Improving Survey Methods with Cognitive Interviews in Small- and Medium-scale Evaluations

Katherine Ryan
Nora Gannon-Slater
Michael J. Culbertson

This manuscript has been published in the American Journal of Evaluation (2012), v 33, no 3, pp 414-430, doi:10.1177/1098214012441499

Credible evidence that informs policy making decisions across all domains in evaluation (e.g., education, health care, and public administration) is of vital importance (Donaldson, 2008). A common method of gathering evidence is through surveys that are used to evaluate intended and unintended policy effects (Gottsfredson & Gottfredson, 2002; Palardy & Rumberger, 2008; Trenholm, et al., 2007). Survey questionnaires are employed in large-scale evaluations because of their capacity to gather maximum amounts of information from large, diverse populations across a wide range of contexts; their potential for producing generalizable findings; and the relative efficiencies (e.g. cost) of the method (Desimone & le Floch, 2004).

Findings from these kinds of self-reported, structured questionnaires are used to inform policy in a variety of domains, including the effectiveness of prevention programs, the efficiency of healthcare systems, and the efficacy of environmental programs (e.g., Abt Associates; Westat; Gottfredson & Gottfredson, 2002). Educational survey findings are used to monitor the status of reform efforts, including policy implementation fidelity, and establishing links between reform practices and desired policy outcomes (Desimone & le Floch, 2004; Supovitz & Taylor, 2005). Because surveys provide a significant source of evidence in policy-making, ensuring the validity and reliability of interpretations derived from self-report questionnaires is crucial. There is evidence that well-designed questionnaires can yield valid questionnaire interpretations in educational surveys, for instance (Mayer, 1999).

However, questionnaires are complex instruments that yield misleading results when not well-designed. In addition to inflexible response formats, the ability to probe responses is limited (Groves, et al, 2009). While there are a variety of issues related to the quality of survey evidence (e.g., sampling precision and sample size), the validity of response processes--how respondents process thoughts, ideas, views, perceptions, and experiences when answering survey questionnaires is critical. Since the 1980s, survey methodologists and psychologists have pursued interdisciplinary research, known as the cognitive aspects of survey methodology (CASM), to investigate how respondents understand and interpret survey questions using a variety of methods (e.g., coding reaction time, cognitive interviews).

While there continue to be unanswered questions about, for example, the extent of error reduction and sampling, CASM research has influenced survey development practices (Schwarz, 2007; Tourangeau, Rips, & Rasinski, 2000). Cognitive interviewing, one of the most widely used methods, refers to a set of techniques (e.g. think aloud protocols, verbal probes) that enable a researcher to deeply analyze how respondents understand the survey questions they are to answer (Tourangeau, et. al., 2000). Currently, cognitive interviews are routinely administered as part of questionnaire design, piloting, and refinement in well-funded, large-scale national evaluations in a variety of domains (e.g., education, healthcare) (Desimone & le Floch, 2004; Irwin, Varni, Yeatts, and DeWait , 2009).

In this paper we consider the benefits and challenges of conducting cognitive interviews in medium- and small-scale evaluations that typically have more limited resources. We argue that despite resource limitations, the process of testing survey questions with cognitive interviews, as part of questionnaire design and refinement, can lead to better informed judgments about the potential quality of survey evidence, thus justifying additional costs. We begin with a

brief overview of surveys and the study of survey response processes. After sketching CASM foundations, we describe the relationships between cognitive theory, the CASM response model, methods, and survey responses. To illustrate how cognitive interviewing can improve respondents' interpretations of a survey, we present examples of cognitive interviews conducted as part of small- and a medium-scale evaluation projects in education and healthcare. The paper concludes with a brief discussion about how the use of cognitive interviews can be enhanced in survey development, refinement, and adaptation to increase the validity of survey questionnaire interpretations used in evaluation studies having limited resources.

The Importance of Cognitive Interviews

Examining whether respondents' interpretations of self-reported items are consistent with intended meanings is fundamental for judging whether survey results provide valid interpretations. Studying survey response processes (response processes aspect of validity) includes empirical investigation of (a) the processes, strategies, and knowledge that underlie item and/or task performance and (b) whether the meanings and interpretations of these item or scale responses remain the same across persons, groups, and contexts (Messick, 1995). Sources of evidence for investigating response processes involve both quantitative and qualitative methods. Quantitative methods (e.g., factor analytic methods) are routinely used to analyze the relationships between questionnaire items and scales as part of the survey development process for psychological scales (Messick, 1995; Trochim, 2006). More recently, response process studies based around cognitive interviews that examine respondents' thought processes while giving verbal reports have become recommended survey practices (American Educational Research Association, American Psychological Association, & National Council of Educational

Measurement [AERA], 1999; American Association for Public Opinion Research, http://www.aapor.org/Best_Practices/1480.htm).

Models and Processes of Answering Survey Questions

While several models of the survey question-answering process have been proposed to better understand survey respondents' cognitive work in answering questions, CASM researchers agree that there are a series of processes or interrelated tasks that survey respondents engage in when answering questions. These can be characterized as having four components (Bradburn, 2004; Schwarz, 2007): (a) understanding the question (comprehension), (b) retrieving relevant information (retrieval), (c) preparing one's answer (judgment), and (d) formatting and editing an answer (response). Below we define and identify key issues associated with each component. Note that while these components and specific processes are presented sequentially, in practice the components cycle back and forth or overlap as illustrated in Figure 1.

[Insert Figure 1 here]

Comprehension. Comprehension includes paying attention to instructions and questions, making sense of the question, determining what information is being asked, and making connections between key terms in the question and relevant concepts (Tourangeau et al., 2000). The goal is for the respondent to understand and interpret the question in the same manner as the questionnaire developer (Tourangeau & Bradburn, 2010). Respondents' misunderstanding or alternative interpretation of a question can be the result of a wide variety of problems. These include missing part of the question, not understanding directions, not reading the directions, or being confused by "complex questions" with detailed qualifiers. Comprehension problems can also be related to unfamiliarity with or misunderstanding of terms (Tourangeau et al., 2000).

