ABSTRACT

Title of dissertation:	A COMPARISON OF EX-ANTE, LABORATORY, AND FIELD METHODS FOR EVALUATING SURVEY QUESTIONS
	Aaron Maitland, Doctor of Philosophy, 2014
Dissertation directed by:	Professor Stanley Presser Department of Sociology

A diverse range of evaluation methods is available for detecting measurement error in survey questions. Ex-ante question evaluation methods are relatively inexpensive, because they do not require data collection from survey respondents. Other methods require data collection from respondents either in the laboratory or in the field setting. Research has explored how effective some of these methods are at identifying problems with respect to one another. However, a weakness of most of these studies is that they do not compare the range of question evaluation methods that are currently available to researchers. The purpose of this dissertation is to understand how the methods researchers use to evaluate survey questions influence the conclusions they draw about the questions. In addition, the dissertation seeks to identify more effective ways to use the methods together. It consists of three studies. The first study examines the extent of agreement between ex-ante and laboratory methods in identifying problems and compares the methods in how well they predict differences between questions whose validity has been estimated in record-check studies. The second study evaluates the extent to which ex-ante and laboratory methods predict the performance of questions in the field as measured by indirect assessments of data quality such as behavior coding, response latency and item nonresponse. The third study evaluates the extent to which ex-

ante, laboratory, and field methods predict the reliability of answers to survey questions as measured by stability over time. The findings suggest (1) that a multiple method approach to question evaluation is the best strategy given differences in the ability to detect different types of problems between the methods and (2) how to combine methods more effectively in the future.

A COMPARISON OF EX-ANTE, LABORATORY, AND FIELD METHODS FOR EVALUATING SURVEY QUESTIONS

Aaron Maitland

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Stanley Presser, Chair
Professor Fred Conrad
Professor Frauke Kreuter
Professor Kent Norman
Dr. Roger Tourangeau

DEDICATION

This dissertation is dedicated to my mother, Kerry Maitland, whose life was too short, but whose influence and inspiration will last forever. I also dedicate this dissertation to my children Lisa and Jake so that they may understand that perseverance and finishing what you started should be valued in life.

ACKNOWLEDGEMENTS

There are many people that played important roles in completing this dissertation. Maybe the most important thing I learned during the writing of this dissertation is that no one does this alone. First and foremost, I thank Stanley Presser for serving as my advisor throughout the whole process. There is no one’s advice on survey methodological topics that I trust more than his. Dr. Presser has shaped this field in innumerable ways. In addition to showing incredible patience, he provided very insightful comments that vastly improved the dissertation along the way. I am incredibly thankful for having had the opportunity to work with someone of his achievement and intelligence. I am also grateful to Dr. Presser and the JPSM Survey Practicum students for providing much of the data that was used in the dissertation.

I am also thankful to the rest of my committee. They also share Stanley’s achievement and intellect. Roger Tourangeau has provided me with so many opportunities in life that I cannot possibly enumerate them in this space. I am fortunate to have had the opportunity to interact so closely with someone who has such a keen understanding of survey methodology. Dr. Tourangeau provided many helpful comments and endless encouragement throughout this whole process. I thank Frauke Kreuter for her advice and also for allowing me to use her question design course to collect much needed data for the dissertation. Fred Conrad also provided many helpful comments and encouragement throughout the process of working on this dissertation. I thank Kent Norman for serving as the Dean’s Representative on the committee.

There are other faculty, staff, and students at JPSM who made it possible to complete this dissertation. Katharine Abraham was a valued mentor throughout the JPSM

program. I relied heavily on my cohorts Jill Dever, Michael Lemay, Benmei Liu, and Santanu Pramanik at various times throughout the JPSM program. I will always value their friendship and advice. Andrew Mercer, both as a JPSM and Westat colleague, provided many suggestions for analysis. I am grateful for his friendship and advice.

I also want to thank Kristen Miller and Paul Beatty at the National Center for Health Statistics who shared many ideas and provided much needed support along the way towards completing this dissertation. I also want to thank Mike Brick and David Morganstein at Westat for providing me with the time and resources that I needed to finish the dissertation.

I also want to thank my family for all of the opportunities and support they have given me over the years. I hope to soon find a way to successfully communicate exactly what a degree in survey methodology is all about to them. I want to especially thank my father, John Maitland, for his dedication and hard work over the years that provided me with so many opportunities in life. Finally, I want to thank my wife, Noriko Maitland, for putting up with my nervous energy when I was spinning my wheels at various times during this dissertation. I am forever indebted to her for her unending care, understanding, and support throughout this whole process. I would never have completed this dissertation without her help.

CHAPTER 1: INTRODUCTION

Many research studies collect data through survey questionnaires. In order to enhance the validity of the findings from these studies, it is important for the studies to employ questions that minimize measurement error. A diverse range of question evaluation methods are available for detecting measurement error in survey questions. Ex-ante question evaluation methods are relatively inexpensive, because they do not require any data collection from actual survey respondents. Other methods require data collection from respondents either in the laboratory or in the field setting.

Some of the ex-ante methods have a long history, whereas others have been developed more recently. Two ex-ante methods that have been in use for a long time are expert review and forms appraisal. The more widely used of the two is expert review, which involves experts reviewing the questionnaire and critiquing the questions. Forms appraisals are somewhat less commonly used and consist of a checklist of problems for evaluating survey questions.

Two other ex-ante methods, both computer-based, are relatively new.   One is the Question Understanding Aid (QUAID) developed by Graesser and colleagues (Graesser et al., 2006). QUAID is a computer program that identifies problems with the wording, syntax, and semantics of survey questions. The program is designed primarily to detect linguistic features of questions that cause problems with question comprehension. The Survey Quality Predictor (SQP) introduced by Saris and colleagues (Saris and Gallhofer, 2007) is another relatively new ex-ante method. The SQP is a computer program that utilizes findings from dozens of multitrait-multimethod (MTMM) studies to predict the validity and reliability of a survey question.

Several other question evaluation methods require data collection from respondents. Laboratory methods such as cognitive interviewing are often used to collect verbal reports of cognitive processes from laboratory participants. Field based methods such as behavior coding are often used to identify questions that are difficult for the interviewer to administer or lead to inadequate answers in a standardized survey interview. Response latency measurements can also be used to identify problematic questions that require an inordinate amount of information processing by respondents.

This diverse mix of old and new methods confronts researchers with key decisions about how to adequately evaluate survey questions. Researchers could better package the methods together if they understood more about how effective the methods are at identifying flawed questions. However, there is currently a dearth of research on these issues. Some researchers have explored how effective some of these methods are at identifying problems with respect to one another (e.g., Presser and Blair, 1994; Yan, Kreuter, Tourangeau, 2012a). However, a weakness of many of these studies is that they do not compare a range of question evaluation methods that are currently available to researchers. Furthermore, the newer methods such as QUAID and SQP have not been compared to some of the more established methods. More importantly, a major problem in the literature is the general lack of evidence that the problems identified by these methods are actually problems as assessed by traditional quality standards such as reliability or validity. Although one would expect these methods to identify questions that produce low quality data, behavior coding is the only technique in the literature that has been shown to consistently predict the reliability and validity survey questions (Dykema, Lepkowski, and Blixt, 1997; Hess, Singer, and Bushery, 1999).

The purpose of this dissertation is to understand how the methods researchers use to evaluate survey questions influence the conclusions they draw about the questions. In addition, the dissertation seeks to identify more effective ways to use the methods together when evaluating survey questions.

A Model of the Question Development Process

Question evaluation methods are often targeted to one aspect of the question development process and thus it is useful to begin by outlining a framework for developing survey questions. This framework combines features from the existing literature explaining the question development process (e.g. Aday, 1989; Blair and Czaja, 2005; Converse and Presser, 1986; Esposito, 2004; Saris and Gallhofer 2007; Wilson 2005). Figure 1.1 illustrates the key components of the question development processes.

Figure 1.1. Key components of the question development process.

Process	**Actor**	**Methods**
Question design	**Researcher** Defines objectives Defines constructs Operationalizes constructs	Focus groups In depth interviews Expert review Expert system
Question administration	**Instrument** Defines task	Expert review Expert system
	Interviewer Delivers question Clarifies question Records response	Behavior coding Cognitive interviewing
	Respondent Comprehends question Recalls information Estimates answer Selects response	Behavior coding Cognitive interviewing Response latency
Data editing	**Researcher**	Consistency checks Quality assurance Imputation
Statistical analysis	**Researcher**	Data analysis Measurement error

The sample survey involves a communication between a researcher and a respondent. The first step in the process is for the researcher to define an overall set of survey objectives. These objectives will define the phenomena to be measured and require that the researcher communicate certain constructs to the respondents. The researcher will then need to operationalize the construct and formulate questions that enable her to understand how the respondents relate to these constructs. These questions become the central task in the communication between the researcher and respondent. In the case of a self-administered survey, the researcher has only the survey instrument itself to communicate with the respondents. In the case of an interviewer-administered survey, communication is delivered to the respondent through the survey instrument and the interviewer. In the next sections, I discuss the role of the researcher, survey instrument, interviewer and respondent in more detail.

Question Design: The Researcher's Role

Defining objectives. The first task for the researcher is to define the objectives of the survey. For example, according to the National Center for Health Statistics, the main objective of the National Health Interview Survey (NHIS) "is to monitor the health of the United States population through the collection and analysis of data on a broad range of health topics" (http://www.cdc.gov/nchs/nhis/about_nhis.htm). This indicates the phenomena that will be measured and also the population in which the measurement will occur. However, this objective is too broad to lead immediately to the development of survey questions so the researcher must undertake a process of conceptualizing the specific constructs that will be measured.

Conceptualization. Conceptualization is the process of defining the constructs that will be measured. The researcher would need to develop a clear definition of health for the purposes of the NHIS. For example, is the health construct limited to physical characteristics or does it also include mental, emotional, or spiritual characteristics? One must also decide the "range of health topics" that would be covered. A health survey like the NHIS might cover topics such as physical activity, physical limitations, or access to health care – just to name a few. Each of these topics would need to be defined to permit the identification of appropriate constructs.

There are at least two general approaches to the conceptualization process. Esposito (2004) describes how top-down and bottom-up processing are important in understanding social phenomena in the context of question construction. With respect to top-down or theory driven processing, subject matter experts are important. They may use existing literature or theory to help define a construct. Since there may be gaps in the existing theories, it is also important to consider bottom-up processing that involves observing how survey respondents understand the constructs. This can be accomplished with the use of qualitative techniques such as focus groups or in-depth interviews with members of the population of interest.

Operationalization. Once conceptualization is complete, the researcher would then be able to operationalize the constructs by creating survey questions. Analytical goals play a central role in the construction of survey questions (Aday, 1989). Fowler (1995) writes that, "…a question objective can be defined only within the context of an analysis plan, a clear view of how the information will be used to meet a set of overall research objectives" (p.11) At this stage, the researcher should be thinking explicitly about the type of data needed to meet the research objectives. For example, the level of measurement that the researcher desires will be important. It will be sufficient in some cases to obtain a count of the number of people who have a physical limitation. This may require a question like, "Are you limited in your ability to carry out physical activities?" In other cases, one may need to order respondents into low, medium, or high levels of limitation. This may require a question like, "How limited are you in your ability to carry out physical activities?" In still other cases, one might need to determine the extent to which some respondents are more or less limited. This requires the creation of a scale with a series of questions tapping the concept of physical limitation. The researcher then uses an implicit mathematical model to transform the responses to the questions into a summary score. One approach is to simply sum the responses to all questions, giving each response equal weight (e.g. Spector, 1992). Another approach is item response theory modeling where the goal is to construct items that tap different levels of a construct and each item has a certain level of difficulty attached to it (e.g. Wilson, 2005).

There are vast resources available to researchers constructing questions. Textbooks provide specific advice on how to construct questions (e.g. Converse and Presser, 1986; Fowler, 1995; Payne, 1951). These texts generally provide advice ranging from best practices in question design to lists of mistakes to avoid. For example, Fowler (1995) outlines five general principles with several subcomponents for designing good survey instruments, such as "ask one question at a time" and "a survey question should be worded so that every respondent is answering the same question."

In addition to question design texts, one could replicate questions from other surveys (Converse and Presser 1986). Such questions are likely to be useful to the extent that they were developed with objectives similar to the current objectives. Hence, one would need to examine the context in which the questions were asked. Ideally one would want supplemental information about how a question was interpreted. For example, the Q-Bank database of question evaluation reports provides access to the questionnaire and links survey questions to question evaluation findings so that researchers can assess how questions were interpreted during an evaluation of the question (http://www.cdc.gov/qbank).

Question administration

Eventually, the draft questionnaire will be administered directly to the survey respondents or indirectly to the respondents through an interviewer. Sudman and Bradburn (1974) "…conceptualize the interview as a microsocial system in which there are two roles, that of respondent and that of interviewer, joined by the common task of giving and obtaining information" (p.6). Understanding how this interaction unfolds can provide important insight into the quality of the resulting data and hence the components of the survey interview that need to be evaluated. This dissertation will focus on various methods that provide information about these different components. Before I discuss these methods and the information they generate, I will focus on each of these elements of the survey interview described by Sudman and Bradburn.

The Survey Instrument and Task Definition. There are numerous features of survey questions that might influence the quality of the survey data that they produce. In addition to the guidance provided by many of the question design texts mentioned in the previous section, there have been many literature reviews and empirical studies on the formal characteristics of survey questions (e.g. Krosnick and Presser, 2010; Alwin, 2007; Saris and Gallhofer, 2007). One approach taken by Schaeffer and Dykema (2011) is to group decisions made about survey questions into broad classes. They outline the following eight classes of decisions shown in Table 1.1.

Table 1.1. Decisions researchers make in designing survey questions.

Decision	Example
Question topic	What topic is studied?
Question type and response dimension	Is the question factual or subjective?
Conceptualization and operationalization of the target object	How do we turn concepts into questions?
Question structure	How do we group questions together?
Question form	Is the question open or closed?
Response categories	How many categories are used?
Question implementation	In what mode will the question be administered?
Question wording	How complex is the question?

A review of Table 1.1 reveals a fairly complex set of decisions that are made with every question that is drafted. It also is clear that many of these decisions are dependent on one another. For example, the question topic influences the question type. In addition, how questions are conceptualized and operationalized influences how they are structured. There are many other dependencies in Table 1.1. These dependencies are inevitable and make it difficult to disentangle the individual effects of any one question characteristic on data quality.

The existing literature on question characteristics is limited to a subset of the universe when one considers all of the complex interactions that might occur between question characteristics. Schaeffer and Dykema (2011) make the following judgment regarding the applicability of the existing body of research on question characteristics: "The usefulness of this research depends, ultimately, on the underlying analysis of the characteristics of questions, which characteristics are compared, and how the dependencies among question characteristics are taken into account in the study design." Furthermore, Willis (2005), discusses how question design rules by themselves are not specific enough, blind to the larger picture, and may fail to produce questions that address our information needs.

The existing literature on the formal features of survey questions and their effect on data quality is undoubtedly useful for developing initial drafts of survey questions. In some circumstances, it might even be all that is needed. However, for many questions, further evidence from question evaluation is needed to understand the quality of the information the questions yield.

The Interviewer. The interviewer also plays a role in determining the quality of the information yielded by questions. Within the standardized interview, the interviewer has the responsibility of delivering the survey question as worded. He or she also records the respondent's answer. The interviewer may also be required to clarify meaning or repeat part of the question. In addition, there are extra-role characteristics such as the interviewer's race, class or gender that may influence the interaction between the interviewer and the respondent (Sudman and Bradburn, 1974). Although these extra-role characteristics are important in some contexts, the existence of these effects does not indicate a fault with a question. Hence, the following discussion of the interviewer focuses on how the behavior of the interviewer may help diagnose problems with questions and data quality.

Standardized interviewers are trained to administer questions as worded. Fowler (1995) writes that one major goal of measurement "is to produce comparable information" (p. 2). Questions that can be consistently read as worded by the interviewer should help to reach this goal. In other words, reading the question as worded ensures that each respondent receives the same stimulus in the survey interview. The ideal standardized survey interview would consist of a simple stimulus-response or question-answer dialogue between the interviewer and respondent. However, there are times when the question-answer process breaks down and the interviewer performs an expanded role. This occurs when the respondent does not provide a response that adequately answers a question. For example, the respondent may provide a 'yes' response when the categories are approve or disapprove. In these cases, the interviewer would normally just repeat the response categories to the respondent to obtain an adequate response. Respondents may also ask to have part of the question repeated. More problematic, are instances when the respondent indicates that they do not understand a question or asks for clarification of a specific term. This can lead to a break down in standardization depending upon how the interviewers are trained to handle such situations and how skilled the interviewers are at their job.

