report, Development Associates.

Hix, D. and H.R. Hartson (1993) *Developing User Interfaces: Ensuring Usability Through Product and Process*, Wiley.

Neilson, J. (2006). "Alertbox," June 26, www.useit.com/alertbox/quantitative_testing.html.

Neilson, J. and T. K. Landauer (1993) "A Mathematical Model of the Finding of Usability Problems," *Interchi 93.*

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

Rubin, J. (1994) *Handbook of Usability Testing*, Wiley.

Willis, G. B. (2005) *Cognitive Interviewing: A Tool for Improving Questionnaire Design*, Sage.

## End Notes

[1] 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 polls.

[2] There were a few instances when questions did not need to be altered because they were judged to already be sufficiently problematic.

[3] The term "protocol" is used differently in the psychological and survey literatures. For the purposes of the present paper, the term indicates the plan for actual conduct of the cognitive interviews, including instructions to respondents, scripted probes, notes to the interviewer or any other instruction or guideline for the interviewer. In psychological studies using think aloud methods, "protocols" refers to participants' verbal reports.

[4] 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, and 50.