Lexical ambiguities are one type of comprehension problem that occurs when words have different meanings (Bradburn, 2004).

Retrieval. After respondents grasp the question, they retrieve information to answer it. Retrieval involves bringing to mind information from long-term memory (stored memory) and short-term memory so it can be used. Retrieval component processes include adopting recall strategies, cueing to activate recollections, recalling memories, and making inferences that complete partial recollections. Survey question wording, definitions, other survey material, and emotions or images evoked serve as retrieval cues that activate survey respondents' memories in searching for information. Semantic memory involves vocabulary, structural language features, and conceptual knowledge and is distinguished from episodic memory, which includes events and actions that occurred in time and space (Tourangeau et al., 2000). Asking respondents to answer questions about behavior is associated with episodic memory while attitude questions are connected with semantic memory although respondents may also draw on episodic memory (Bradburn, 2004).

Retrieval is influenced by a wide variety of factors. These include memory sources (first-hand experience or second-hand knowledge), length of time since actions occurred, the quality and number of question cues (examples), and the fit between question terminology and survey participants' experiences (Tourangeau et al., 2000).

Judgment. In addition to processing question and survey context cues (e.g., wording, question location, question type) in formulating an answer to a question, survey respondents engage in a variety of cognitive tasks to extend and integrate what they retrieve from memory. These include evaluating the importance and completeness of recalled information or knowledge, making inferences about any gaps in what was retrieved from long-term memory, synthesizing

the information retrieved in order to answer the question, and estimating gaps to adjust for what is missing (Tourangeau et al., 2000). The crucial issue is whether the respondent is able to appropriately assess the relevant information. Most researchers agree that these judgment "processes" are applicable whether survey respondents assess facts, behaviors, or attitudes (Tourangeau, et al, 2000; Grove et al., 2009) while acknowledging noteworthy distinctions when respondents answer these different kinds of questions (e.g., facts, attitudes) (Bradburn, 2004; Schwarz, 2007).

Respondents may be unwilling or unable to make a judgment based on the information they possess. Others individuals may take short cuts to bypass the cognitive tasks required to make a thorough judgment or simply interpret a question superficially, which is known as satisficing (Krosnick & Presser, 2010). Decreasing task difficulty in questions can lead to more accurate respondent self-reports (Krosnick & Presser, 2010).

Response. After preparing a judgment, the respondent chooses an answer and communicates it (Tournageau, et al, 2000). A key task for people is to fit their answer to the response format offered. Most important, there may not be a good fit between the answer respondents formulate and the response options provided. Since questionnaire items are often precoded, selecting the appropriate scale or response option introduces a variety of issues regarding boundaries between response categories such as "strongly agree" and "agree," order effects (first option selected more frequently), and others. Once an answer is selected, people may "edit" their answer to see if it is consistent with how they answered other questions and for its social desirability (e.g., respondents' claiming a behavior or attitude that puts them in more favorable circumstances) (Bradburn, 2004). Administering anonymous questionnaires diminishes this effect, but it does not eliminate social desirability bias.

Cognitive Interview Methods

The CASM theoretical foundations and the early cognitive interview approach are formulated around studying cognition with verbal protocol methodology (Sudman, Bradburn, & Schwarz, 1996; Tourangeau et al, 2000). The argument for using verbal protocol methodology (a verbal report and protocol analysis) to study cognition is based on a multitude of studies from diverse areas (e.g. decision making, text comprehension). (See Ericcson & Simon, 1993, for a review of these investigations.) A key assumption when collecting verbal accounts of individuals' thinking is that people can report on what is occurring in their working memory, where active thinking takes place (Bradburn, 2004; Conrad & Blair, 2009). Active thinking includes both short-term memory and what is retrieved from long-term memory. Thus, the respondent's verbal report helps the researcher gain insight into the relevant cognitive processes that are taking place.

Over time, the scope of cognitive interviewing has evolved substantially through focused study of how survey respondents answer questions. Cognitive interview techniques have included concurrent think-aloud protocols, focus groups, verbal probes, hybrid approaches, and others. In the next section, we present an overview of the cognitive interview methodology. We describe three main techniques: the think-aloud protocol, intensive verbal probes and a hybrid model.

Cognitive Interview Techniques

Think-aloud approach. The think-aloud protocol asks participants to provide an account of what they are thinking as they respond to a survey item (the think-aloud protocol) or just after responding to the item (the retrospective protocol) (Tourangeau et al., 2000). Think-aloud protocols are typically administered under standardized conditions. After a brief description of

the survey, participants receive training in how to "think-aloud" using sample questions. Participants then read survey items and think-aloud to report what they are thinking as they respond to a question. While a small number of participants may have difficulty thinking-aloud (Willis, 2005), most adults and even children can think-aloud successfully with a modest amount of training. The evidence regarding whether thinking aloud interferes or changes individuals' response processes is mixed (Blair & Conrad, 2009; Willis, 2005).

The think-aloud cognitive interviewer is primarily an observer and data collector requiring modest training and skill. The interviewer merely provides instruction in thinking out loud and intervenes only to remind people to think aloud after a period of silence. In addition, the cognitive interviewer records what the respondent says, as well as any participant gestures or other informal communication (e.g., sighs or hesitations).

Verbal probing. As an alternative to the think-aloud technique (Beatty & Willis, 2007; Willis, 2004), the verbal probing technique asks pointed questions about participants' thinking, rather than just recording what they spontaneously report (Blair & Presser, 1993). Probing can be used to specifically target the various cognitive processes (e.g., comprehension, retrieval, etc.,) involved in answering survey questions, but may change the context for subsequent questions (Willis, 2005). Probes may be standardized or non-standardized, though empirical research provides some support in favor of standardized probes (Conrad & Blair, 2009).

Hybrid model. Although the two cognitive interviewing approaches are often presented in the literature as mutually exclusive, the boundaries between these approaches are blurred in practice. Research reviews and findings from an empirical investigation of cognitive interview practices suggest that, more often than not, survey developers advocate using probing techniques in conjunction with the think-aloud method (Blair & Brick, 2009; Beatty & Willis, 2007).