In the strictest form of the standardized interview, the interviewers are trained to reply with “whatever it means to you.” In other cases, interviewers are allowed to clarify terms through question by question instructions or based on their understanding of the intent of the questions. There is ongoing research about how interviewers can clarify the meaning of survey questions and how this influences data quality (Schober and Conrad 1997; Conrad and Schober 2000).

The Respondent. The respondent plays the central role in the survey interview since it will ultimately be representations of his or her responses that constitute the final data. These responses will be shaped by the cognitive processes central to answering questions. Tourangeau (1984), building on earlier models, proposed a four-fold model of these processes.[1] The first process is comprehension, which refers to how respondents assign meaning to survey questions. The second process is retrieval, which includes how respondents recall information from memory. The third process is judgment, which includes how respondents combine or supplement information recalled from memory. The final process is reporting, which includes how the respondent communicates an answer.

[1] Jobe and Hermann (1996) review other cognitive models of the survey response process. These models generally break out sub-processes of the four major processes in the Tourangeau (1984) model. Some of the models also suggest that motivation is an underlying factor in the process that a respondent uses to answer a survey question.

Question Evaluation Criteria

Statistical estimates of measurement error (e.g. reliability and validity) should be of primary concern for the question design process, but they are expensive to obtain. In addition, the researcher has an interest in detecting and correcting any problems prior to the actual fielding of the survey instrument. Hence, obtaining knowledge about the components of the survey process shown in Figure 1.1 is useful in gaining an understanding of the data that is likely to result from a question.

In addition to the components shown in Figure 1, Fowler (1995) outlines a few characteristics of a good survey response process. First, the measurement process needs to be consistent. This means that questions need to be consistently understood and communicated to respondents. In addition, what constitutes an adequate answer should be consistently communicated. Second, unless measuring knowledge is the goal of the question, all respondents should have access to the information needed to answer the question accurately. Last, respondents must be willing to provide the answers to the question. Many of the methods that collect supplementary information about questions are attempts to assess these characteristics.

Question evaluation methods differ with respect to the type of problems that they identify. Hence, in order to evaluate the methods, it is important to determine the type of problems that the different methods detect. Ideally one would want to code the problems that each method identifies into one of the four processes from the Tourangeau (1984) model. Unfortunately the four processes in the Tourangeau model are too broad to work as a coding scheme. For example, there are several different types of comprehension problems that respondents might encounter. Presser and Blair (1994) created a coding scheme with four major categories: respondent semantic, respondent task, interviewer, and analysis problems. Respondent semantic problems occur when respondents have difficulty understanding a question, remembering the question, understanding the meaning of particular words or concepts in the question or when respondents have different understandings of what a question refers to. Respondent task problems referred to difficulty recalling, formulating, or reporting an answer. Interviewer problems refer to problems reading the question or recording the answer. Analysis problems occurred when the problem creates difficulties with data analysis (e.g. lack of variation in responses).

Methods for Question Evaluation

This dissertation examines six methods for question evaluation. This section summarizes the literature on each method with respect to the aims of the dissertation.

Expert Review. Expert questionnaire reviews vary in at least three ways. First, expert reviews can be conducted by questionnaire design experts, subject matter experts, or both. Subject matter experts are most helpful for establishing that a survey is collecting the information needed to meet the analytic objectives of a survey. In contrast, questionnaire design experts are more helpful for evaluating whether a question is likely to be problematic according to questionnaire design principles or because they may cause problems with the survey response process. Second, some expert reviews are conducted in a panel format, whereas others are conducted in an individual format. A chairperson of the panel summarizes the findings from the panel, whereas with the individual format the experts review the questionnaire and provide feedback independently. Third, expert reviews can be unstructured or structured. An unstructured expert review might simply ask the experts to indicate whether or not a question has a problem and then describe the problem. In contrast, more structured reviews might ask the experts to examine each question according to some predefined criteria and indicate the presence of specific problems.

The majority of the problems found by expert reviews are comprehension problems followed distantly by respondent task problems such as recall or response selection problems (Presser and Blair 1994; Rothgeb, Willis, and Forsyth 2001; Willis, Shechter, and Whitaker 1999). Expert reviews have also been found to detect question flaws that can lead to analytical problems (Presser and Blair 1994). There is some evidence that experts are generally able to identify problems with survey questions that lead to lower data quality (Olson, 2010). A common problem with expert reviews is that experts often disagree about the presence of a problem with a question. This low level of agreement has been primarily observed with individual expert reviews (Demaio and Landreth, 1993). This is probably due to the fact that many expert reviews are conducted in a fairly unstructured manner requiring the reviewer to do little more than describe the problem with a question. Expert panels or more structured review forms may improve the reliability of problem detection with experts.

Forms Appraisal. Forms appraisal methods are more structured than most expert reviews. The primary goal behind forms appraisal methods is to provide a systematic method for evaluating survey questions that can be employed by those who are less experienced with the principles of questionnaire design. A forms appraisal is conducted by evaluating each individual question for a specified set of problems. The Questionnaire Appraisal System or QAS (Willis and Lessler 1999) was developed to detect problems with the four cognitive processes described in the Tourangeau (1984) model. In addition, the QAS is designed to focus attention on those problems that are likely to affect accuracy (Lessler and Forsyth 1996). The QAS requires the evaluator to check each question for seven classes of problems involving question reading, question instructions, question clarity, assumptions, knowledge or memory, sensitivity or bias, and response categories. In total, each question is evaluated for the presence of 26 potential problems.

Although the majority of problems found by QAS are usually comprehension problems (Rothgeb, Willis, and Forsyth 2001; Forsyth, Lessler, and Hubbard 1992), one might expect that the systematic focus of the QAS on all four processes would make it a more effective tool at finding other types of problems. For example, there is some evidence that QAS finds more retrieval problems than other methods such as expert review (Rothgeb, Willis, and Forsyth, 2001).

QUAID. The Question Understanding Aid (QUAID)[2] is a computer tool that was developed by Graesser and colleagues (Graesser et al. 2006). QUAID is inspired by computational models developed in the fields of computer science, computational linguistics, discourse processing, and cognitive science. The software identifies technical features of questions that have the potential to cause question comprehension problems. The current version of QUAID critiques each survey question on five classes of comprehension problems: unfamiliar technical terms, vague or imprecise predicate or

[2] The QUAID tool can be found online at http://mnemosyne.csl.psyc.memphis.edu/QUAID/quaidindex.html.

relative terms, vague or imprecise noun phrases, complex syntax, and working memory overload.

QUAID generally identifies these problems by comparing the words in a question to several databases or data files (e.g., Coltheart's MRC Psycholinguistics Database). QUAID identifies a word as unfamiliar if it falls below a threshold level of frequency or familiarity metrics in several lexicons. Vague or imprecise predicate or relative terms (e.g. frequently) are identified by QUAID if their hypernym value is less than a threshold (i.e., the word is abstract), polysemy value is greater than a threshold (i.e., the word has multiple senses), concreteness value according to Coltheart's (1981) MRC Psycholinguistics Database is less than a threshold, or they are found in a list of vague terms. QUAID identifies complex syntax if the number of words before the main verb or main clause exceeds a threshold, the number of modifiers of a noun exceeds a threshold, or the average number of higher level constituents per word exceeds a threshold. Last, working memory overload is detected if the number of higher-level constituents per word exceeds a threshold, the number of conjunctions exceeds a threshold, or the number of words that signify logical operations exceeds a threshold. Expert ratings of a corpus of survey questions were critical in the development of QUAID. The corpus consisted of 505 questions on 11 surveys developed by the US Census Bureau. The threshold levels of the computer program were determined by identifying values that maximized the correlations with the expert ratings.

SQP. The Survey Quality Predictor (SQP) created by Saris and colleagues is based on a meta-analysis of Multi-Trait Multi-Method (MTMM) studies (Saris and Gallhofer 2007). The program uses the results from these studies to predict the quality of a survey question. The program outputs coefficients for reliability, validity, and method effects. It also computes a total quality indicator as the product of reliability and validity. In order to obtain these coefficients, the researcher codes each question according to the variables from the MTMM studies. The current version of the program requires the researcher to code the question according to approximately 50 variables ranging from fairly objective factors such as mode of administration and type of response options to more subjective factors such as degree of social desirability and how central the question is to the respondent.

Cognitive Interviewing. Cognitive interviewing is an umbrella term for a number of techniques conducted in a laboratory. Beatty and Willis (2007) propose that the most common application of cognitive interviewing involves, "administering draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that the author intends." The additional verbal information or verbal protocols are usually produced by either encouraging the interview subjects to think-aloud while answering the survey questions or by probing the subjects about their answers afterwards. Common examples of probes used during cognitive interviewing include "How did you come up with your answer?" or "What does [term] mean to you?" Other techniques such as card sorting, paraphrasing, and confidence ratings are also often included in formal definitions of cognitive interviewing. However, a study of academic and federal government research organizations by Blair and Presser (1993) concluded that probing and think-alouds were the most commonly used techniques. In fact, probing was utilized by the most organizations, think-alouds were used by a few organizations, and the other procedures were rarely used.

Even though cognitive interviewing is generally thought of as a product of the CASM movement, “follow-up” or “special” probes to understand respondents’ interpretation of survey questions had been used prior to CASM. Belson (1981) conducted probably the most widely known study using similar techniques. He had interviewers administer a questionnaire and complete a second “intensive interview” with the same respondents one day later. The intensive interview consisted of the interviewer reading the respondent the questions asked on the previous day and reminding the respondents of their answers. The interviewer would then ask follow-up probes such as, “When you were asked that question yesterday, exactly what did you think it meant?” The primary purpose of these follow-up questions was to understand the respondents’ interpretations of the survey questions. Belson (1981) found significant variation in how respondents interpreted many of the questions. Schuman (1966) used a technique called the “random” probe to evaluate whether respondents understood closed questions as intended. He had interviewers non-directively ask respondents to explain what they had in mind when answering the closed questions.

The think-aloud method for collecting verbal protocols is derived from a technique from cognitive psychology called protocol analysis, which was most famously utilized by Ericsson and Simon (1980) to study how people solve fairly complex problems such as puzzles or mathematical problems. According to the Ericsson and Simon (1980) framework, verbal protocols were valid to the extent that the reported information exists in short-term memory and the process of verbalization does not interfere with the task being reported. The authors drew distinctions between Type 1, Type 2, and Type 3 verbalizations. Type 1 verbalizations are direct reports of information as it is processed. Type 2 verbalizations are the result of information that has been recoded from a nonverbal format to verbal format. Type 3 verbalizations are those that require more explanation or are the result of the subject attending to information to which he or she would not normally attend. Ericsson and Simon conducted numerous experiments that provide evidence that Type 1 and Type 2 verbalizations produce valid protocols, but Type 3 verbalizations are more prone to reactivity. In other words, Type 3 verbalizations are more likely to be influenced by the nature of the task being performed.

Loftus (1984) was probably the first to apply think-alouds to the survey setting. She asked subjects to think-aloud while giving a response to a question such as, "In the last 12 months, how many times have you gone to a doctor, or a dentist, or a hospital, or utilized any health care specialist or facility?" Her main interest was in determining the order in which the subjects recalled autobiographical events. Her conclusion from the interviews was that "subjects tend to retrieve autobiographical memories in a predominantly past to present, or forward, direction" (p.64).

However, the applicability of the Ericsson and Simon (1980) framework to the survey interview has been questioned. Practitioners of cognitive interviewing are typically investigating problems with comprehension, retrieval, judgment, or response selection, whereas the Ericsson and Simon (1980) framework covers retrieval of events from short-term memory. Hence, as Willis (2004) indicates, it does not necessarily follow from the Ericsson and Simon framework that think-aloud methods will produce valid verbal protocols describing the mental processes that respondents use to answer survey questions. Willis (2004) highlights three important ways that the verbal protocols from a cognitive interview differ from those obtained during a typical task carried out by Ericsson and Simon's subjects. First, comprehension was not of concern for Ericsson and Simon. Their subjects performed tasks that were well defined and understood. However, survey cognitive interviews are commonly investigating how respondents assign meaning to a question. Second, the information retrieved more most survey questions relies heavily on semantic or long-term memory. In contrast, Ericsson and Simon were interested in how subjects process information in short-term memory. Last, survey cognitive interviews entail more social interaction than Ericsson and Simon advocated. Ericsson and Simon (1998) argue that social speech is likely to lead to Type 3 verbalizations. Hence, they provided instructions to respondents that limited or discouraged social interaction prior to their experiments. In contrast, the process of answering survey questions, at least in interviewer administered surveys, involves a social interaction between the respondent and interviewer (Schaeffer 1991).

Practical considerations have also led to a style of cognitive interviewing that in most situations diverges from the Ericsson and Simon framework. One frequently mentioned drawback of the think-aloud interview is that subjects vary in their ability to perform the task (Von Thurn and Moore 1994). In particular, less educated subjects tend to have more difficulty thinking out loud than those with higher levels of education (Bickart and Felcher 1996; Wellens 1994).

For the reasons stated above, most cognitive interviewers tend to rely on probing methods to elicit verbal information from subjects. Precise guidelines for probe development have yet to be derived; however, one recommendation is to develop probes that target the response processes that are most likely to be problematic for a question. Willis (2005) provides the following example of probes that target certain response processes. The question shown below (EX.1) includes technical terms which might cause difficulty with comprehension and lacks a reference period which might cause difficulty with retrieval.

> EX.1 Has anyone in your household ever received vocational rehabilitation services from:
>
> the State Vocational Rehabilitation Program?
>
> any other vocational rehabilitation program?

Willis (2005) recommends a probe such as “What, to you, is a “vocational rehabilitation program?” to address comprehension of the technical terms in the question and a probe such as “How sure are you that [person] got (or didn’t get) this type of service?” to determine the subject’s ability to retrieve the information with confidence.

Several other features are essential to an accurate description of cognitive interviewing. Cognitive interviews are usually performed in a lab setting where the environment is controlled and the interview is usually video or audio recorded. Purposive samples of interview subjects are recruited to ensure that relevant members of the target population are interviewed. For example, individuals with a range of physical disabilities should be recruited to test a questionnaire on the topic of physical disabilities. In practice, the typical cognitive interviewing project interviews approximately 10-30 subjects. However, recent research suggests that significant new problems can be uncovered even after 50 or more interviews (Blair and Conrad, 2011). The training of cognitive interviewers varies widely from interviewers who have advanced degrees in fields such as psychology, sociology, or survey methodology to survey staff or standardized interviewers who have been specially trained.

Behavior Coding. A goal of standardized interviewing is to produce comparable information across a random sample of the population (Fowler, 1995). A perfectly standardized question-answer sequence occurs when the interviewer reads the question as worded and the respondent provides an answer that meets the question's objectives.

However, there are times when this question-answer process does not go according to plan. For example, the interviewer might change the wording of the question or the respondent might reply with a request for clarification rather than an answer that meets question objectives. Although these behaviors are not always direct indicators of problems with questions, frequent deviations from the ideal question-answer process indicate the potential for problems with a question. In addition, Hess, Singer, and Bushery (1999) found that respondent behavior codes were significant predictors of the reliability of a question.

Several behavior coding schemes have been used to evaluate survey questions. Ongena and Dijkstra (2006) outline a few of the important decisions that must be made when developing an appropriate behavior coding system. First, the researcher must decide at which level the coding should be conducted. One option is to code every utterance in the interview. This is most useful when the sequence in which an utterance occurs is needed for analysis. This type of coding is often applied to the study of interviewer-respondent interaction. Another option is to code certain exchanges within the interview. For example, many behavior coding schemes for evaluating questions only code the initial exchange between the interviewer and respondent. Second, coding can be done live during the interview or from tapes. Third, coding can be completed by a variety of individuals. For example the coders can be the researchers themselves, field staff such as interviewers, or specially trained coders who do not have interviewing experience.