Further, empirical evidence suggests that even within interviews that are solely probing, participants still think-aloud, although much more spontaneously (DeMaio & Landreth, 2004). Thus, practical decisions when using a think-aloud should consider how much probing is appropriate and to what extent probing should be standardized or determined by interviewer judgment. After these critical decisions, a variety of design considerations remain, including whom to interview, the number of interviews and data analysis approaches.

Cognitive Interview Sampling and Data Analysis

Although cognitive interviews are not conducted for the purpose of generalizing or to statistically represent a larger population, there is no consensus about participant selection and adequate sample sizes (Willis & Beatty, 2007). There is an assumption that conducting a modest number of cognitive interviews will reveal the most critical problems. For large-scale national survey panels, current practice recommendations suggest that cognitive interviews be conducted in rounds ranging between 5 and 15 (Willis, 2005). In practice, time and resource constraints typically determine the number of cognitive interviews that can be conducted (Beatty & Willis, 2007).

In selecting the cognitive interview sample, convenience or quota sampling strategies are typically employed (Ackerman & Blair, 2006), with some attention paid to diversity. Purposive sampling approaches are emerging in specific domains, such as healthcare, when a survey is targeting a specific population of interest (Irwin, et al., 2009). For example, in the development of a health outcomes survey, cognitive interviews were conducted with those who do (children with asthma) and do not (children without asthma) have a key characteristic being studied with the survey questionnaire (Irwin, et al., 2009).

In general, the cognitive interview literature is not specific about data analysis procedures. The goal of analyzing cognitive interview data is to reveal problems respondents have with (a) the survey context, (b) understanding questions, (c) retrieving and integrating information used to answer questions, and (d) communicating answers in order to revise or repair questions. Some form of coding is often used, such as codes based on cognitive models (see Figure 1). Although useful for summarizing overall results, coding is criticized for being time consuming and so reductive that the information obtained is inadequate for repairing questions (Collins, 2007; DeMaio & Landreth, 2004; Willis, 2005). In place of elaborate coding, an in-depth response analysis at multiple levels might include, for example, looking at individuals' cognitive processing problems during individual interviews, consistencies and inconsistences in question response processes and patterns across interviews, and comparisons of subgroup response differences to explore potential bias issues.

Can Cognitive Interviews Improve the Quality of Survey Evidence in Evaluation?

Cognitive interviews are widely used by university survey centers, government agencies, and commercial survey enterprises (Beatty & Willis, 2007). Cognitive interviews are also employed in questionnaire development for large-scale evaluation projects using large, well-defined probability samples to investigate and generalize about the outcomes or impacts of large policies or programs. Requiring robust human, financial, and time resources, these kinds of large-scale evaluations include, for example, healthcare outcome evaluations (Irwin, et al., 2009, American Institute for Research, 2009), youth development outcomes evaluations (Sabaratnam & Klein, 2006), and educational impact evaluations (Desimone & le Floch, 2004).

The question remains whether there are potential benefits that outweigh constraints and costs when using cognitive interviews more broadly, in questionnaire development and

refinement for small- or medium-scale evaluations. Characteristics of small- and medium-scale evaluations, such as evaluation purpose, sampling, and resources (financial, human, etc.) may be different than they are in large scale evaluations (Anderson & Postlethwaite, 2007; Howell & Yemam, 2006; Robert Wood Johnson Foundation, http://www.rwjf.org/files/research/022908knickmanhuntanthology.pdf). Nevertheless, evaluators conducting small- or medium-scale evaluations—often focused on studying program or policy implementation or needs assessment with representative samples with restricted resources (e.g., fiscal, human, etc.)—also rely on questionnaires to collect evaluation evidence.

In the next section, we present examples from two different types of evaluations (needs assessment and policy implementation) from two domains (healthcare and education) to demonstrate the utility of cognitive interviewing for improving survey response interpretations in small- and medium-scale evaluations. In these examples, cognitive interviews revealed significant problems related to the cognitive model components: retrieval (specifically, the use of the non-substantive response category "I'm not sure" and context effects), and judgment (in formulating a single judgment). Furthermore, we show how cognitive interview data can be used to improve and refine survey questions. We underscore the benefits as well as the technical and practical constraints of implementing cognitive interviews in these kinds of evaluations.

Medium-Scale Education Policy Implementation Evaluation

The first two examples reflect cognitive interview data gathered over two years during the *development* and *refinement* of a survey questionnaire used for a comprehensive, single state, four-year No Child Left Behind (2002) (NCLB) policy implementation evaluation ($300,000 per year). The questionnaire's purpose was to describe stakeholders' (e.g., teachers and principals) experiences and perspectives across five areas of assessment consequences relevant to

instructional practices, local assessment practices, use of test data, school policies and practices, and the teaching profession. The questionnaire assessed abstract concepts or beliefs such as perceived changes in educator collaboration.

Further, we thought conducting cognitive interviews was especially important because we were administering this questionnaire in 2009 and forward during the later years of NLCB policy implementation. Some schools had already received significant sanctions (e.g., reduction in federal funds allocated to the school or a district or state takeover of the school) based on school-wide NCLB accountability testing (U.S. Department of Education, 2009). Educational researchers have illustrated that there can be differences among schools in how teachers understand and interpret various concepts terms, etc. used in survey items assessing educational policy (Desimone & Floch, 2004). Acknowledging that disparate contexts could compound any problems in response validity, we systemically investigated variations in question meaning with respect to student and school/community contextual factors (e.g., income level, population density) as well as institutional role (e.g., principal vs. teachers). In addition to cognitive interviews, the survey development process followed recommended steps in survey design (e.g., extensive research synthesis, reviews by content and survey experts, etc.).

Cognitive interview design and analysis. The cognitive interviews, based on the hybrid model above, consisted of a think-aloud protocol and retrospective, non-standardized probes used to clarify issues with specific items. For example, non-standardized probes might include such questions as *How do you understand the word (phrase) … ? (or) What do you think of when you hear … ?* that were asked after the participant finished 'thinking aloud' about the item. The cognitive interviews were conducted by advanced doctoral students (e.g., evaluation, psychometrics) (N=4) who participated in cognitive interview training (readings and practice)

adapted from the cognitive interview and verbal protocol literature (e.g., Ericcson & Simon, 1993). Each cognitive interview was audio-taped with a team of two individuals (interviewer and note-taker) carrying out the interview.

The cognitive interviews were conducted before and after the questionnaire pilot administration. There were not adequate resources or time to conduct the number of cognitive interviews that are recommended for a large-scale national survey panel (rounds of 5 or more). In total, 15 cognitive interviews were conducted over two years—11 with elementary and middle school teachers and four with principals. The interviews were conducted in rounds of 2-4 interviews each with revisions between each round targeting specific populations of interest. Respondents came from large, diverse districts that were experiencing greater difficulties in meeting NCLB established annual achievement performance targets (AYP)—overall and for subgroups (e.g., low-income)—as well as in smaller, more homogenous districts. Interviewees included both novice and veteran educators. All but four interviewees were women.

In analyzing the data, we organized the issues that emerged during the interviews into four levels: general, construct dimension (items that address a specific area of assessment consequences), item scales, and individual items. If two or more cognitive interview participants within or across rounds exhibited difficulty in responding or expressed confusion in understanding content on any four of the levels, the survey development team would examine the problem and suggest improvements, which were also tested. All revisions were subject to re-testing, expert review, and final revision prior to questionnaire administration.

In most cases, revisions were straightforward and included word changes or new graphic presentation of the item. The survey would have yielded ambiguous or misleading questionnaire

results if the questions had not been subject to revision and retesting, as example 1 below illustrates.

Example 1: Detecting judgment complexities, 'better and/or worse'. It is typically assumed that the underlying attitude about a concept exists on a one-dimensional scale. That is, even if a person retrieves pieces of information on a complex issue (e.g., attitudes about accountability) that are contradictory or thought to represent "both sides of the argument," he/she will still integrate that information and compute a single judgment about the specific issue (Tourangeau, 1984; Sudman, et al., 1996). Based on the question's content and response alternatives, a respondent is expected to decide if something has increased or decreased, improved or worsened, but *not* both. Researchers studying the early years (2004 and 2005) of NCLB implementation used global survey items to assess a few broad changes such as 'teaching practice', 'principal's effectiveness as an instructional leader', etc. with a one-dimensional scale (e.g. changed for the worse, no change, changed for the better) and specific items (Hamilton, et al., 2007). For example, when asked to assess how their teaching practice had changed, only 5% (GA) and 10% (CA) of the teachers reported their own teaching practice had changed 'for the worse' (Hamilton et al., 2007). However, some of their comparisons of findings from specific items (e.g. changes in test item formats) did suggest teachers' responses to NCLB were mixed.

In addition to items assessing changes to specific educational practices (e.g. teaching of tested topics), we constructed three "global items" that required that participants make overall judgments about three broad areas of NCLB high stakes assessment policy consequences (instructional practices, school policy and practices, and the teaching profession). Each participant was asked "how much overall change had taken place, and was it for the better or worse?" The scale was constructed under the assumption that participants would respond that

instruction was to some degree better *or* worse as a result of NCLB testing. These global items with the one-dimensional scale were fielded in the first round of cognitive interviews with teachers (Figure 2). As the following cognitive interview example illustrates, participants' perceptions of the overall effects of NCLB policy were not easily characterized.

[Insert Figure 2 about here]

In the first round each teacher exhibited difficulty in forming a single opinion on the issue and expressed frustration with having to respond to the item. For example, in responding to the instructional practices item, one teacher struggled to integrate her negative views of a narrowed curriculum with the positive outcomes she saw in the increased emphasis on math and literacy skills:

> *Teacher*: Basically, how I have altered [my instruction] is in terms of what I teach. How much time I spend on teaching since [state] testing—[it is] just my impression that reading and math drive curriculum. Science and social studies are core time-focused losers.
>
> *Facilitator*: How would you respond to this question?
>
> *Teacher*: I would say it changed. …what I see it goes both [negative and positive] ways. Teachers are finding time for math instruction…something that was shut away in the past. And there is an emphasis on really helping kids develop the literacy skills they need. What's being given up is that we think about the literacy and math…
>
> *Facilitator*: So if you were taking this survey at home would you mark anything?
>
> *Teacher*: I'd probably mark a 4 and a 2.

We found similar ambiguities in other global items including one that asked teachers how NCLB policy implementation had changed their school. Clearly, the question content and response alternatives for these global items were insufficient in describing the views teachers would have on this issue. As a solution, we split each global question into two items—one asked about the extent of positive NCLB consequences and the other assessed negative effects (Figure

3). We fielded these items in the second round of cognitive interviews with both teachers and principals.

[Insert Figure 3 here]

In contrast to Round 1 cognitive interview results, Round 2 teachers provided reasons for their responses to each item, further reinforcing our observations that a single judgment about an issue would be too difficult for participants to make. For example, one teacher explained after each item:

> Positive effects item: I thought it really focused instruction and getting students to understand the math and getting away from rote learning.
>
> Negative effects item: When I first started, the school I used to work at before did a couple workshops on [state] testing. Basically, the workshops give you the format of the test and basically told you to teach to the test. I didn't think that was such a good thing...so I marked very little negative effects [2]. I don't think [state] testing affected [my instruction] in a huge way but when I first started I think it did.

We fielded the revised items during the pilot administration (N=860). As predicted from the cognitive interview findings, teachers reported both positive and negative NCLB consequences in three areas: instructional practices, school policy and practices, and the teaching profession. Overall, teachers perceived more negative accountability influences than they did positive influences (standardized differences (SD), ranging from -.5 to -1.2). There were also notable standardized differences (0.3 to 1.2) between teachers and principals' perspectives regarding negative *and* positive consequences in the three areas. In all cases, teachers reported *more* negative effects and *less* positive effects than did principals. These results were purposefully compared with focus group findings to elaborate *why* teachers reported negative and positive effects in this way (Author, Gandha, ZZZZ, Lim & Wakita, 2009).