One of the most commonly used schemes is presented by Fowler and Cannell (1996). This scheme is a good example of what is often referred to as classical behavior coding, in which the initial exchange at a question is coded for several interviewer and respondent behaviors. Their scheme has three interviewer codes: interviewer reads the question exactly as worded, interviewer reads the question with minor changes, interviewer changes the question so that the meaning is altered. There are also seven respondent codes: respondent interrupts initial question reading, respondent requests clarification, respondents gives answer that meets question objective, respondent qualifies answer, respondent gives an answer that does not meet question objective, respondent gives a don't know answer, and respondent refuses to answer. One criticism of this scheme is that many of the behaviors are rare, because the standardized interview discourages overt expressions of problems. Schaeffer and Maynard (2002) find that behaviors such as hesitations, reports, and feedback are more effective indications of problems with questions than are behaviors such as explicit requests for clarification. Reports and feedback refer to instances when respondents provide some indication about certainty of their answer. For example, if Schaeffer and Maynard give the example of a respondent who is asked the question, "Do you have your own business or farm?" and the respondent reports that he owns the farm in partnership with his sister. This report may indicate that the respondent is unsure whether this qualifies as having his “own business or farm.” In short, there are many more elaborate behavior coding schemes available, but there is very little guidance on the advantages and disadvantages of the various schemes. Very little research has been done that compares and contrasts the use of different behavior coding schemes (see van der Zouwen and Smit 2004 for an exception).

An advantage of behavior coding is that it provides a quantitative estimate of the problematic nature of a question. The results from behavior coding can be analyzed in different ways that primarily depend on the nature of the scheme that was used. The most common type of analysis involves analyzing the frequency with which codes occur for the questions being evaluated. This type of analysis is typically performed on classical behavior coding schemes like Fowler and Cannell’s (1996) that code the initial interviewer-respondent exchange. An intuitively plausible percentage is decided upon to indicate questions that are candidates for revision. Other schemes that fully code all of the utterances in an interview tend to perform a sequence analysis on the resulting codes.

Response Latency. Another approach to diagnosing problems with a question is to measure how long it takes respondents to answer the question. There are two assumptions behind the use of response latencies as a question evaluation method. The first assumption is that response latency is an indicator of the amount of information processing required to answer a question. This includes the amount of time that it takes a respondent to comprehend a question, retrieve information from memory, integrate that information into a summary judgment, and select a response option. A second assumption is that problems with a question lead to slower response times, because resolution of the problem requires processing time (Basilli and Scott 1996). Like behavior coding, response latencies provide a quantitative assessment of the amount of difficulty that respondents are having with a question. However, the second assumption has been questioned because it is often difficult to tell whether longer response latencies are due to a problem with a question or to more careful processing of a question. More recently, researchers have been interested in shorter response latencies as indications of satisficing or shortcutting the response process with very short response times.

Response latency can be measured in different ways. One method is to have the interviewer press a button on the computer keyboard following the reading of the question and again when the response is given by the respondent. A variant of this approach is to have the interviewer press the button once and then have a voice activated key stop the timing when the respondent answers. A third method is to record the interviews and measure the latencies from the audio recording. One problem with response latency measurement is that certain utterances that the respondent makes before providing an answer can invalidate the measurement. For example, this can happen if a respondent requests clarification or produces speech disfluencies before answering. A fourth method is to use latent timers embedded within the software of computer-assisted telephone interviewing systems. These latent timers measure the amount of time from when a question was presented to when the respondent provides an answer. Hence in Web surveys, it is possible for the computer to measure the amount of time from when the question is presented until the respondent indicates an answer.

There is some debate in the literature about the type of timings that should be used for response latency. Some argue for more active approaches beginning at the end of the question and ending at the moment that the answer is given. Others argue for the latent approach that begins at the initial reading of the question and ends after the answer is completed (Mulligan et al., 2003). As Yan and Tourangeau (2008) indicate, each approach makes an assumption about when the survey response process starts. On the one hand, some argue that latent timings are contaminated with too many processes to be meaningful (Bassili, 2000). On the other hand, there is ample evidence from behavior coding studies that frequent interruptions of survey questions indicate processing does begin well before the end of a question (Draisma and Dijkstra, 2004). In spite of these differences, research has shown that similar conclusions can be drawn from either approach (Mulligan et al., 2003).

Statistical Models of Data Quality

Statistical estimation of measurement error has been influenced by theory from psychometrics and sampling statistics. Although these two perspectives have overlapping goals they often use different terminology. Psychometricians often refer to the validity and reliability of questions, whereas sampling statisticians refer to the bias and variability of questions. All of these notions of measurement error are based on the measurement model shown in Equation 1 where each individual response (y_{it}) is equal to a true value plus some error. In addition, each administration of a question is seen as one trial (t) within an infinite set of trials or administrations of the survey question.

Equation 1. Measurement error model.

$$y_{it} = \mu_i + \varepsilon_{it}$$

Validity refers to how well the answers to a survey question relate to some criterion or gold standard. Validity can be represented statistically by the correlation between responses to a survey question and some gold standard that is external to the survey responses. The validation of factual and attitudinal questions is somewhat different. The correlation between the responses to a factual survey question and administrative records, which serve as a gold standard, could be an estimate of the validity of the survey question. In contrast, there is no gold standard for attitude questions. However, one can estimate the validity of attitude questions by correlating the answers of one question with other answers to establish the construct validity of a question. Campbell and Fiske (1959) introduced the concepts of convergent and discriminant validity. In order to establish construct validity, responses to one attitude question should be correlated with the responses to another question measuring a related construct (convergent validity) and uncorrelated with the responses from a question measuring an unrelated construct (discriminant validity).

Validity is often confused with the related concept of bias from sampling statistics. Bias refers to the extent to which the mean or expected value of the survey responses averaged across a set of respondents differs from the expected value of the true values for the same set of respondents. Similar to validity, the measurement of bias also requires a gold standard external to the survey response. However, bias is different from validity because a question can elicit consistent underreports or overreports from a set of respondents and still be perfectly correlated with the respondents true values. Although, measurement of validity or bias may be the ultimate goal of any question evaluation process, the use of a gold standard often proves infeasible. This is particularly the case for subjective questions, but even administrative records used to validate factual questions can also suffer from inadequacies as the match process can introduce errors and the records themselves may be wrong.

Reliability, by contrast requires only two parallel measurements. The test-retest method is often used to assess the reliability of a question. According to this method, one assumes the measurement model in equation 1, where the variance of the observed values consists of some true score variance plus some error variance. This is often referred to as the "Classical True Score Model" from the psychometric literature (Lord and Novick 1968). Repeated interviews are conducted with the same respondents to evaluate the consistency of the responses at time 1 and time 2. The method assumes that the expected values of the responses are constant over time. This implies that there are no changes in the underlying construct and that the essential survey conditions are the same at both measurements.[3] Additionally, one must assume that the first measurement does not affect the second measurement. This implies that respondents do not remember their answer from the first interview and simply repeat it in the second interview. These assumptions together allow for the calculation of several measures of the consistency of responses over time. Psychometricians prefer to measure the positive side of reliability and define it as the ratio of the variance of the true scores to the variance of the reported values. The true score variance is equal to the covariance between repeated measurements. Reliability is equal to the correlation between the responses over time, which is shown in equation 2.

[3] The essential survey conditions refer to characteristics of the interview such as question context, question wording, interviewing procedures, and mode of data collection (Groves 1989).

Equation 2. Formula for test-retest correlation coefficient.

$$\rho_{y_{i1},y_{i2}} = \frac{Cov\left(y_{i1}, y_{i2}\right)}{\sqrt{Var\left(y_{i1}\right)*Var\left(y_{i2}\right)}}$$

As indicated by Groves et al. (2004), sampling statisticians focus on the negative side of reliability and refer to statistics such as the Index of Inconsistency (*IOI*). *IOI* is equal to (1 - ρ).

Equation 3. Formula for the Index of Inconsistency (*IOI*)

$$\text{IOI} = \frac{\frac{1}{2n}\sum_{i}\left(y_{i1} - y_{i2}\right)^{2}}{\sigma_{y}^{2}}$$

Although the formulations above refer to continuous outcomes, dichotomous formulations of reliability and the index of inconsistency have also been derived. Hess, Singer, and Bushery (1999) show that the index of inconsistency is equal to 1-kappa. Kappa is a commonly used measure chance corrected agreement.

Another commonly used measure of unreliability is the Gross Discrepancy Rate (GDR). Figure 1.2 shows an interview-reinterview table to illustrate how to calculate the GDR on a binary variable. The GDR is the proportion of individuals who answer differently on two occasions (O'Muircheartaigh, 1991). The columns of the figure illustrate how the respondent answered during the original interview and the rows of the table illustrate how the respondent answered during the reinterview. From figure 1.2, the GDR would be equal to (b + c) / n.

Figure 1.2. Interview-reinterview table.

	yj1=1	yj1=0	
yj2=1	a	b	a+b
yj2=0	c	d	c+d
	a+c	b+d	n

Reliability does not ensure validity, but is still a useful indicator of the quality of a survey question since a question must be reliable in order for it to be valid. In the case of subjective questions, reliability may be the most appropriate indicator of quality. The accuracy of the test-retest reliability coefficients will depend upon how well the assumptions of the model above hold. True changes in the characteristic being measured may lead to underestimates of reliability. Practice and memory effects can lead to overestimates of test-retest reliability. Research designs with respondents interviewed on three occasions allow the researcher to reduce memory effects by lengthening the time period between measurements and modeling the true change (Alwin, 2007). The researcher must attempt to balance these two concerns by carefully timing the reinterview when this is not possible.

Review of Method Evaluations

Table 1.2 lists published method evaluation studies that were found by searching the archives of the Journal of Official Statistics, articles in the electronic database JSTOR, edited survey methodological volumes on measurement error and questionnaire design, and the proceedings of the Survey Research Methods Section of the American Statistical Association. Additional studies were found in the bibliographies of these sources.

The evaluations used a variety of research designs. Forsyth, Rothgeb, and Willis (2004) distinguish three general approaches to method evaluation: exploratory, confirmatory, and reparatory. In addition, to these approaches there are also numerous examples of studies that describe the types of problems that were discovered by each method without an attempt to compare or evaluate the method. Hence, there are four general approaches to question evaluation highlighted in Table 1.2. Reparatory studies are really a special case of exploratory study, so I show three types of studies in the table. The study type column of the table illustrates the type of study that was conducted. Descriptive evaluations are denoted with the letter 'D' in column 3 of Table 1.2, exploratory evaluations with the letter 'E', and confirmatory evaluations with the letter 'C'. A quick glance at the table reveals that descriptive and exploratory studies are more common than confirmatory studies. The literature review to follow describes the relevant studies.

Table 1.2. Method evaluations in the existing literature.

Study	Methods Evaluated	Study Type
Bassili and Scott (1996)	BC, RL	E
Bischoping (1989)	BC, CP, ID	D
Blair et.al. (2007)	CI, BC	C
Campanelli, Martin, and Rothgeb (1991)	ID, RD	D
Draisma and Dijkstra (2004)	BC, RL	C
Dykema, Lepkowski, and Blixt (1997)	BC, RC	C
Eisenhower (1994)	FG, CI	D
Esposito et al. (1991); Esposito and Rothgeb (1997)	BC, ID, RD, RDA	D
Forsyth, Lessler, and Hubbard (1992)	CI, FA	D
Forsyth, Rothgeb, Willis (2004)	CI, ER, FA, ID, BC, IN	C
Fowler and Roman (1992)	FG, CI, BC, ID, RD	D
Graesser et al. (2000)	ER, QUAID	E
Graesser et al. (2006)	ER, ET, QUAID	E
Hess, Singer, and Bushery (1999)	TR, BC	D
Hughes (2004)	BC, CI, RD	D
Hunt, Sparkman, and Wilcox (1982)	RD	E
Lessler, Tourangeau, and Salter (1989)	CI, CP	D
Miller (2002)	CI, FG	D
Oksenberg, Cannell, and Kalton (1991)	BC, RD	D
Presser and Blair (1994)	BC, CI, ER, ID	E
Rothgeb, Willis, and Forsyth (2001)	CI, ER, FA	E
Stapleton Kudela et al. (2006)	BC, CI	D
Sykes and Morton-Williams (1987)	BC, RD	D
van der Zouwen, Saris, Draisma, and van der Veld (2001)	BC, CS, ER, FA, SQP	E
van der Zouwen and Smit (2004)	BC, ER, FA, SQP	E
van der Zouwen and Dijkstra (2002)	BC, ER, FA	E
Willis (1991)	BC, ID, OD	D
Willis and Schechter (1997)	CI, FE	C
Willis, Schechter, and Whitaker (1999)	BC, CI, ER	E
Yan, Kreuter, and Tourangeau (2012)	ER, CI, SQP, LCM, TR, V	E

Abbreviations: Behavior Coding (BC), Cognitive Interviewing (CI), Conventional Pretest (CP), Computer Simulation (CS), Expert Review (ER), Eye Tracking (ET), Forms Appraisal (FA), Field Experiment (FE), Focus Group (FG), Interviewer Debriefing (ID), Item Nonresponse (IN), Latent Class Models (LCM), Observer Debriefing (OD), Question Understanding Aid (QUAID), Response Distribution Analysis (RDA), Record Check (RC), Respondent Debriefing (RD), Response Latency (RL), Survey Quality Predictor (SQP), Test-Retest (TR), Validity (V)

There are some important differences between the approaches to question evaluation. In general, descriptive studies demonstrating the use of different question evaluation methods in a question development process provide an overview of the contribution of each method to the overall process. These studies provide an important contribution by describing current practices in the field; however, they are of little value in helping to determine the relative effectiveness of each method since there is no empirical comparison of the methods.

Empirical method evaluations begin with what Forsyth, Rothgeb, and Willis (2004) label as exploratory studies. These studies are typically designed to compare methods using metrics such as the number and types of problems that the methods detect. Agreement or correlational statistics are then used to measure the extent to which the methods agree or disagree on individual problems or overall conclusions about questions. A challenge with this approach is that it often involves comparing methods that produce very different types of results from very different environments. For example, some methods produce qualitative results with rich descriptions of problems, whereas other methods produce purely quantitative results. Researchers have dealt with these differences in different ways, but it is unclear what approach, if any, is the best one. In addition, the literature has done very little to foster an understanding of the circumstances under which the methods should agree either in theory or in practice. For example, the methods may be more likely to agree on certain types of problems or on problems with certain types of questions. As will be described later, the literature using these types of studies has provided very mixed results. The best advice from these studies is that the methods are complementary and should be used in combination (Yan, Kreuter, and

Tourangeau, 2012; Presser et al., 2004). However, the literature has not been successful at determining how to package the methods together most effectively.

The remaining – confirmatory – approach has the potential to provide greater clarity regarding which problems and methods are the most likely to influence data quality and should be given the most weight by researchers. The confirmatory approach focuses on using the results from one or more methods to predict the quality of questions in the field. This approach asserts a model of the question evaluation process that gives priority to methods that assess data quality in a realistic field setting. Ideally, the researcher would prefer to use direct assessments of reliability and validity. This would be done using a reinterview design or obtaining record checks for a set of survey questions. These methods are often too expensive or impractical to implement. Hence, researchers often rely on indirect measures of data quality collected in the field such as item nonresponse, behavior coding results, response timings, or field experiment predictions. Indeed, researchers have shown some links between method results and data quality in the field (e.g. Hess, Singer, and Bushery, 1999; Forsyth, Rothgeb, and Willis, 2004). However, a major weakness of the existing literature is that the studies tend to evaluate one method at a time. This gap in the literature prevents researchers from understanding the relative effectiveness of different methods. In addition, this gap prevents researchers from understanding how to package methods together.