Validity of questionnaire interpretations. In the above example, cognitive interview findings revealed that participants were unable to integrate retrieved information to form a single judgment about overall NCLB assessment and accountability policy effects. Had we fielded the items as they were originally written, respondents would have been forced to choose a scale point that did not fully reflect their judgment about the information in the item. Instead of revealing teachers' complex views about the negative *and* positive assessment consequences, the questionnaire findings would have inevitably led to our making inappropriate conclusions about how teachers view accountability assessment consequences as either negative *or* positive. Fundamentally, the interpretations of the questionnaire results would be compromised.

Further, noteworthy differences between teachers and principals' views on important issues would have been masked. Breaking the original item out into two items that independently assessed positive and negative effects allowed divergent teacher and principal perspectives to be disclosed. The resultant findings are likely related to the distinctive institutional roles or identities held by principals and teachers. Although both kinds of educators seek to improve student learning, classroom teachers are primarily responsible for the planning and delivery of instruction and for evaluation of specific learning outcomes of a specific group of students. In contrast, principals aim to improve student learning through a broad deployment of resources, managing a team of teachers and support staff, and setting daily routines in the school setting. The questionnaire revisions made as a result of the cognitive interview findings allowed for those different perspectives to be revealed; this enriched our interpretation of the questionnaire findings. Nevertheless, cognitive interviews are not a remedy for all questionnaire problems as we show in example 2.

Example 2: Detecting issues related to retrieval, 'I don't know'. A quality item on a questionnaire is one that a participant can answer because s/he has the information required (Czaja & Blair, 1995). If a participant fails to retrieve relevant information because it is not available, s/he is going to provide an inaccurate response or omit responding to the item altogether. The survey literature on whether to include a 'don't know' (DK) response option is mixed. Some researchers are concerned that that including a DK option can encourage satisficing or choosing "I don't know" simply because a participant does not *want* to engage the cognitive processes required to answer the item (Krosnick & Presser, 2010). Other survey researchers suggest that if a respondent cannot provide an answer to an item, then that item should probably be eliminated from the questionnaire *unless* there is sufficient reason to provide DK options (or filters) (Czaja & Blair, 2005).

Recent literature on accountability contends that in order for accountability policies (like NCLB) to be effective, a school-wide systematic effort is needed that requires teachers to know about and be supportive of reform beyond their own classroom (Elmore, 2004; Hawley & Rollie, 2007). During the initial survey questionnaire development (Year 1 cognitive interviews and pilot administration), all items including nine items that examined changes in school-wide policies and practices were studied. The analysis of Year 1 cognitive interview data revealed few issues with the school policy and practice items; only two out of eight teachers communicated that they were unable to answer two of the nine items.

Based on these results and the survey literature (e.g., Krosnick & Presser, 2010), the DK option was purposefully omitted from these items during the pilot administration to discourage satisficing. However, analysis of pilot administration data revealed little variance in this item series. Specifically, more than 50% of respondents marked 'no change' and more than 10% of

respondents chose not to answer most of the items in the set; this was substantially different than response patterns in surrounding items. Further, this pattern was observed with only teachers and not principals.

In order to *refine* the questionnaire, the research team used additional cognitive interviews (Year 2) to specifically understand why teachers were all responding one way or omitting responses to these items. When presented with the item in Figure 4, teachers gave responses such as "I would have to say that I don't know on this one. So I would say not changed but more that I don't know." For six of the nine items in the topical series, teachers explained during interviews that they did not know how to answer the particular item.

[Insert Figure 4 about here]

The responses to these items in the cognitive interviews indicated that teachers had trouble retrieving relevant information needed to respond to the item. They understood what was being asked of them but did not contain the knowledge required in order to respond. As a result, teachers chose the middle choice of "no change" believing it to be the closest to what they knew.

Clearly, the response behavior undermined questionnaire interpretations since teachers reporting that *no change had occurred* is arguably different from teachers *not knowin*g if certain kinds of change were taking place at the school as a result of NCLB. Based on the survey pilot results and the Year 2 cognitive interview data, we included a DK option as a response option for each of the items in this subset for Year 2 survey administration. Data analysis of these items in the Year 2 survey administration revealed a moderate (6-7%) or substantial (13-17%) decrease in the endorsement of "no change" across most items. The percent of "I don't know" responses ranged from 5% to 30%, indicating that teachers purposefully chose the DK option to reflect their knowledge about the content of a particular item. By including the DK option in the Year 2

questionnaire, were more confident about questionnaire interpretations regarding teachers' perceptions of and knowledge about changes taking place at the school level as a result of NCLB.

Validity of questionnaire interpretations. Example 2 illustrates some of the limitations of cognitive interviews. While the questionnaire was clearly in the development stage, the Year 1 cognitive interview design (e.g., sampling, scope) was not robust or sensitive enough to detect issues revealed later in the pilot administration. However, the validity of questionnaire interpretations was strengthened when evidence from the Year 2 cognitive interviews, survey and accountability literature, and survey pilot data was used to make a decision about using the DK option. While multiple sources of evidence are often used as a guide in decision-making and judgments, this may be especially important for medium- and small-scale evaluations with limited resources for conducting cognitive interviews.

Small-Scale Healthcare Needs Assessment

Example 3 comes from a small-scale evaluation of the healthcare needs of engineering and natural sciences students at a large, Midwestern, public university. Although the survey would eventually be sent electronically to 10,000 students (about a quarter of the student population), this evaluation was small-scale due to its almost non-existent budget and limited human resources. Providers at the campus healthcare facility had noticed that science-related students participate in health services at lower rates than do other students, so they were interested in reaching out to this group of students that represent almost half of the University student population. However, health services personnel had little experience with the particular healthcare needs of these students, and thus, not enough information to inform targeted outreach activities. Since little previous work has investigated health issues facing science-related

students, we conceived of *health* very broadly to obtain a wide snapshot of student health that could be used to identify areas of concern for future study. Following a review of the student-health literature, we developed a survey questionnaire that covered an extensive set of topics around physical and mental health, specifically tailored to undergraduate and graduate student life.