Exploratory Research

Presser and Blair (1994) conducted a study that compared behavior coding, cognitive interviewing, expert review, and conventional pretesting. They tested five supplements from the National Health Interview Survey on a variety of topics including food knowledge, dietary behavior, medical care, general health knowledge, and knowledge about AIDS. Each question evaluation method was replicated as part of the study in order to measure the reliability of the methods. First, the results of the conventional pretests consisted of two sets of interviewer debriefings. Teams of four interviewers with previous question evaluation experience completed roughly 40 undeclared telephone pretests. Each team of interviewers then reviewed their overall and question by question experience with a senior interviewer. Second, observers behavior coded the interviewer-respondent behavior interaction in real time during the pretest interviews. The interviewer-respondent interactions were coded for major changes in reading the question, interviewer probing, respondent requests for clarification or other difficulty, and uncodable answers. Third, three sets of 10-12 face to face cognitive interviews were conducted by interviewers who had previous experience with cognitive interviewing in questionnaire development. The cognitive interviews consisted of a combination of follow-up probes and concurrent and retrospective think-alouds. Last, expert panels consisting of one psychologist, one specialist in questionnaire design, and one general survey methodologist were asked to review the questionnaire in a tape recorded 2-3 hour panel discussion.

Summary reports of the problems identified by each method in the Presser and Blair study were coded by type: respondent semantic, respondent task, respondent behavior, interviewer, and analysis problems. Respondent semantic problems occurred when a problem summary indicated that respondents had (or would have) difficulty understanding a question, remembering the question, understanding the meaning of particular words or concepts in the question or when respondents had (or would have) different understandings of what a question refers to. Respondent behavior problems referred to the behaviors recorded during the behavior coding. Respondent task problems referred to difficulty recalling, formulating, or reporting an answer. Interviewer problems referred to problems reading the question or recording the answer. Analysis problems occurred when the problem statements anticipated problems with data analysis.

Presser and Blair evaluated the four methods by analyzing the number of problems identified, type of problems identified, consistency of problems identified between trials of the same method, consistency of problems identified between different methods, and cost of each method. Overall, averaging across trials of the different methods, expert reviews identified the most problems. On average, expert reviews uncovered almost twice as many problems as the other methods. However, there was significant variability in the number of problems identified between trials of the conventional pretests and cognitive interviews. There was very little variation in the number of problems identified between trials of expert reviews and behavior coding.

The methods also differed in the type of problems that were detected. Conventional pretesting and behavior coding were the only methods to detect a substantial number of interviewer problems. In general, respondent semantic problems were the most prevalent problem found by conventional pretests, cognitive interviews, and expert review. These methods also detected, to a somewhat lesser extent, respondent task problems. Cognitive interviews and expert reviews detected the highest number of analysis problems. The distribution of the types of problems detected varied between trials of the conventional pretests and cognitive interviews.   In comparison, there was very little variation in distribution of the types of problems detected between trials of the expert reviews and behavior coding. The authors also assessed the extent to which the methods detected the same problems both within trials of the same question evaluation method and between methods. Unexpectedly, the between-method correlations as measured by Yule's Q were not much lower than the within-method correlations for both conventional pretests and expert reviews. Behavior coding was by far the most reliable method between trials followed distantly by cognitive interviews and expert review. Finally, Presser and Blair (1994) evaluated the cost of each method after analyzing the number and types of problems detected. The authors found that conventional pretesting and behavior coding were the most expensive and similar in cost. Cognitive interviews cost roughly 20 percent less and expert panels roughly 50 percent less.

A similar study was conducted by Willis, Schechter, and Whitaker (1999) that compared cognitive interviewing, behavior coding, and expert review. Their study design was similar to Presser and Blair's (1994) with some notable exceptions. First, the authors included cognitive interviews from two different survey organizations. One set of interviewers from the National Opinion Research Center (NORC) were specially trained for the study and the other set of interviewers from the National Center for Health Statistics (NCHS) were more experienced cognitive interviewers. This difference in experience between cognitive interviewers at each organization led to rather different implementations of cognitive interviewing. For example, the NCHS interviewers developed their own probes and used them in either a scripted or spontaneous fashion, whereas the NORC interviewers were instructed to use a set of scripted probes. A second difference from Presser and Blair is that larger sample sizes were used for cognitive interviewing and behavior coding. Third, all questions were evaluated at the individual level. For example, in the Presser and Blair (1994) study, expert reviews were conducted by panels of experts, whereas in the Willis, Schechter, and Whitaker (1999) study experts individually reviewed the questionnaire. This individual level analysis was facilitated with a problem box next to each question in the questionnaire. Cognitive interviewers checked a box if a problem was observed during the interview or if the interviewer noted that a problem might exist and expert reviewers checked a box if the reviewer thought that some problem might exist. Space was also provided for comments next to each question.

Willis, Schechter, and Whitaker reported the number of problems identified by each question evaluation method, the correlation between and within methods in terms of the number of problems identified, and the type of problems identified. Like Presser and Blair (1994), they found that expert reviews detected the most problems. Behavior coding detected the second most problems and cognitive interviewing the third most problems. Next, the authors examined the extent to which the methods agreed on the number of problems. The correlation between trials of behavior coding was found to be the highest (.79). The correlation between trials of cognitive interviewing were somewhat lower, but was still quite high (.68). Importantly, these within method correlations were higher than the correlations between methods. The authors hypothesized that the correlations between methods would vary according to a continuum of objectivity. The authors ordered the methods from most objective to most subjective. The most objective method was behavior coding followed by NORC cognitive interviewing, NCHS cognitive interviewing, and expert review. Contrary to hypothesis, although the correlation between the behavior coding trials and expert review was relatively low (~.54), it was not the lowest correlation. However, in support of the hypothesis the authors did find that the correlation between the NCHS cognitive interviews and expert review was higher than the correlation between the NORC cognitive interviews and expert review. In addition the correlation between the NORC cognitive interviews and behavior coding trials was higher than the correlation between the NCHS cognitive interviews and the behavior coding trials.

Last, Willis, Schechter, and Whitaker (1999) analyzed the distribution of the types of problems detected by cognitive interviewing and expert reviews. The authors used a different coding scheme from Presser and Blair (1994) so the results from the two studies are not directly comparable. Their coding consisted of five general problem types: comprehension / communication problems (including administration problems for the interviewer, problems with question length, problems with specific terms, problems with question difficulty, and problems related to question vagueness), recall-based problems, bias/sensitivity problems, response category problems, and logical/structural problems. The authors coded the comments that described each problem that was encountered. The results are shown in Table 1.3. They concluded that the overwhelming majority of the problems encountered by cognitive interviewing and expert review were communication or comprehension problems followed by response category problems, recall problems, logical problems, and sensitivity problems.[4] This pattern held for both cognitive interviewing and expert review.

[4] The authors did not include behavior coding in this part of the analysis.

Table 1.3. Summary of qualitative nature of problems found in Willis, Schechter, and Whitaker (1999).

Problem category	NCHS CI	NORC CI	ER
		Percent	
Comprehension / communication	70.5	58.1	75.1
Recall	11.0	13.3	7.8
Bias	1.9	1.3	3.3
Response categories	12.1	19.8	9.1
Logical	4.5	7.5	4.7
	(n=471 problems)	(n=626 problems)	(n=551 problems)

The authors did not comment on some differences in the distribution of problems across methods. The results also demonstrated that the cognitive interviews, particularly those done by the less experienced NORC interviewers, found higher proportions of recall and response category problems than the expert review. The authors did not conduct the analysis required to support this finding, so I have conducted the analysis on my own with the data from Table 2 using Chi-squared statistics to compare the distributions of the three methods. Overall, there is a significant difference between the distribution of the problems across all three methods ($\chi_8^2 = 64.77, p < .0001$). The distribution of the NORC cognitive interviews differs from both the NCHS cognitive interviews ($\chi_4^2 = 25.88, p < .0001$) and the expert review ($\chi_4^2 = 56.64, p < .0001$). The difference in the distribution of the problems found by the NCHS cognitive interviews and expert review approaches significance (X_4^2= 7.64, p = .10). Both trials of cognitive interviewing found lower proportions of comprehension/communication problems and higher proportions of recall and response category problems than were found in the expert review. Furthermore, the relatively inexperienced NORC interviewers accentuated this pattern.

Rothgeb, Willis, and Forsyth (2001) undertook a study that compared cognitive interviewing, expert review, and forms appraisal. They examined the effect of question evaluation method, survey organization, and questionnaire topic in a Latin Square experimental design. The design included the three question evaluation methods, three researchers each from three different survey organizations, and three questionnaire topics. All three researchers within each organization eventually conducted all three question evaluation methods. However, each researcher evaluated each individual questionnaire only once. The expert reviews were conducted on an individual level and the researchers checked a problem indicator box and summarized the problem when one was encountered. The forms appraisal consisted of the Questionnaire Appraisal System (QAS), developed by the Research Triangle Institute. The researchers used the QAS to evaluate each question for 26 potential problems. Last, each researcher conducted three cognitive interviews on their assigned questionnaire. Similar to the expert reviews, the researchers were asked to check a problem box and summarize the problem when one was encountered.

Rothgeb, Willis, and Forsyth (2001) conducted two types of analyses. First, they compared the number of times that a problem was found across organizations and question evaluation methods. Averaging across organizations, the QAS found the most problems. In fact, the QAS found a problem with nearly every item, whereas cognitive interviewing and expert review found problems for nearly half the items. The authors cautioned that the extraordinary sensitivity of the QAS might come at the expense of the low specificity of the method. Averaging across question evaluation methods, there were no significant differences in the number of times that organizations found an item to be problematic. Although the organizations found similar numbers of problems, there was only a moderate level of agreement with respect to which specific questions were problematic according to a wide range of correlation statistics.

Second, the authors compared the qualitative type of problems that were found by each method and across organizations. Their coding scheme consisted of comprehension and communication, memory retrieval, judgment, evaluation, and response selection problems. Overall, the authors concluded that there were no significant differences in the type of the problems detected by different methods or organizations. The vast majority of problems found by all three methods were comprehension / communication problems. However, as shown in Table 1.4, there was a tendency for forms appraisal and cognitive interviews to detect a higher proportion of retrieval problems and somewhat lower level of comprehension / communication problems than expert reviews. Relatively small cell sizes make it difficult to draw any strong conclusions about this finding. However, the differences between cognitive interviewing and expert review are similar to those in Willis, Schechter, and Whitaker (1999).

Table 1.4. Summary of qualitative nature of problems found in Rothgeb, Willis, and Forsyth (2001).

Problem category	ER	QAS	CI
		Percent	
Comprehension / Communication	79.3	66.3	69.7
Retrieval from memory	9.4	18.9	21.2
Judgment and evaluation / response selection	11.3	14.9	9.1
	(n=53)	(n=175)	(n=66)

Bassili and Scott (1996) compared response latency with behavior coding. They completed a telephone survey of 200 students that asked questions that contained superfluous negatives, were double barreled, or had previously been shown to elicit high percentages of problem behaviors. Alternative versions of the questionnaire contained either damaged or repaired versions of the questions. Response latency was measured for each question by having the interviewer press a key on a computer keyboard at the end of reading the question and again when the respondent began to answer the question. Three behavior codes measured requests to have the question repeated, requests for clarification, and in the case of double barreled questions, whether or not the respondent asked which aspect of the question should be answered. These codes were included in the study because they refer to behaviors that interrupt the ideal question-answer sequence and thus call into question the response latency measure. The authors found that questions with superfluous negatives took longer to answer and elicited more requests for clarification or repetition than those that did not contain superfluous negatives. Similar results were obtained for double barreled questions, which took longer to answer and elicited more requests for clarification or repetition. However, the response latency and behavior coding did not agree on other types of problems. For example, the authors examined four questions that were shown by past research to exhibit a high degree of problematic behaviors. The repaired versions of these questions actually took longer to answer than the original questions even though the original versions were again associated with more problematic behaviors than the repaired questions. This finding casts some doubt on the assumption that longer response latencies are indicators of problems with questions.

A series of studies by van der Zouwen and colleagues evaluated how interaction analysis, a particular form of behavior coding, compares with other techniques. The goal of interaction analysis is to analyze question-answer interactions between interviewers and respondents and then code the interactions for how adequately the interactions fulfill the goals of standardization. Interaction analysis leads to a description of every question answer sequence as being paradigmatic, problematic, or inadequate. Paradigmatic sequences occur when the question-answer sequence does not differ from the ideal standardized interview. Problematic sequences occur when the sequence deviates from standardization, but it is repaired. Inadequate sequences occur when the sequence deviates from standardization and is not repaired. Van der Zouwen and Dijkstra (2002) compared interaction analysis with expert review and a forms appraisal method using 37 questions from a questionnaire about advertising. For the expert review, they asked ten experienced survey researchers to evaluate the quality of the 37 questions on a scale from 1 (excellent question) to 7 (worthless question). The authors also assigned a task difficulty score (TDS) to each question by evaluating each question according to twelve criteria about the difficulty of the question and clarity of the task. They found that at least one of the three methods found a problem with 14 of the 37 questions. Generally, the TDS and expert reviews agreed on what questions were problematic. For example, the TDS identified five problematic questions and the expert reviews identified six problematic questions. Four of these questions identified as problematic by either method were found to be problematic by both. In contrast, the interaction analysis found seven problematic questions, but none of them were identified as problematic by the TDS or expert review. The interaction analysis mainly identified interviewer problems such as directive or inadequate probing, whereas the TDS and expert review found problems with questions that required retrieval of detailed information.

Van der Zouwen et al. (2001) expanded on the work by van der Zouwen and Dijkstra (2002) by adding a comparison with SQP.[5] SQP identified six of the questions as problematic. The three ex ante methods (expert review, TDS, and SQP) agreed on four questions being problematic. However, interaction analysis found as problematic seven completely different questions than the ex-ante methods.

A study by van der Zouwen and Smit (2004) compared interaction analysis with classical behavior coding, expert review, forms appraisal (QAS and TDS), and the survey quality predictor (SQP). Eight income questions were pretested for this study. Four questions were found to be problematic by interaction analysis and classical behavior coding. The two methods agreed on three out of the four questions being problematic. The Spearman rank-order correlation was .95 for these two methods when all eight questions were rank ordered according to how problematic they were. The expert review, QAS, SQP, and TDS had much lower levels of agreement with the interaction analysis. The rank order correlations between interaction analysis and these other methods were all less than .2. The two forms appraisal methods (QAS and TDS) agreed that the same five questions were problematic and had a rank order correlation of .91. Four of the five questions identified as problematic by the forms appraisal methods were also identified as problematic by the expert review. The rank order correlation between the expert review and each of the forms appraisal methods was approximately .75. The SQP did not correspond with any of the other methods. In fact, the authors found negative rank order correlations between the SQP and all of the other methods.

[5] van der Zouwen et al. (2001) analyzed the same data for expert review, TDS, and interaction analysis as van der Zouwen and Dijkstra (2002).

Graesser and colleagues have undertaken a series of studies in an attempt to establish the validity of the Question Understanding Aid (QUAID). Graesser et al. (2000) suggest that an advantage of QUAID is its ability to identify problems that would be missed by respondents during pretesting. Graesser et al. (2006) assert, "There are inherent limitations in methodologies that exclusively use focus groups and one on one interviews with samples of respondents during pretesting, at least with respect to dissecting linguistic problems with question interpretation (p. 14)." According to the authors, pretest respondents are only able to reliably identify unfamiliar technical terms and vague or ambiguous noun phrases. Graesser et al. (2000) also find the judgments of experts to be problematic, because they can also be unreliable at assigning judgments to questions as either problematic or not problematic. However, they do feel experts are in a better position to judge the adequacy of a question.

Therefore, in order to validate QUAID, the authors compared the results from QUAID to an expert review. That is, three experts who were extensively trained in questionnaire design and had graduate training in linguistics, discourse or cognition evaluated several questions. Overall, 550 questions provided by the United States Census Bureau were evaluated by the experts for six problems that are identified by QUAID.[6] The authors used techniques from signal detection theory to evaluate the effectiveness of QUAID at diagnosing problems that were previously identified by experts. For this type of analysis, one essentially uses the expert review of a question as the "Gold Standard." QUAID is viewed positively if it identifies problems that were identified by experts (hits)

[6] The six problems identified by QUAID are unfamiliar technical terms, vague or imprecise relative terms, vague or ambiguous noun phrases, complex syntax, working memory overload, and misleading presuppositions. Misleading presuppositions were later dropped from QUAID.

and negatively if it identifies problems that were not identified by experts (false alarms).[7] Two findings emerged from the study. First, the hit rates for QUAID, assuming the experts were correct about the problems identified, were high (>.85) with respect to unfamiliar technical terms, vague or imprecise relative terms, and vague or ambiguous noun-phrases. However, this promising finding is tempered by the high false alarm rate for these problems. That is, QUAID was very sensitive to these types of problems at the expense of lower specificity. QUAID was also very sensitive with respect to misleading presuppositions, though to a lesser extent. In contrast, the authors found very low hit rates for QUAID with the problems of complex syntax and working memory overload. Second, the authors analyzed d prime scores: measures of QUAID's ability to discriminate signal (hits) from noise (false alarms). The d` scores are essentially the normalized ratio of the proportion of hits to the proportion of false alarms. The analysis revealed positive d` scores that were significantly different from zero. This indicates that QUAID is able to distinguish problematic questions from non-problematic questions as identified by experts.