Cognitive interview design and analysis. Despite limited resources, we conducted cognitive interviews as a crucial step in questionnaire development to better understand the survey respondents' question-answer process. The interviewer, an advanced graduate student in psychometrics, had previous experience conducting cognitive interviews for another project, which formed the basis for the health interview protocol. The cognitive interview model consisted of a hybrid of open-ended thinking aloud, followed by unscripted probes for particular items based on interviewee responses. In addition to audio-taping the cognitive interviews, a graduate student in evaluation observed the each interview while taking field notes. A small number of cognitive interviews were conducted—two rounds of three interviews each, including one undergraduate male (two interviews), one graduate male, and three graduate female students, one of whom was an international student.

Results were analyzed to determine where participants explicitly voiced confusion and where participants expressed evidence of an understanding of the item that differed from the researchers' intended meaning. Due to the length of the questionnaire, each participant could think-aloud for only about a third of the items in a one-hour session. Consequently, most items were tested only once in each round due to the limited budget. Since we received data for most items from only one participant in each round, we drew on a variety of resources—the cognitive interview findings, researcher opinion (based on the student-health literature), and consultation

with expert health educators (expert review) to judge whether responses were idiosyncratic or likely represented difficulties many respondents might confront. Approximately 35 items were revised after both rounds of cognitive interviews. Most of the revisions were minor changes in wording to clarify the intention of the item, provide examples, or simplify syntax. While most of these revisions were straightforward, example 3 below illustrates the kinds of item context effects that can be difficult to detect in expert reviews but are revealed in cognitive interviews.

Example 3: Detecting context effects, 'carryover in what is retrieved'. Sometimes a respondent's ability to retrieve information relevant to an item is limited not by the item itself, but by the context in which the item is placed (McFarland, 1981; Todorov, 2000). Preceding items may prime the respondent's thinking around a particular subset of past experiences or knowledge, effectively blocking other potentially relevant information. Previous items can serve as an interpretive framework for subsequent items affecting what the respondents think the item is asking. These kinds of 'carryover effects' essentially entangle meaning from one item to 'carry over' to another item (Tourangeau & Rasinski, 1988). This occurred in the health assessment questions concerning nutritional attitudes and behaviors.

In the Round 1 draft, the item "How often are you aware of the nutritional content of the food you eat?" followed an item asking how often the subject ate out (Figure 5). We intended this item to apply to all of the respondent's food choices, and initially did not expect the item to be affected by the preceding item about eating at restaurants, particularly given the general wording ("the food you eat"). However, when the cognitive interview participant began thinking aloud for this item, it was immediately clear that her retrieval of relevant experience was restricted:

> Rarely. I don't think a lot of restaurants have [nutritional information], or I haven't read the menu too closely. I know that Applebee's sometimes will list calories.

The cognitive interviewer probed this response to determine if the participant was thinking only of food eaten in restaurants or also food eaten at home, and the participant was able to confirm quite explicitly the relationship between the two items:

> Well, actually, I was thinking of food at the restaurant, because I was following up on "How often do you eat out?"

[Insert Figure 5 here.]

Here, the restaurant item cued the respondent to think primarily of her experience eating out when responding about being aware of nutritional content, believing that since the items were next to one another, they must be related (not an unreasonable inference when taking a survey). To break the association, we simply reordered the items in this section, placing the item about nutritional content closer to other items concerning attitudes about food choice and moving the restaurant item to the end of the section. The Round 2 cognitive interview findings suggested that the respondent was consequently able to retrieve a wider range of experiences for this item.

Validity of questionnaire interpretations. As shown in example 3, even a modest number of cognitive interviews during questionnaire development can highlight unexpected interference in the response process. In this case, the range of experiences explored during the retrieval phase was limited due to cuing from the preceding item, which suggested that only eating in restaurants was relevant. Had this section stood as originally ordered, the item would have not provided information about respondents' overall nutritional awareness as was intended. Moreover, since it is likely that not all respondents would make this connection, the item would effectively conflate nutritional awareness when eating out with overall nutritional awareness, muddying the interpretation of results from this item. While the pilot statistics for this item might have indicated some kind of problem, identifying this type of measurement error would have

been difficult without the cognitive interviews. While there is no guarantee that other, undetected problems were not introduced, the cognitive interview data provide some evidence of improving questionnaire quality.

Conclusion

Cognitive interviews isolate "problems in the underlying cognitive processes through which respondents generate their answers to survey questions," (Tourangeau, 2003, p. 5). In the previous sections, we illustrated how cognitive interviews can be used to support the validity of questionnaire interpretations in a small- and medium-scale evaluations while recognizing limitations. Potential limitations include undiagnosed problems that could have been detected with additional cognitive interviews (Blair & Conrad, 2011). While not a panacea, there are particular types of problems with surveys (e.g. carryover effects) that are challenging to detect without cognitive interviews—suggesting a singular benefit to doing these interviews (Krosnick & Presser, 2010). Examples 1 and 3 illustrated how cognitive interviews can identify these types of measurement errors that are not likely to be revealed through other questionnaire testing methods (e.g., expert review). In spite of the value added by cognitive interviews, the potential advantages of this approach must be weighed against practical costs (time and material or personnel resources) and constraints, which will be especially challenging in the resource lean environment of evaluation. The restricted resources that characterize small- and medium-scale evaluations, as we showed in example 2, will constrain the power of cognitive interview designs.