Graesser et al. (2006) conducted a similar study comparing QUAID and expert review. For this study, sixty of the most problematic questions from the Graesser et al. (2000) study were divided into three sets of questions. First, twelve expert survey methodologists critiqued the questions. The conditional probabilities that the experts identified unfamiliar technical terms, vague or imprecise predicates or relative terms, vague or ambiguous noun phrases, complex syntax, and working memory overload given that QUAID identified the problems were .10, .11, .46, .47, and .37. In other words,

[7] The hit rate is p(QUAID finds a problem | expert finds a problem) and the false alarm rate is p(QUAID finds a problem | expert finds no problem).

QUAID was more likely to identify these problems than the expert reviewers. The authors used this as support for the claim that QUAID identifies problems that are not typically found by other question evaluation methods. Next, the experts revised one set of questions with the use of QUAID and another set of questions without the use of QUAID. Another set of 12 expert survey methodologists were then shown two alternatives of a question and asked to choose which question would be easier to comprehend for most respondents. Each pairing was either a question revised by experts with the assistance of QUAID or experts alone versus an original question. The results were mixed. Preference scores from the experts revealed that both the questions revised by the experts alone and the experts using QUAID were preferred over the original questions.[8] However, there was not a significant difference in preference scores between those revised with QUAID and those revised by expert reviewers alone at the traditional .05 alpha level (though the questions revised with QUAID did have higher preference scores than those revised by the experts alone, which is in the correct direction if QUAID helps the experts revise questions).

Yan, Kreuter, and Tourangeau (2012) compared a combination of five quantitative and qualitative question evaluation methods. They compared expert reviews, cognitive interviews, quantitative measures of reliability and validity, and error rates from latent class models. They generally found low consistency across the methods in how they rank ordered the items in terms of quality. There was, however, considerable agreement between the expert ratings and the latent class method and between the cognitive interviews and the validity estimates. They concluded that the methods yield

[8] The preference score indicates the extent to which the choice of the revised question over the original was above .50 or random. It is computed as [(Revised p-.50)/(1-.50)]

different and sometimes contradictory conclusions with regard to the 15 items pretested. The findings raise the issue of whether results from different testing methods should agree. In their discussion, the authors put for the idea that agreement could be related to the nature of the problems that the methods detected. A post-hoc analysis revealed higher correlations between the proportion of experts and cognitive interviews that found recall problems compared to other types of problems like comprehension problems. However, there was no explanation for this pattern of correlations. The notion that agreement could be influenced by problem type provides some future direction to the literature to begin looking beyond simple agreement between methods though.

Overall, the main conclusion from Yan, Kreuter, and Tourangeau was consistent with what Presser et al. (2004) concluded regarding best practice in using question evaluation methods. In the authors' own words, "…until we have a clearer sense of which methods yield the most valid results, it will be unwise to rely on any one method for evaluating survey questions" (p. 523). The authors also recommended three aims for further research on question evaluation methods: (1) understand how to reduce inconsistencies, (2) investigate how to best combine different evaluation methods while capitalizing on the strengths of each, and (3) compare the outcomes of evaluation methods to traditional psychometric measures of reliability and validity. Hence, this suggests what is needed is more confirmatory research identifying the conditions under which question evaluation methods predict data quality in the field. I will now review the relevant studies from the literature using a confirmatory research approach.

Confirmatory Research

There are relatively fewer studies in the literature that attempt to confirm the results of question evaluation methods in the field. Two studies used indirect indicators of data quality to confirm the results from ex-ante or laboratory methods. Forsyth, Rothgeb, and Willis (2004) conducted follow-up research to the initial study reported by Rothgeb, Willis, and Forsyth (2001). They tallied the number of problems found cumulatively by expert review, forms appraisal, and cognitive interviewing. Next, the authors conducted a follow-up field study to see if questions that had relatively more problems identified by these methods resulted in higher levels of item nonresponse, problematic behavior, and problems identified by the field interviewers. Their results showed that the questions identified with more problems by expert review, forms appraisal, and cognitive interviewing did tend to have higher levels of problems in the field. Specifically, they found that interviewer problems found by expert review, forms appraisal, and cognitive interviewing were related to interviewer problems in the field identified by behavior coding and interviewer ratings. Respondent problems found by expert review, forms appraisal, and cognitive interviewing were related to respondent problems found by behavior coding. Finally, recall and item sensitivity problems were related to item nonresponse. In fact, some of the strongest relationships in this study occurred when the problem type was more specific. This suggests that future confirmatory research studies should carefully consider what type of results should be predictive of specific outcomes in the field. This type of analysis is in contrast to what is done in most exploratory studies where overall agreement is given more attention than more specific types of agreement.

Unfortunately, it is not possible to tell from their analysis the relative effectiveness of the methods since the results from all methods were combined in the analysis.

A study by Blair et al. (2007) investigated the extent to which problems identified by cognitive interviewing show up in the field. Their study included 24 questions with problems identified by cognitive interviewing that might be evident in the field. They identified the problems in the field using behavior coding techniques. They coded if the problematic verbatim interviewer-respondent exchange matched problem descriptions from the cognitive test report. Overall, they found that 47% of the problematic interviewer-respondent exchanges matched a problem described in the cognitive test report. However, the authors rely on a strong assumption that the causes of problematic interviewer-respondent exchanges can be easily mapped back to specific cognitive test findings. In most cases, there probably will not be enough verbal information from a standardized interview to draw this conclusion. In addition, the authors did not provide clarification of how they determined which problems were likely to be evidenced in the field. This in and of itself is an empirical question and should be further explicated. Given these challenges in understanding the link between cognitive interviewing and behavior coding, a perhaps more valuable analysis would have focused on the extent to which questions flagged as problematic by cognitive interviewing will result in problematic interviewer exchanges versus questions that were not flagged as problematic by cognitive interviewing.

Hess, Singer, and Bushery (1999) behavior coded 34 questions on food security for which they obtained test-retest measurements. Next, they used the behavior codes to predict the reliability of these questions. They found that two respondent behavior codes were significantly related to the Index of Inconsistency (IOI). The percentage of adequate answers was negatively related and the percentage of qualified answers was positively related to the *IOI*.

Dykema, Lepkowski, and Blixt (1997) behavior coded 10 medical history questions for which they obtained medical records to verify responses. The authors attempted to predict inaccurate responses with the behavior codes at the respondent level. They found no consistent relationship between interviewers' misreading questions and the inaccuracy of the answers. In contrast respondent behavior codes including qualified or don't know answers and interruptions were significant predictors of inaccuracy.

Draisma and Dijkstra (2004) conducted a similar study to Dykema, Lepkowski, and Blixt (1997) that included both behavior coding and response latency. The authors had access to records which provided the true value for several dichotomous questions. Hence, they were able to examine how well these techniques predicted the probability of a correct answer. An analysis of the response latencies illustrated that longer response latencies were associated with incorrect answers. In addition, Draisma and Dijkstra (2004) behavior coded the interviewer-respondent interactions to see how different linguistic and paralinguistic indicators of response uncertainty related to response error. Based on bivariate analyses, they found that linguistic indicators of doubt such as "I think" or "I believe" were associated with incorrect answers. Likewise, some paralinguistic indicators such as answer switches and the number of words used by the respondent to answer a question were also associated with incorrect answers. Last, the authors fit a multivariate logistic regression model that predicted response accuracy and included the response latencies, linguistic indicators, and paralinguistic indicators as predictors in the model. The response latencies and expressions of doubt were significant predictors of response error, whereas the number of words in an answer and answer switches were not significant.

Willis and Schechter (1997) performed a study on questions that were tested for the National Health Interview Survey and a Women's Health Survey. Their study confirmed that problems identified in the lab appeared in the field and that repaired versions of the items performed better in the field. For example, cognitive test results illustrated that a question on strenuous activity that did not include a filter for whether the respondent does any strenuous activity at all, tended to lead to over reporting of the amount of strenuous activity that respondents did. They added a filter to this question and fielded an experiment that compared the response distributions between a question with the filter and another question without the filter. The difference in the response distributions between the two versions was consistent with their predictions. That is, the inclusion of the screening question significantly increased the percentage answering that they did not do strenuous activity. Overall, Willis and Schechter found support for four out of five hypotheses that they tested in this manner.

Although these findings from these confirmatory studies have shown that problems identified by experts or cognitive interviews do appear in the field, various limitations make it difficult to draw firm conclusions from them. Most confirmatory studies to date either focus on the results from a single method or combine results across methods in a way that makes it impossible to understand the relative effectiveness of each method.

In addition, some confirmatory method evaluations focus exclusively on the relationship between method results and quality at the question level (e.g. Hess, Singer, and Bushery, 1999), with small samples of questions. Most of the existing studies in the literature involve as few as 10-12 questions. Thus analyses at the question level make it hard to reliably distinguish what is significant from what is not significant. This issue occurs, because researchers sometimes summarize the data up to the question level. For example, a researcher might analyze the index of inconsistency at the question level rather than modeling item discrepant answers at the question exchange level. The latter approach requires a data set with both respondents and questions repeated in the dataset and multilevel models to appropriately estimate the standard errors in the data set. Finally, virtually all of the existing studies tend to ignore the characteristics of the questions such as whether the questions are factual or subjective and the type of response categories used. This is presumably because of the small number of questions used in combination with the statistical techniques that are used; however, this raises concerns about the findings being robust across different types of questions.

Some studies have recognized some of the issues above, but have not utilized appropriate statistical techniques in conducting analyses (Forsyth, Rothgeb, and Willis, 2004; Blair et al., 2007). Some authors resort to presenting results descriptively without conducting hypothesis tests and reporting standard errors ((e.g. Blair et al. (2007) present the percentage of problematic exchanges that match cognitive interview problems)). Others have conducted bivariate analyses on repeated observations that underestimate the size of the standard errors (Forsyth, Rothgeb, and Willis, 2004). In any event, none of the analyses in the literature utilize the flexibility of multilevel models to conduct hypothesis testing, control for question or respondent characteristics, and explain the variability in data quality. Mulitilevel models have become popular tools in the survey methodological literature on data quality, but so far the power has not been utilized in the question evaluation literature (e.g., Couper and Kreuter, 2012; Pickery and Loosveldt, 2001; Yan and Tourangeau, 1998).

Summary of Findings from the Literature

Table 1.5 summarizes the results that have been reviewed from the literature. Column 1 lists the method of interest and column 2 specifies the comparison. The negative, positive, and equal signs in column 3 indicate whether previous studies found that the focal method identifies fewer, more, or about the same number of problems as the method of comparison. The negative, positive, and zeros in column 4 indicate whether previous studies have found negative, positive, or no agreement between the methods. The negative, positive, and zeros in column five indicate whether the problems found by a method are negatively, positively, or not related at all to reliability. Cells denoted with an "NS" indicate that the relationship has not been studied. In fact, the most disconcerting finding from Table 4 is the lack of evidence the problems identified by these methods lead to less reliable survey questions. Behavior coding is the only method for which I was able to find any evidence that the problems identified by the method leads to less reliable questions.

Table 1.5. Summary of findings from the literature.

Focal Method	Comparison	Number of Problems	Agreement Between Methods		Prediction of Reliability
ER	QAS	-[c]	+[a]		NS
	QUAID	NS	NS		
	SQP	NS	0[a], +[e]		
	CI	+[b, d]	-[e], +[b, c, d]		
	BC	+[b],=[d]	0[a], +[b, d]		
	RL	NS	NS		
QAS	ER	+[c]	+[a]		NS
	QUAID	NS	NS		
	SQP	NS	0[a]		
	CI	+[c]	NS		
	BC	NS	0[a],		
	RL	NS	NS		
QUAID	ER	NS	NS		NS
	QAS	NS	NS		
	SQP	NS	NS		
	CI	NS	NS		
	BC	NS	NS		
	RL	NS	NS		
SQP	ER	NS	0[a], +[e]		NS
	QAS	NS	0[a]		
	QUAID	NS	NS		
	CI	NS	0[e]		
	BC	NS	-[a],		
	RL	NS	NS		
CI	ER	-[b, d]	+[b, c, d]		NS
	QAS	-[c]	NS		
	QUAID	NS	NS		
	SQP	NS	0[e]		
	BC	=[b, d]	+[b, d]		
	RL	NS	NS		
BC	ER	-[b],=[d]	0[a],+[b, d]		+[g]
	QAS	NS	0[a],		
	QUAID	NS	NS		
	SQP	NS	0		
	CI	=[b, d]	+[b, d]		
	RL	NS	+[f]		
RL	ER	NS	NS		NS
	QAS	NS	NS		
	QUAID	NS	NS		
	SQP	NS	NS		
	CI	NS	NS		
	BC	NS	+[f]		

Note. NS = Not Studied. [a]vander Zouwen and Smit 2004 [b]Presser and Blair, 1994 [c]Rothgeb, Willis, and Forsyth, 2001 [d]Willis, Schechter, and Whitaker, 1999 [e]Yan, Kreuter, and Tourangeau, 2012a [f]Draisma and Dijkstra, 2004 [g]Hess, Singer, and Bushery, 1999

Hypotheses

Empirical comparisons between question evaluation methods have produced inconsistent findings. Some have concluded that agreement is generally low (Presser and Blair, 1994; Yan, Kreuter and Tourangeau, 2012a). Others have found moderate agreement (Willis, Schechter, and Whitaker, 1999; Rothgeb, Willis, and Forsyth, 2001). In addition, there is a dearth of evidence suggesting that many of the commonly used question evaluation methods predict data quality that is achieved in the field.

The next step in this line of research is to understand the circumstances under which the findings from the methods converge (or diverge) and to also understand the circumstances under which the methods provide useful results (Presser and Blair, 1994; Yan, Kreuter, and Tourangeau, 2012b; Madans and Beatty, 2012). Furthermore, there is a need for this continued evaluation to occur in a context where multiple methods are compared on questions with known psychometric properties such as reliability or validity (Yan, Kreuter, and Tourangeau, 2012a; Krosnick and Presser, 2010). In order to move forward with this line of research it is important to place the question evaluation method results in the context of the evaluation process, the psychological processes that respondents use to answer survey questions, and the properties of the survey questions that are being analyzed. This dissertation will use both exploratory and confirmatory method evaluations to address these issues.

I first examine how some of the new computer based methods such as QUAID and SQP relate to traditional methods of question evaluation such as expert review, forms appraisal, and cognitive interviewing. These new methods are able to reliably assess the formal characteristics of survey questions. It is often argued that the new computer based methods are adding a dimension to question evaluation that would typically be ignored by question design developers by looking at these form characteristics. Saris (2012) argues that SQP is best thought of as a model-based procedure for analyzing the form of a question. The coding procedures used by SQP are based on theories and results from the fields of linguistics and statistical modeling. In contrast, Saris argues that experts and cognitive interviews are best thought of as methods that use personal judgment to assess the accuracy of a question at measuring a concept. This explains many of the findings from the literature that shows high rates of disagreement between SQP and other methods (e.g. van der Zouwen and Smit, 2004; Yan, Kreuter, and Tourangeau, 2012a). QUAID is, in part, based on models of syntactic theories from the field of linguistics. Graesser et al. (2006) argues that "…syntactic analyses are rather subtle and therefore detectable by few individuals. It is conceivable that experts might learn from QUAID and thereby become more sensitive to these problems." Graesser and colleagues have presented some evidence that QUAID identifies problems that are not identified by experts and some that are identified by experts. However, QUAID results have never been empirically compared to other methods such as cognitive interviewing and forms appraisal. Similar to SQP, one might expect that the results from QUAID may only weakly correlated with the results from expert review, QAS, and cognitive interviewing that are based on a similar cognitive model of the response process and these traditional methods also involve personal judgment.

Hypothesis 1: (Model-based method hypothesis) There will be higher levels of agreement between the traditional methods (e.g. expert review, QAS, cognitive interviewing) than between the model-based methods (e.g. QUAID and SQP) and traditional methods.

In order to move forward, we must begin to understand some of the sources leading to convergence or divergence. The literature has been focused primarily on overall agreement between methods. For example, it is common for the agreement between qualitative methods to be assessed by looking at the correlation in results across all different types of problems (Presser and Blair, 1994; Rothgeb, Willis, and Forsythe, 2001; Willis, Schechter, and Whitaker, 1999). Examining overall correlations between methods is sensible to the extent that the methods have comparable abilities to detect different classes of problems. Most studies find that comprehension problems are the most prevalent among the traditional methods of question evaluation (Conrad and Blair, 1996; Presser and Blair, 1994; Rothgeb, Willis, and Forsythe, 2001; Willis, Schechter, and Whitaker, 1999). However, there are often differences in the methods ability to detect other types of problems such as recall problems, problems with response categories, or analysis problems (Presser and Blair, 1994; Rothgeb, Willis, and Forsyth, 2001). Furthermore, analyses by Yan, Kreuter, and Tourangeau (2012a) suggested that agreement could vary by the type of problem detected for a question. They found higher correlations between the proportion of experts and cognitive interviews that found recall problems compared to other types of problems like comprehension problems. This finding is contrary to what I would expect, given that most of the traditional methods of question evaluation such as expert review, cognitive interviewing, and forms appraisal are generally based on a similar cognitive information processing model. I would expect that these methods would be more likely to agree on comprehension problems since they are the most proficient at detecting these types of problems. I refer to this hypothesis as the problem nature hypothesis.

> Hypothesis 2: (Problem nature hypothesis) The rate of agreement between qualitative methods will vary by type of problem.

An important gap to bridge in our current understanding of question evaluation methods is the extent to which problems found by ex-ante and laboratory methods predict data quality in the field. Confirmatory research designs are needed to bridge this gap. One can use either indirect or direct assessments of data quality in the field depending on the available data. Some examples of indirect assessments include behavior codes, response timings, and item nonresponse. A reinterview study used to provide measures of reliability is an example of direct assessment of data quality. One important point to address from the literature is how to best use the methods together. Esposito (2004) posits the idea that replication of problems across methods is important for a researcher to decide which questions are problematic and may need to be revised. In addition, it has been suggested that it is best to use multiple methods until we know more about the situations in which question evaluation methods converge or diverge (Presser et al., 2004; Yan, Kreuter, and Tourangeau, 2012a). This leads to the complementary method hypothesis.

> Hypothesis 3: (Complementary Method Hypothesis) Using multiple methods together will be better at predicting data quality in the field than using individual methods.

Even though the literature suggests that the methods are complementary and should be used together, it is likely that some methods are more predictive of actual data quality than others. Some authors make a distinction between methods that require data collection and those that do not require data collection (Esposito and Rothgeb, 1997; Saris, 2012). One reason that this distinction might matter is that, in theory, methods that actually observe the process of responding to the survey should be more predictive of actual data quality than methods that simply review questions. Furthermore, methods that are closer to actual survey conditions should be the most predictive. As shown in Table 1.5, this is supported by the current knowledge in the literature since field based methods have been shown to be the most predictive of accuracy and reliability (e.g. Hess, Singer, and Bushery, 1999; Draisma and Dijkstra). This dissertation includes methods that are based on different levels of knowledge about the response process. Computer based methods are based on the least amount of knowledge about the response process for any specific survey question. Expert methods are not based on direct observation of the response process, but are based upon the researchers experience with the survey response process in general. Cognitive interviews are based on direct observation of the response process, but in a perhaps unrealistic setting such as the laboratory where probing behavior by interviewer and the resulting mental processes might lead to a different response process from what occurs in the field. Field based methods such as behavior coding and response latency occur in the most realistic environment compared to the other methods in this dissertation and should be most closely related to data quality. This suggests the following hypotheses:

> Hypothesis 4: (Test Environment Hypothesis) Methods that are implemented in a more realistic survey setting will be most closely related to data quality.

Research has shown that those with lower levels of cognitive ability such as lower educated respondents and older respondents tend to have more difficulty with survey questions (Krosnick, 1987; Alwin, 2007). In addition, it is often suggested that the question evaluation process should focus on evaluating questions particularly for respondents with lower levels of cognitive ability (Esposito, 2004; Willis, 2005). Therefore one might expect that questions identified as problematic may be more likely to cause response problems for those with lower levels of cognitive ability compared to those with higher levels of cognitive ability. Since existing question evaluation studies have focused on question level analyses only, we have not been unable to test this key objective of many question evaluation methods. Cross-classified multilevel models that include effects for both respondents and questions offer a framework to test this hypothesis. I refer to this as the respondent and question problem interaction hypothesis.

> Hypothesis 5: (Respondent and question problem interaction hypothesis) Respondents with lower levels of cognitive ability will have more difficulty with questions identified as problematic by ex-ante and laboratory methods than respondents with higher levels of cognitive ability.

Outline of Remaining Chapters

A diverse mix of qualitative and quantitative methods was analyzed in this dissertation. Each method has unique requirements to prepare for analysis. For example, the problems identified by qualitative methods such as expert review or cognitive interviewing had to be coded into a consistent coding scheme. The procedures that were used and decisions that were made are detailed in Chapter 2.

Chapter 3 presents an analysis of four different types of ex-ante question evaluation methods and one laboratory method: expert reviews, forms appraisal, QUAID, SQP, and cognitive interviewing. The findings provide an understanding of what we can learn about the problematic nature of questions at a relatively low cost since no respondents are needed for the ex-ante methods and only a small number of respondents are typically utilized for cognitive interviewing. The chapter addresses three important research questions from the literature: (1) How much do the methods agree? (2) What circumstances affect the level of agreement? (3) Can the methods detect differences in data quality? The chapter tests the model-based method hypothesis, problem nature hypothesis, complementary method hypothesis, and test environment hypothesis.

Chapter 4 begins to address the question of whether question evaluation method results predict data quality in the field. I employ field-based measures such as behavior coding and response timings that provide information about the quality of survey data from the field. I then look at whether the results from QUAID, SQP, expert review, forms appraisal, and cognitive interviewing predict behavior codes and response times. I also look at whether the question evaluation method results interact with question characteristics and respondent characteristics. This chapter tests the complementary method hypothesis, test environment hypothesis, and the respondent and question problem interaction hypothesis.

Chapter 5 addresses the question of whether evaluation method results are predictive of results from traditional psychometric methods. The analyses in this chapter investigate how effectively QUAID, SQP, expert review, forms appraisal, cognitive interviewing, behavior coding, and response latency predict the consistency of survey questions over time. I look at whether the results interact with various question and respondent characteristics in order to understand the circumstances under which the methods may predict reliability. This chapter tests the complementary methods hypothesis and test environment hypothesis

CHAPTER 2: METHODS

The data for this dissertation come from the 2006 Joint Program in Survey Methodology (JPSM) Survey Practicum. The JPSM Survey Practicum is a two-semester course in which graduate students gain experience developing a questionnaire, sampling a population, collecting and analyzing data, and reporting results. The practicum exposes students to realistic problems in survey design and implementation. The course begins in the spring semester with the questionnaire development process, main data collection occurs over the summer months, and the fall semester is devoted to data analysis and the reporting of results.

The sponsors for the 2006 Survey Practicum were interested in why people give inconsistent answers to survey questions. The aim of the research was to examine the consistency of responses to attitude and behavioral questions over short periods of time. The students developed a questionnaire during the spring of 2006 that included questions on a variety of topics. In general the questionnaire included four types of questions: two types of attitudinal questions — questions asking about relatively familiar issues (the Iraq war and wiretapping) and questions about a new or unfamiliar issue (a new school-based program in mathematics or English); quasi-attitudinal questions (self-ratings of health or disability status); behavioral questions (e.g. doctor visits, trips to movies and restaurants); and a few demographic questions.

Data preparation

Three sets of questions were evaluated for this study. One set of questions was from a questionnaire that was evaluated with cognitive interviews by the practicum students. The second set of questions were the questions that were actually fielded for the practicum. The third set of questions come from a review of studies in the literature in which survey responses were compared with administrative records. A total of 231 questions were evaluated for the study.

Computer based systems. Two computer-based systems were used to evaluate the survey questions. First, the Question Understanding Aid (QUAID)[9] is a computer tool that was developed by Graesser and colleagues (Graesser et al. 2006). It is based on computational models developed in the fields of computer science, computational linguistics, discourse processing, and cognitive science. The software identifies technical features of questions that have the potential to cause question comprehension problems. The current version of QUAID critiques each survey question on five classes of comprehension problems: unfamiliar technical terms, vague or imprecise predicate or relative terms, vague or imprecise noun phrases, complex syntax, and working memory overload. QUAID generally identifies these problems by comparing the words in a question to several databases or data files (e.g., Coltheart's MRC Psycholinguistics Database). I entered the text of the questions and response options into the QUAID tool for evaluation.

Second, the Survey Quality Predictor (SQP) is a computer program created by Willem Saris and colleagues. The program uses results from a meta-analysis of previously conducted multitrait-multimethod (MTMM) studies to predict the reliability, validity, method effects and total quality of a survey question (Saris and Gallhofer 2007). Total quality is the product of reliability and validity. In order to use SQP, the researcher

[9] The QUAID tool can be found online at http://mnemosyne.csl.psyc.memphis.edu/QUAID.

codes each question according to the variables from the MTMM studies in order to obtain these predictions. I coded the questions using the latest version of SQP2 tool that was released in 2012.

Expert Review. Three experts reviewed each questionnaire. They were e-mailed copies of the questionnaire to which they were assigned. Each of the reviewers had more either had a Ph.D. in survey methodology or a related discipline or had more than five years of experience as a survey manager or researcher. The reviewers who were asked to review the questionnaire for the practicum questions were given the following instructions:

> Question wordings, introductions associated with questions, and response categories are considered in scope for this evaluation. For each survey question, identify and briefly explain each specific problem you find. Please type a brief description of the problem immediately following the question in the attached document. You may observe multiple problems with a question. Please describe each one. You do not need to type anything after questions for which you do not observe a problem.

Another questionnaire included questions from the literature that had record checks done on them. The reviewers who reviewed the record check questions were given similar instructions with the following information to preface how the questions were adapted for this dissertation.

> The questions are not intended to be part of a single questionnaire, but are compiled from various surveys that have been conducted over time. They were chosen from studies that used administrative records to check the accuracy of the questions. Some of the questions have been modified slightly. For example, the reference year of a question might have been changed so that the question is relevant for today. I have attempted to group questions appropriately by topic in the attached document. Generally, questions from similar studies are grouped together.

> I realize that the nature of the attached questionnaire makes it difficult to assess question context for most of the questions. Please do your best to evaluate the specific questions given this constraint.
>
> In terms of the mode of data collection, you may assume that the questions are being fielded over the telephone. You may also assume that the questions were fielded at a time prior to the development of cell phone technology.

Forms Appraisal. Students from a graduate level course on questionnaire design at JPSM were asked to complete the Questionnaire Appraisal System (QAS) for each of the questions in this study. The students were not specifically asked about their level of experience with question design, but it is expected that they had less experience than the survey experts. This is in accordance with the purpose of the QAS, which is to give a survey novice a structured tool to review questions. The questionnaire was divided into different sections so that the students evaluated roughly 30-40 survey questions. The form also had room for a brief description of each problem found. Typically three or four students were assigned to a section of the questions. The students were given the following instructions.

> Forms appraisals are tools that allow question designers to identify common problems that occur with survey questions. One of the most prevalent forms appraisals in the question design literature is the Questionnaire Appraisal System (QAS) introduced by Willis and Lessler (1999) [QAS manual available on CTOOLS].   The QAS looks for seven different types of problems that might occur with survey questions, including, problems with reading, instructions, clarity, assumptions, knowledge or memory, sensitivity or bias, response categories. The QAS asks a series of questions within each type of problem to help the question designer understand any potential errors that might occur with a question. Your assignment for problem 1 is to use the QAS to evaluate several questions as we have discussed in class.
>
> Each question is evaluated by proceeding down Column B and checking to see if you think the question has that particular problem. Enter 1 in Column C if you think the question has the problem and enter a BRIEF note in Column D describing the problem. Enter 2 in Column C if you do not think the question has

the problem. Enter codes for all QAS problems in steps 1-8 and then proceed to the next question by clicking on the next worksheet.

The exact form that the students filled out is shown below in figure 2.1.

Figure 2.1 QAS form used by students to evaluate survey questions.

Code Description	**Code**	**Comment**
	1=Yes; 2=No	
WHAT TO READ: Interviewer may have difficulty determining which parts of the question should be read.		
MISSING INFORMATION: Information the interviewer needs to administer the question is not contained in the question.		
HOW TO READ: Question is not fully scripted and therefore difficult to read.		
CONFLICTING OR INACCURATE INSTRUCTIONS, introductions, or explanations.		
COMPLICATED INSTRUCTIONS, introductions, or explanations.		
WORDING: Question is lengthy, awkward, ungrammatical, or contains complicated syntax.		
TECHNICAL TERM(S) are undefined, unclear, or complex		
VAGUE: There are multiple ways to interpret the question or decide what is to be included or excluded		
REFERENCE PERIODS are missing, not well specified, or in conflict		
INAPPROPRIATE ASSUMPTIONS are made about the respondent or about his/her living situation.		
ASSUMES CONSTANT BEHAVIOR or experience for situations that vary.		
DOUBLE-BARRELED: Contains more than one implicit question.		
KNOWLEDGE may not exist: Respondent is unlikely to know the answer to a factual question.		
ATTITUDE may not exist: Respondent is unlikely to have formed the attitude being asked about.		
RECALL failure: Respondent may not remember the information asked for.		
COMPUTATION problem: The question requires a difficult mental calculation.		
SENSITIVE CONTENT (general): The question asks about a topic that is embarrassing, very private, or that involves illegal behavior.		
SENSITIVE WORDING (specific): Given that the general topic is sensitive, the wording should be improved to minimize sensitivity.		
SOCIALLY ACCEPTABLE response is implied by the question.		
OPEN-ENDED QUESTION that is inappropriate or difficult.		
MISMATCH between question and response categories.		
TECHNICAL TERM(S) are undefined, unclear, or complex.		
VAGUE response categories are subject to multiple interpretations.		
OVERLAPPING response categories.		
MISSING eligible responses in response categories.		
ILLOGICAL ORDER of response categories.		
Other problems not previously identified.		

Cognitive interviews. Students enrolled in two graduate level courses in the Joint Program in Survey Methodology conducted cognitive interviews on different sets of questions. First, students enrolled in the practicum conducted cognitive testing on the initial questionnaire. Each student was instructed to develop their own cognitive protocol. The cognitive protocol consisted of the initial questionnaire and any think-aloud exercises and cognitive probes that the student utilized during the testing of the questionnaire. Thirteen students interviewed four subjects each for a total of 52 completed cognitive interviews. The students recruited subjects among their friends, neighbors, co-workers, and/or other convenient populations. All interviews were recorded so that the students could review the recordings when preparing their reports on the findings from the cognitive interviews. The students' reports and audio tapes were turned in to the JPSM instructors when completed. Revisions were made to the questionnaire following a classroom discussion about the findings from the cognitive interviews.

The same design was implemented for the testing of the questionnaire used in the field. The same students from the graduate course on question design who conducted the QAS coding of the questions also cognitively tested the fielded questionnaire. This questionnaire included revised versions of the questions that were tested by the Practicum students. It excluded some questions from the final practicum questionnaire that were not part of the original practicum questionnaire. A total of thirteen students completed four interviews each and also wrote a report summarizing their findings.

Four remaining students from the graduate level course on questionnaire design tested the remaining questions. This included questions added to the final practicum survey that had not been included in earlier cognitive interviews and those from record check studies. These students conducted four cognitive interviews each and also wrote a report summarizing their findings.