What complementary methods are available and how might these methods be used to strengthen cognitive interview design and implementation? As a partial remedy, we propose a multi-method, multi-iterative approach to questionnaire development and refinement (Krosnick & Presser, 2010). Using cognitive interviews in conjunction with and as a complement to other

questionnaire testing methods and other information (e.g., expert review, survey and content literature findings) will provide multiple sources of evidence for discerning survey question quality that will be especially valuable in small- and medium-scale evaluation. Further, it is desirable to retest revised questions (multiple testing iterations) to assess whether the item repair was successful. To fully capitalize on the value of cognitive interviews while balancing scarce resources, evaluators will need to be judicious and carefully plan how, when, and why they are conducting cognitive interviews as part of a multi-method, multi-iteration testing of survey questions. While there will be on-going tensions and constraints about new problems that could be introduced and remain undetected, this kind of multi-method, multi-iterative approach to questionnaire development and refinement offers an important set of resources that can deployed to improve the credibility and quality of survey evidence in small- and medium-scale evaluations.

In addition, given an escalating population diversity and a trend towards increased attention to cross-cultural, multicultural, and multiracial contexts, the use of cognitive interviews in developing or adapting instruments for new populations, to better fit the range of respondent interpretations, is likely to increase (e.g., Irwin, et al., 2009; Willis & Zahnd, 2007).

References

Abt Associates. (2011). Retrieved from http://www.abtassociates.com/page.cfm?PageID=1468

Ackerman, A. & Blair, J. (2006, May). Efficient respondent selection for cognitive interviewing. Paper presented at the Annual Meeting of the American Association of Public Opinion Research, Hollywood, FL.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anderson, L. W. & Postlethaite, N. (2007). *Program evaluation :large and small studies.* Paris, France: International Institute for Educational Planning/UNESCO.

Beatty, P. C. & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), 287-311.

Blair, J. & Brick, P.D. (2009, May). Current practices in cognitive interviewing, Paper presented at the Annual Meeting of the American Association of Public Opinion Research, Hollywood, FL.

Blair, J. & Conrad, F.G. (2011). Sample size for cognitive interview pretesting, *Public Opin Quarterly 75*(4), 636-658.

Blair, J. & Presser, S. (1993). Survey procedures for conducting cognitive interviews to pretest questionnaires: A review of theory and practice. *Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association,* 370–75. Alexandria, VA: American Statistical Association.

Bradburn, N. M, (2004). Understanding the question-answer process. *Statistics Canada, 30*(1),

5-15.

Collins, D. (2007). Analysing and interpreting cognitive interview data: a qualitative approach. *Proceedings of the 6th Questionnaire Evaluation Standard for Testing Conference*. Ottawa, Statistics Canada

Conrad, F. G. & Blair, J. (2009). Sources of errors in cognitive interviews, *Public Opinion Quarterly, 73*(1), 32-55.

Czaja & Blair (2005) *Designing surveys: A guide to decisions and procedures*. CA: Pine Forge Press.

DeMaio, T. J., & Landreth, A. (2004). Cognitive interviews: Do different methods produce different results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al.(Eds.), *Methods for Testing and Evaluating Survey Questionnaire* (pp. 89-108). Hoboken, NJ: John Wiley and Sons.

Desimone, L. M. & le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Education Evaluation and Policy Analysis, 26*(1), 1-22.

Donaldson, S. I. (2008). In search of the blueprint for an evidence-based global society. In S. I. Donaldson, C.A., Christie, and M. M. Mark. (Eds.) *What counts as credible evidence in applied research and evaluation practice?* (pp. 2-18), Thousand Oaks, CA: Sage Publications.

Ericcson, K. A., & Simon, H. A. (1993). *A protocol analysis: Verbal reports as data, second edition*. Cambridge, MA: MIT Press.

Elmore, R. (2004). *School reform from the inside out: Policy, practice, and performance*. Cambridge, MA: Harvard Education Press.

Gottfredson, D.C. & Gottfredson, G.D. (2002). Quality of school-based prevention programs: Results from a national survey, *Journal of Research in Crime and Delinquency*. *39*(1), 3-35.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology (2nd edition)*. New York: Wiley.

Hamilton, L., B. Stecher, J. Marsh, J. McCombs, A. Robyn, J. Russell, S. Naftel, & H. Barney. (2007). Standards-based accountability under No Child Left Behind. Santa Monica, CA: Rand Education.

Hawley, W. D. & Rollie, D. L. (2007) *The keys to effective schools: Educational reform as a continuous improvement*. Thousand Oaks, CA: Corwin Press.

Howell, E. M. & Yemane, A. (2006). As assessment of evaluation designs, *American Journal of Evaluation, 27*(2), 219-236.

Irwin, D. E., Varni, J. M., Yeatts, K., & DeWalt, D. A. (2009). Cognitive interviewing methodology in the development of a pediatric item bank: a Patient Reported Outcomes Measurement Information System (PROMIS) study. *Health and Quality of Life Outcomes, 7:3*. (PMCID: PMC2642767)

Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of Survey Research* (2nd Edition) (pp. 263-313). West Yorkshire, England: Emerald Group.

Mayer, D. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1) 29-45

McFarland, Sam G. (1981). "Effects of question order on survey responses." *Public Opinion Quarterly,45*, 208–215.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning, *American Psychologist*, 50(9), 741-749.

No Child Left Behind Act of 2001, Pub L. No. 107-110, 115 Stat. 1435 (2002).

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in the first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis*, 30, 111-140.

Author, YYYY, Gandha, T., Culbertson, M., Lim, E., & Wakita, S. (2009). *IL Assessment consequences evaluation: Year one report* (Research Report No. 6*).* USA: University of Illinois at Champaign-Urbana, Illinois Assessment Consequences Evaluation.

Sabaratnam P, Klein JD. **(**2006). Measuring youth development outcomes for community program evaluation and quality improvement: findings from Dissemination of the Rochester Evaluation of Asset Development for Youth (READY) tool. *Journal of Public Health Management Practice, 12,* S88-94.

Schwarz, N. (2007). Cognitive aspects of survey methodology, *Applied Cognitive Psychology, 21*, 277-287.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Supovitz, J. A., & Taylor, B. S. (2005) Systemic education evaluation: Evaluating the impact of systemwide reform in education. *American Journal of Evaluation 26(2)*: 204-230, 2005.

Todorov, Alexander (2000). The accessibility and applicability of knowledge: Predicting context effects in national surveys. *Public Opinion Quarterly, 64*, 429–451.

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. In T.

Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.

Tourangeau, R. (2003). Cognitive aspects of survey measurement and mismeasurement. *International Journal of Public Opinion Research*, *15*, 3-7.

Tourangeau, R. & Bradburn, N. M. (2010). The psychology of survey response. Wright & P. V. Marsden (Eds.), *Handbook of Survey Research* (2nd edition) (pp. 315-346). West Yorkshire, England: Emerald Group.

Tourangeau, R. & Rasinsk, K. A. (1988). Cognitive processes underlying context effects in attitude measurement, *Psychological Bulletin, 103*(3), 299-314.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Tourangeau, R., Singer, E., & Presser, S. (2003). Context effects in attitude surveys, *Sociological Methods and Research, 31*(4), 486-513.

Trenholm C., B. Devaney, K. Fortson, L. Quay, J. Wheeler and M. Clark, Impacts of four Title V, section 510 abstinence education programs, Mathematica Policy Research, Princeton, NJ (2007).

Trochim, (2006). Construct Validity, Retrieved from http://www.socialresearchmethods.net/kb/constval.php.

U.S. Department of Education. (2009). *Elementary & secondary education: No child left behind*: *A desktop reference*. Retrieved from http://www2.ed.gov/admins/lead/account/nclbreference/index.html

Westat (2011). Retrieved from http://www.westat.com/westat/research_areas/health_and_medical_studies/health_and_medical_pe.cfm

Willis, G. B. (2004). Cognitive interviewing revisited: A useful technique, in theory? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al.(Eds.), *Methods for Testing and Evaluating Questionnaires* (pp. 23-44). Hoboken, NJ: John Wiley and Sons.

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

Willis, G. & Zahnd, E, (2007). Questionnaire design from a cross-cultural perspective: An empirical investigation of Koreans and Non-Koreans. J*ournal of Health Care for the Poor and Underserved 18*(4) 197-217.

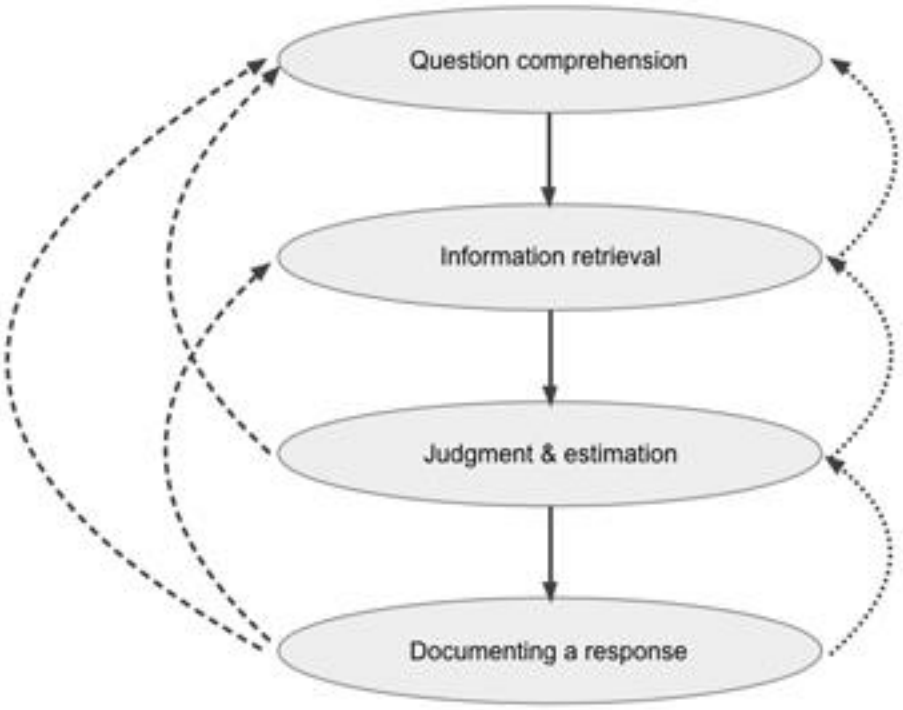

Figure 1. A four step model of cognitive processing in answering questions

Figure 2. Global Item Tested with Cognitive Interviews: Round 1

To what extent, and how, have NCLB testing changed your **instruction** since it began in all grades 3-8?	Changed it a lot, *for the better*	Changed it somewhat, *for the better*	Did not change it	Changed it somewhat, *for the worse*	Changed it a lot, *for the worse*
	5	4	3	2	1

Figure 3. Global Item Tested with Cognitive Interviews: Round 2

1. Has NCLB testing had *positive* effects on your instruction since it began in grades 3-8? *[Circle one number]*	**A great deal of positive effects**	**Some positive effects**	**Very little positive effects**	**No positive effects**
	4	3	2	1
2. Has NCLB testing had *negative* effects on your instruction since it began in grades 3-8? *[Circle one number]*	**A great deal of negative effects**	**Some negative effects**	**Very little negative effects**	**No negative effects**
	4	3	2	1

Figure 4. School Policy and Practice Item Tested with Cognitive Interview: Round 1

	My school's use of this strategy has…						**To what extent was it a result of NCLB testing?**			
	Increased a lot	**Increased somewhat**	**Not changed**	**Decreased somewhat**	**Decreased a lot**	***We don't use this strategy***	**To a great extent**	**To a moderate extent**	**To a small extent**	**Not at all**
Placement of students in pull-out programs.	5	4	3	2	1	❑	4	3	2	1

Figure 5. Nutrition Items Tested with Cognitive Interviews: Round 1

How often do you eat out?

Less than once per week	About once per week	2-3 times per week	4-6 times per week	About once per day	More than once per day
○	○	○	○	○	○

How often are you aware of the nutritional content of the food you eat?

Rarely	Sometimes	Often	Almost Always
○	○	○	○