Problem Coding. The problems identified from QUAID, QAS, expert review and cognitive interviewing were then coded according to the same coding scheme used by Presser and Blair (1994). The coding scheme has four basic categories: respondent semantic, respondent task, interviewer, and analysis problems. Respondent semantic problems occur when respondents have difficulty understanding a question, remembering the question, understanding the meaning of particular words or concepts in the question or when respondents have different understandings of what a question refers to. There are two subtypes of semantic problem. The first type refers to problems with wordiness, question structure, or relationships between questions. This can generally be thought of as problems with the structure of the question or the questionnaire. Another type of semantic problem occurs when the respondent has difficulty interpreting the meaning of questions due to terms or concepts within the question. There are also three subtypes of Respondent task problems. The first refers to difficulty recalling information or formulating an answer. The second type of respondent task problem is due to insufficient response categories. The last type of respondent task problem deals with a question being sensitive for the respondent to answer. Interviewer problems refer to problems reading the question or having difficulty understanding how to implement a question. Analysis problems occurred when the problem creates difficulties with data analysis (e.g. lack of variation in responses).

Table 2.1. Presser-Blair problem coding scheme.

Semantic I: Problems with question structure	
Information overload	Contains too much text or too many response categories to be retained or understood ("information overload")
Structure / organization	Words or ideas are structured or organized unclearly.
Transition problem	Question's intelligibility affected by an earlier question or questions (lacks needed transition).
Semantic II: Problems with meaning of terms	
Boundary lines	Respondents differ on what the question includes or excludes or are uncertain what the question refers to.
Technical term not understood	Technical term is not understood
Common term not understood	common term is not understood (e.g., used an unusual way)
Double-barreled	A single question asks about more than one subject, each of which could be answered differently ("double barreled")
Respondent Task I: Problems with recalling information	
Recall/response difficult	Difficult--the level of response detail, demand on memory, or some other feature of the task is too difficult
Recall/response impossible	Impossible--information requested is not known
Recall/response redundant	Redundant--answer has (or seems to have been given to an earlier item
Recall/response resisted (assumptions)	Resisted by respondent--makes an assumption that is inappropriate or not sensible
Respondent Task II: Problems with response categories	
Overlapping response categories	overlapping response categories
Insufficient response categories	response categories being insufficient (category is missing)
Too fine distinction between categories	response categories making too fine a distinction
Response categories not appropriate to Q	response categories not appropriate to question
Respondent Task III: Problems with question sensitivity	
Sensitivity	item requires admitting ignorance, undesirable behavior, or something else that leads to discomfort
Interviewer problems	
Procedural	Unclear how the question is supposed to be asked
Reading problem	caused by length, awkward syntax, pronunciation, etc
Coding to open question	Coding answers to an open question
Analysis issues	
Question answered same by all respondents	
Question suggests answers	
Acquiescence	
Order of response categories	

Two research assistants and the study author coded the problems from the cognitive interviews and expert reviews into the Presser-Blair coding scheme. There were 586 problems identified by cognitive interviewing. The two research assistants coded all of these problems. Approximately twenty percent of the cognitive interview problems were double-coded by to measure the reliability of the coding scheme. The overall kappa for the coding process was .76. The Kappa values for each category are as follows: respondent semantic (.91), respondent task (.73), interviewer problems (.68), and analysis problems (.49). The study author adjudicated any discrepancies between the initial coder and second coder on this 20 percent. There were 960 problems identified by expert review and I coded all of these problems. A crosswalk was used to systematically code the QUAID and QAS problems into Presser-Blair categories. These crosswalks are shown in Table 2.2. Finally, I then determined which problems matched across methods and assigned an identifier to each problem.

Table 2.2. Crosswalk between QUAID, QAS, and Presser-Blair codes.

Presser Blair (1994) Codes		QUAID	QAS
Semantic I	Information overload	Working memory overload	
	Structure / organization	Complex syntax	Conflicting or inaccurate instructions, Complicated instructions, Wording
	Transition problem		
Semantic II	Boundary lines	Vague or imprecise relative or technical term, Vague or ambiguous noun phrase	Vague, Reference period
	Technical term not understood	Unfamiliar technical term	Technical term
	Common term not understood		
	Double-barreled		Double barreled
Respondent Task I	Recall/response difficult		Computation
	Recall/response impossible		Knowledge, Attitude, Recall
	Recall/response redundant		
	Recall/response resisted		Inappropriate assumptions, Assumes constant behavior
Respondent Task II	Overlapping response categories	Vague or imprecise relative or technical term, Vague or ambiguous noun phrase	Vague, Overlapping
	Insufficient response categories		Open ended, Missing
	Too fine distinction between response categories		
	Response categories not appropriate		Mismatch
Respondent Task III	Sensitivity		Sensitive content, Sensitive wording, Socially acceptable
Interviewer	Procedural		
	Reading problem		What to read, Missing information, How to read
	Coding answers to open		
Analysis	Question answered same by all respondents		
	Question suggests answers		
	Acquiescence		
	Order of response categories		
			Other

Behavior coding. A total of 377 survey interviews were randomly selected for behavior coding in this study. The interviewer-respondent exchanges were coded according to the coding scheme shown in Table 2.3. The coding scheme includes interviewer codes that capture the extent to which the interviewer read the question exactly as printed in the instrument and whether or not the interviewer had to probe to record a final answer. The coding scheme includes codes to indicate the adequacy of the respondents answer, whether the respondent requested clarification, whether the respondents used pauses or fillers, and whether the respondent interrupted the reading of the question.

Table 2.3. Behavior coding scheme used in the study.

Variable	**Short Description**	**Detailed Description**	**Kappa**
Interviewer Codes			
EX	Exact	Interviewer initially reads the question exactly as printed.	.64
SC	Slight Change	Interviewer initially reads the question changing a minor word that does not alter question's meaning. For example, the interviewer leaves out the article, "a" or "the."	.53
MC	Major Change	Interviewer initially changes question such that the meaning is altered. Interviewer does not complete the reading of the question. Interviewer skips a question that should have been asked. Interviewer skips continuous words that are not articles or prepositions. Interviewer paraphrases question.	.80
PB	Probing	Interviewer probes during any interaction in the in the question answer sequence. Interviewer repeats all or part of the question, including response categories.	.79
Respondent Codes			
AA	Adequate Answer	Respondent's initial answer meets question objective.	.87
QA	Qualified Answer	Respondent initially gave a qualified answer that indicated doubt or uncertainty on the part of the respondent. Examples include "I think," "Maybe," "probably," or "about."	.74
IA	Inadequate Answer	Respondent's initial answer does not meet question objectives.	.87
DK	Don't Know	Respondent initially gives a "don't know" or equivalent response.	.76
RF	Refusal	Respondent initially refused to answer the question.	.85
RI	Respondent Interruption	Respondent interrupted the initial asking of the question to provide an answer or request clarification.	.88
PF	Pauses or Fillers	Respondent pauses for longer than one second or uses a filler such as "ah," "um," or "well" immediately after the initial reading of the question.	.42
RC	Respondent Clarification	Respondent asks for clarification of question or makes a statement indicating uncertainty about question meaning at any point in the question answer sequence. Respondent asks to have all or part of the question repeated, including response categories.	.91

Two research assistants coded 292 interviews and I coded 85 interviews. Research assistants coded the interviews after reviewing the coding scheme with the study author and practicing the coding scheme on a few interviews. Approximately six percent of the cases were double coded to obtain a measure of the reliability of the coding process. Kappa values for each of the codes also shown in Table 2.3.

Response latency. Response latency measurement involves a researcher measuring the amount of time from the end of the interviewers reading of a question to the time when a respondent begins to answer. Response latency timings were recorded on a subsample of 111 of the same cases that were behavior coded. The first 111 case IDs that were behavior coded were selected for response latency measurement. One research assistant coded 88 interviews and I coded the remaining 23 interviews.

Test-Retest measures. An important feature of the Practicum was the implementation of a reinterview. Respondents to the initial interview were asked to repeat the interview two weeks later. A total of 53 questions were repeated between time 1 and time 2. The responses to these questions over time were used to compute measures of consistency or reliability. I computed measures of agreement at both the question level and individual level. At the question level, the index of inconsistency was computed. At the individual level I computed discrepancies between time 1 and time 2 for each respondent at each question administered at both points in time. I allowed a difference of 1 for questions that asked about continuous information such as the number of times that the respondent exercised, went to a movie, or went out for dinner.

Record check data. One way to judge the validity of survey questions is to compare the answers to survey questions with records. There are several studies in the literature that have made types of record checks. A total of 51 questions from the literature with record checks were analyzed for this dissertation (see Appendix C). These questions came from 8 different studies. I abstracted the percent of correct answers for each question with a record check. This was usually relatively straight forward in most cases. Occasionally studies reported this figure for several conditions (e.g. percent correct within each mode). In this case, I averaged across the conditions and recorded the average percent correct across all conditions. In chapter 3, I examine the relationship between the problems that are identified with survey questions and the reports of the accuracy of the survey questions from published record check studies.

Analytic Approach

One goal of this dissertation is to improve the analytical techniques that are used to evaluate methods. Most prior studies have used a relatively small number of questions, and even when a larger number of questions have been used potentially important characteristics of the questions themselves have not been taken into account. Ideally, one wants to sample from the universe of survey questions when conducting a study of question evaluation methods. This is difficult, if not impossible, since the universe is relatively undefined and expanding over time. However, one can use appropriate models to control for key characteristics such as question type or response format.

There is also a tendency for existing studies to ignore the structure of the data in which the findings arise. Most analyses aimed at understanding question evaluation methods occur at the question level. As outlined in chapter one, the question-answer process involves a complex interaction between the survey instrument and the survey respondent. This means that both indirect and direct indicators of data quality such as behavior codes, response latencies, item nonresponse, and response consistency are measured on individual respondents who answer a set of survey questions. In other words, these measures are nested within the cross-classification of respondents and questions. This suggests a certain structure for a data file that is looking at the quality of questions across the survey instrument. Table 2.4a shows the structure of the typical data file where respondents are represented in the rows of the data file and questions are represented in the columns.

Table 2.4a. Typical survey data file.

Respondent	Question 1	Question 2	Question 3
1	1	9	1
2	1	2	NA
3	1	9	9

We need to construct summary measures from these data in order to examine differences in data quality. For example, we could calculate the average item nonresponse rate for each question and then predict this using the results from the question evaluation methods. This results in a relatively small data set, which means that the power to detect any differences will be weak. Aggregating over respondents for a question also means that the researcher will not be able to look at potentially interesting interactions between question characteristics and respondent characteristics.

A preferred method of analysis is to recognize the context in which the data arise and model the data appropriately. This can be done by transforming the data file as shown in Table 2.4b, which repeats questions and respondents. Column 3 of the data file includes a data quality indicator for item nonresponse that indicates whether a specific respondent did not respond to a specific question. The remaining columns in the data set then represent indicators for question characteristics or respondent characteristics.

Table 2.4b. Transformed survey data file.

Respondent	Question	Item Nonresponse	Question characteristic (e.g. pretest problem)	Respondent Characteristic (e.g. low education)
1	1	0	1	1
1	2	1	0	1
1	3	0	1	1
2	1	0	1	0
2	2	0	0	0
3	1	0	1	1
3	2	1	0	1
3	3	1	1	1

This creates a long form data set where the observations are not independent. Intra-respondent and intra-question correlations due to the clustering of observations within respondents and questions act to increase the variance of parameter estimates. Therefore, analyzing the data using standard statistical techniques such as regression or bivariate crosstabulations will lead to grossly understated standard errors. Fortunately, multi-level models can be used to appropriately model this type of data. Multi-level models have become more popular over recent years to evaluate similar data sets evaluating data quality across respondents and questions. For example, similar models have been used to look at variability in item nonresponse (e.g., Pickery and Loosveldt, 2001) and question timings (Yan and Tourangeau, 1998; Couper and Kreuter, 2012). We are particularly interested in cross-classified multilevel models since the data quality indicators of interest in this dissertation are nested within the cross-classification of both questions and respondents. Chapter 4 examines the relationship between method results and three indirect indicators of data quality from the field: behavior codes, item nonresponse, and response timings. Chapter 5 examines the relationship between method results and a direct indicator of data quality – the consistency of responses over time. The cross-classified model used to predict a binary data quality indicator such as item nonresponse is summarized below in equation 2.1. This model represents the null model that includes no predictors.

Equation 2.1

$$\text{Logit}\left(\pi_{ij}\right) = \beta_0 + u_{1j} + u_{2i}, \quad u_{1j} \sim N(0, \sigma_{u1}^2),\ u_{2i} \sim N(0, \sigma_{u2}^2)$$

The logit of the probability of respondent j not responding to item i equals an overall mean plus a random effect for respondents (u_{1j}) and a random effect for questions (u_{2i}). The random effect for respondents represents the amount of variability around the overall mean level of item nonresponse for the respondents. It comes from a distribution with mean equal to 0 and variance σ_{u1}^2. The random effect for items represents the amount of variability around the overall mean level of item nonresponse for the items. It comes from a distribution with mean equal to 0 and variance σ_{u2}^2. This model provides a baseline estimate of the amount of variability in item nonresponse and partitions the variance into variability due to the respondent and to the question. Predictors or fixed effects can easily be added to this model. For example, in order to understand how much the predicted probability of item nonresponse will increase when a question is flagged with a problem by cognitive interviewing we would run the following model in Equation 2.2.

Equation 2.2

$$\text{Logit}\left(\pi_{ij}\right) = \beta_0 + \beta_{1i}CIProb + u_{1j} + u_{2i}, \quad u_{1j} \sim N(0, \sigma_{u1}^2), \qquad u_{2i} \sim N(0, \sigma_{u2}^2)$$

CIProb equals 0 if cognitive interviewing does not find a problem with a question and 1 if cognitive interviewing does find a problem with a question. The addition of this predictor not only allows us to understand the fixed effect of knowing that cognitive interviewing has found a problem with a question, but we can also understand how much of the variability in item nonresponse is reduced when we had this predictor to the model. Similarly, one can add respondent level predictors to the model and respondent by question level predictors to understand cross-level interactions.

The models are also very flexible because they allow the testing of a number of relevant hypotheses. First, the traditional significance tests for each coefficient in the model indicate whether a variable is a significant predictor of data quality. Contrast statements can be constructed to test for the difference between individual coefficients in the model. Second, more complex hypotheses can be tested using model fit statistics. For example, when models are nested within each other, likelihood ratio tests can be conducted using the difference in the -2 log likelihood statistics for the models. Non-nested models can be compared to each other using statistics such as the Akaike Information Criterion. Models were run using either SAS PROC GLIMMIX or SAS PROC MIXED.

CHAPTER 3: A COMPARISON OF EX-ANTE AND LABORATORY METHODS OF QUESTION EVALUATION

Introduction

There are often very few resources allocated to the pretesting stage of a survey project. Hence, question evaluation at this stage must often rely on ex-ante methods or methods that require no data collection. It may also be feasible at this stage to bring a few respondents into the laboratory to participate in cognitive interviews. The goal of this chapter is to provide clarification about how to best utilize ex-ante and laboratory question evaluation methods in the question development process.

A few methodological comparisons between ex-ante and laboratory methods were discussed in the literature review in chapter one. These studies typically compare methods on the number and type of problems found. These studies have achieved mixed results in terms of the amount of agreement between the methods. Most of the studies find either weak or moderate agreement between the methods. Other times agreement is in the opposite direction from what is expected. In general, there is a consensus that the next step in this line of research is to explore causes of the inconsistent agreement between the methods (Madans and Beatty, 2012; Presser and Blair, 1994; Yan, Kreuter, and Tourangeau, 2012a). The goals of this chapter are to address some possible sources of the inconsistent results.

I first examine the extent of agreement between methods. I will focus attention specifically on the new computer based methods, such as QUAID and SQP, and how closely they agree with traditional methods such as expert review, QAS, and cognitive interviewing. The model based method hypothesis suggests that the computer based methods detect new types of problems and will have a high rate of disagreement with the traditional methods.

Next, I address the extent to which the nature of the problem affects agreement between methods. As discussed in chapter one, most methods focus on comprehension or interpretation issues, but are capable of detecting a multitude of problems. The problem nature hypothesis suggests that the level of agreement varies by the nature of the problem.

Last, I begin to examine the important question of whether the question evaluation methods are able to provide insight that leads to better question design. This section of the chapter looks at whether the methods are able to detect the difference (find fewer problems) between original questions and revised questions. In addition, I look at whether the methods can predict the accuracy of questions with record checks. The goal of these two analyses is to begin to understand how the methods can be used together to understand data quality. I will examine the complementary methods hypothesis, which suggests that multiple methods used together will be better than using individual methods to understand data quality. Finally, I will examine the test environment hypothesis, which suggests that the results from cognitive interviewing where the response process is observed the closest will be better at predicting data quality than the ex-ante methods.

Results

Ultimately a researcher is concerned about the conclusions that a method draws about a question; however, it is important to recognize that methods can agree on whether a question is problematic without agreeing on the specific cause of the problem. Hence, the following analysis begins at the question level, but proceeds down to the level of specific problems. These analyses include a total of 151 questions that were included in either the questionnaire that was cognitively tested for the survey practicum or fielded in the final practicum survey.

Detection of problematic questions

The probability of detecting any problem with a question was first examined. As shown in Table 3.1, there was a high probability of problem detection with each method. As in Rothgeb, Willis, and Forsyth (2001), QAS found a problem with nearly every question (144/151 or 95.4% of questions). The other methods found a problem with approximately 80% of the questions.

Table 3.1. Percent of questions with a problem detected by method (N = 151).

Problem detected?	QUAID	QAS	Expert review	Cognitive interviewing
Yes	84.8%	95.4%	81.5%	81.5%
No	15.2	4.6	18.5	18.5
	100	100	100	100

Note. McNemar's test of marginal homogeneity significant (p<.05) between QAS and all other methods.

One can also look at the overlap in problem detection between methods. Several different measures of agreement are used in the literature. I show only a few of them in Table 3.2 and Table 3.3. Similar to Presser and Blair (1994), the Yule's Q statistic is shown in Table 3.2. It shows the extent to which the methods agree that there was any problem with a question. According to this statistic, the level of agreement at the question level is quite high for most pairs of methods. The highest level of agreement is between expert review and QAS (Yule's Q = .94). Cognitive interviewing also has quite high correlations with both QUAID and expert review.

Table 3.2. Overlap in problem detection (Yules Q) between methods (N=151).

	QUAID	Expert review	QAS	Cognitive interviewing
QUAID	-	.65*	.40*	.81*
Expert review		-	.94*	.91*
QAS			-	.46*
Cognitive interviewing				-

*p<.05

Based on the results from Table 3.1, there is a high probability that the methods will agree that a question is problematic based on chance since each method has at least an 80 percent chance of finding a problem with a question. The kappa statistic adjusts for this chance probability (Cohen, 1960). Table 3.3 shows the values of Kappa for each pair of methods above the diagonal and the numbers in the lower diagonal represent the proportion of questions where the each pair of methods agrees whether or not a question is problematic. Kappa values are in the range of .07 to .56 for the cells above the diagonal. Based on benchmarks proposed by Landis and Koch (1977), the kappa values indicate poor agreement between expert review and QAS. There is fair agreement between expert review and QAS and also between QAS and expert review. The data show moderate agreement between cognitive interviewing and both QUAID and expert review. There is fair agreement between cognitive interviewing and QAS.

Table 3.3. Overlap in problem detection (Kappa) between methods (N=151).

	QUAID	Expert review	QAS	Cognitive interviewing
QUAID	-	.27*	.07*	.41*
Expert review	.79	-	.29*	.56*
QAS	.83	.85	-	.35*
Cognitive interviewing	.83	.87	.86	-

*p<.05

The Survey Quality Predictor uses the results from multitrait-multimethod experiments to predict the quality of survey questions. Table 3.4 shows the average total quality of the questions as predicted by the Survey Quality Predictor by whether or not each method predicted a problem with the question. The point-biserial correlation between the SQP total quality score and whether or not the other methods (QUAID, expert review, QAS, cognitive interviewing) detected a problem is also shown in the table. There is essentially no difference between the total quality predicted by SQP when the other methods detect a problem and when they do not detect a problem.

Table 3.4. Survey Quality Predictor total quality by detection of method specific problems.

Method	Point biserial correlation	Mean Quality		Significance
		Problem	No problem	
QUAID	.10	.57	.59	n.s.
Expert review	.06	.58	.57	n.s.
QAS	.17	.62	.57	n.s.
Cognitive interviewing	.02	.58	.58	n.s.

These results at the question level provide a somewhat unclear evaluation of the model-based method hypothesis. On the one hand, depending on the statistic, QUAID has a similar level agreement with expert review, QAS, and cognitive interviewing as the methods do amongst themselves. On the other hand, SQP shows very low levels of agreement with all of the methods on the level of agreement. The correlations between SQP and expert review or QAS mirror the size of the correlations found by van der Zouwen and Dijkstra (2004).

Detection of specific problems

The next analyses examine the agreement between methods on specific problems. SQP is excluded from this analysis, because it provides evidence of overall quality for a question and does not indicate specific problems. Table 3.5 shows the percentage of all problems that were identified by each method. QAS identified two-thirds of all problems, whereas the other methods all identified less than one third of all problems. The differences between expert review and cognitive interviewing are not statistically significant. Each of these methods only identified about one fourth of all of the problems.

Table 3.5. Percent of problems detected by method (N = 1,107).

Problem detected?	QUAID	QAS	Expert review	Cognitive interviewing
Yes	32.4%	66.9%	25.9%	27.8%
No	67.6	33.1	74.1	72.2
	100	100	100	100

Note. McNemar’s test of marginal homogeneity significant (p<.05) for each pair of methods except expert review and cognitive interviewing.

Table 3.6 illustrates the overlap in problem detection between methods. The table shows the Yule’s Q coefficient measuring the correlation of problem detecting between each pair of methods. QUAID has a negative correlation with all other methods when looking at specific problems that are detected. This supports the model based method hypothesis and the finding from Graesser et al. (2006) that QUAID tends to detect different problems than methods such as expert review and cognitive interviewing. We now know that this finding extends to other methods such as cognitive interviewing. There is a moderate, but significant correlation between expert review and cognitive interviewing (Yule's Q = .25) with respect to the specific problems that each method detects. This is consistent with Presser and Blair (1994) who found correlations between .07 and .48 between expert review and cognitive interviewing.

Table 3.6. Overlap in problem detection (Yules Q) between methods (N=1,107).

	QUAID	Expert review	QAS	Cognitive interviewing
QUAID	1	-.40*	-.73*	-.49*
Expert review		1	-.05	.25*
QAS			1	.15
Cognitive interviewing				1

*p<.05

Table 3.7 illustrates the kappa statistic (upper diagonal) and proportion agreement (lower diagonal) on the presence of specific problems. The negative kappa values indicate systematic disagreement between QUAID and all other methods. There is slight agreement between cognitive interviewing and both expert review and QAS.

Table 3.7. Overlap in problem detection (Kappa) between methods (N=1,107).

	QUAID	Expert review	QAS	Cognitive interviewing
QUAID	1	-.16*	-.32*	-.20*
Expert review	.52	1	.02	.11*
QAS	.26	.41	1	.05*
Cognitive interviewing	.50	.65	.45	1

Note. Numbers above the diagonal are kappa statistics. Numbers below the diagonal indicate proportion that agree whether or not a problem exists.
*p<.05

All of the previous analyses are looking at overall problem detection. We next examine the problem nature hypothesis that suggests that the agreement between methods will vary depending on the nature of the problem. While the literature suggests that comprehension problems are the most prevalent problems with all of these methods, they may have different abilities to detect other types of problems (Presser and Blair, 1994; Rothgeb, Willis, and Forsyth, 2001; Willis, Schechter, and Whitaker, 1999). One can examine the distribution of the types of problems found by each method. As shown in Table 3.8, the majority of the problems found by each method are problems with question structure or ambiguity. These are primarily the types of problems that would affect comprehension, which is consistent with the existing literature. However, there are some significant differences between each of the distributions shown in Table 3.8. This is not surprising for QUAID, because it focuses almost exclusively on problems with comprehension. For example, there still are not very effective algorithms for identifying sensitive topics or words. There are more subtle differences between expert review, QAS, and cognitive interviewing. For example, expert review tends to find a higher percentage of problems with question structure compared to QAS and cognitive interviewing. Expert review also tends to find a higher percentage of analysis problems compared to QAS and cognitive interviewing.

Table 3.8. Distribution of problem type by method.

Problem type	QUAID	Expert review	QAS	Cognitive interviewing
Question structure	26.2%	18.6%	9.5%	9.4%
Ambiguity	64.4	42.4	42.4	50.3
Recall	0.0	17.6	21.6	19.8
Response categories	9.5	8.6	11.7	11.7
Sensitivity	0.0	3.1	12.0	5.8
Interviewer	0.0	0.7	2.3	2.6
Analysis	0.0	9.0	0.5	0.3
	100	100	100	100
N	(359)	(287)	(741)	(308)

Note. Chi-square tests reveal significant differences between all distributions shown in the table except for QAS and cognitive interviewing.

Since the majority of problems found by each of these methods is related to comprehension or meaning, it is important to focus on any differences that might occur between the methods in detecting these types of problems. The Presser-Blair problem coding scheme involves two general types of issues that are primarily related to comprehension or meaning. Semantic I problems relate to the structure of the question and semantic II problems refer to problems with the meaning of concepts or terms in a question. Table 3.9 shows the distribution of these two types of problems by method. Expert review detects a significantly higher percentage of problems with question structure compared to QUAID. Both QAS and cognitive interviewing detect a lower percentage of question structure problems compared to both QUAID and expert review.

Table 3.9. Distribution of problem type by method

Problem type	QUAID	Expert review	QAS	Cognitive interviewing
Semantic I: Question structure	28.9%	30.5%	18.2%	15.8%
Semantic II: Problems with meaning	71.1	69.5	81.8	84.2
	100	100	100	100
N	(325)	(177)	(409)	(190)

Note. Chi-square tests reveal significant differences between expert review and all methods. QUAID is also significantly different from all methods (p <.05).

Variation in agreement by nature of the problem

Table 3.10 illustrates the level of agreement between methods by problem types. The results, once again, show that QUAID tends to detect different types of semantic problems than all of the other methods. The highest level of agreement between expert review, QAS, and cognitive interviewing is on semantic problems involving comprehension of terms or other interpretation issues. There is lower or even negative agreement between these three methods when looking at semantic issues related to the structure or organization of the question. Hence, even though Table 3.8 showed that QUAID and expert review are equally likely to detect issues with question structure, they identify different structural issues. Overall, the results in Table 3.9 support the problem nature hypothesis. The rate of agreement between methods varies depending on the nature of the problems that are identified.

Table 3.10. Level of agreement (Yules Q) on problems between methods by problem type.

Comparison	Semantic I (Structure)	Semantic II (Meaning)	Resp. Task I (Recall)	Resp. Task II (Response)	Other
QUAID/Expert Review	-.49*	-.42*			
QUAID/QAS	-.48*	-.82*			
QUAID/Cognitive interviewing	-.39*	-.60*			
Expert Review/QAS	.11	.56*	-.32	-.59*	-.96*
Expert Review/Cognitive interviewing	.25	.50*	.28	.18	-.85*
QAS/Cognitive Interviewing	-.60*	.61*	-.68*	-.56*	.26
*p < .05					

Comparison of original versus revised items

The previous sets of analyses provide insight into the types of problems that each method identifies. However, these analyses provide very little insight into the effectiveness of the methods at improving survey questions. Ideally one would want to know something about the accuracy of the survey questions to address this issue. For example, the researcher would like to have a direct measure of the reliability or validity of the survey questions such as a test-retest correlation or record check.

In the absence of direct reliability or validity evidence, researchers have taken other approaches to examine how effective the methods are at improving survey questions. One approach is to intentionally "damage" survey items with problems and evaluate the extent to which the methods detect these problems (e.g. Blair and Conrad, 2011). An alternative approach is to determine the extent to which the methods detect a difference between the original versions of a set of questions versus a revised version of the same questions. Fortunately the questionnaires used for this study included 33 items that were revised from their original form. The instrument was first cognitively tested and then the class along with the professor made revised versions of the questions where necessary. This allows for an analysis of the number of problems detected in both the original and revised form of these questions.

Table 3.11 shows the differences in the average number of problems detected for the 33 revised questions. The main question to address is whether or not the methods detect fewer problems with the revised versions of the questions than the original versions. In general, the methods do detect fewer problems with the revised questions. This is true for QUAID, expert review, and cognitive interviewing. However, the differences are statistically significant for only expert review. Expert review found fewer than half of the number of problems on the revised questionnaire compared to the original questionnaire. The QAS actually found more problems with the revised version of the questions than the original version of the questions.

Table 3.11. Differences in number of problems detected between original versions of questions and revised version of questions (n=33).

Method	Mean number of problems per question		Difference
	Original Questionnaire	Revised Questionnaire	
QUAID	2.91	2.76	.15
Expert review	3.21	1.55	1.66*
QAS	5.42	6.30	-.88
Cognitive interviewing	2.97	2.78	.19
SQP total quality	.58	.58	.00

*p<.05

The analysis in Table 3.11 provides only weak evidence that on average the methods can detect differences between an original and revised question. The table gives us our first look at complementary method hypothesis. It appears that a single method would be the best at identifying problematic questions as opposed to a combination. Only expert review could tell the difference between the original and revised questions. There are many caveats to this analysis though. First, there really is no independent evidence that the revised questions are actually of higher data quality than the original questions. Next, in light of Table 3.9 and 3.10 it is interesting that expert review shows the largest reduction in problem identification between the original questions and the revised questions. Those tables show that expert review is relatively proficient at finding structural issues with the questions. In other words, it finds problems that are "fixed" presumably by changing the syntax of the question. In contrast, problems identified by a method like cognitive interviewing are more likely to involve matters of interpretation or ambiguity that may or may not be as easily addressed with question wording. The results from QAS also align closely with what was found earlier in Table 3.1 and Table 3.5. Although QAS finds the most problems, there seems to be evidence that a number of these problems either are not significant problems or are not easy to fix. Hence, use of the QAS may be useful if the researcher is casting a wide net at the beginning of question design process; however, it may not be so useful at more advanced stage of question development where the research is making finer adjustments to question wording.

Prediction of the accuracy of survey questions

I also examined whether the problems can predict the accuracy of survey questions. In order to do this I searched the literature for survey questions with record checks. This analysis includes a total of 51 questions from 8 studies. Table 3.12 shows the correlations between method results overall and the proportion of respondents who were found to answer the question correctly from the literature. The table shows that the only types of results that are significantly correlated with the percent correct are problems identified about the sensitivity of the questions. A number of the record checks from the literature included questions with socially desirable or sensitive content.

Table 3.12. Correlation between the number of times that different types of problems were detected with items and the percent correct.

	Semantic I: structure	Semantic II: meaning	Resp. Task I: recall	Resp. Task II: resp.cat.	Resp. Task III: sensitivity	Int. Prob.	SQP Total Quality	% Correct
Semantic I	1	.04	-.16	.34	-.26	.00	.03	.21
Semantic II		1	-.24	.05	-.39*	-.14	-.08	.22
Resp. Task I			1	.00	.00	.19	.27	-.13
Resp. Task II				1	-.16	.00	.11	.10
Resp. Task III					1	-.10	-.06	-.55*
Interviewer						1	-.08	-.14
SQP							1	.03
% Correct								1
*p<.05								

I next investigated the complementary methods and test environment hypotheses by examining which methods are the best predictors of the percent correct. I include expert review QAS and cognitive interviewing in this analysis since they are the only methods that can detect this type of problem. The correlations between these variables are shown in Table 3.13. There is a moderate correlation between expert review and QAS. There is also a weaker correlation between QAS and cognitive interviewing. There is no correlation between expert review and cognitive interviewing.

Table 3.13. Correlation between the number of times that different methods detect sensitivity problems.

	Expert Review	QAS	Cognitive Interviewing	% Correct
Expert Review	1	.52**	.03	-.60**
QAS		1	.34**	-.32**
Cognitive Interviewing			1	-.26*
** p < .05, * p <.10				

I next conduct a regression analysis to test the complementary method hypothesis and the test environment hypothesis in Table 3.14. Model 1 is the full model that includes all three methods. The R-squared for this model is .41. The expert review and cognitive interviewing results are significant in the model. The next model shows that the R-squared does not change by dropping the QAS results. The final three models show how the individual methods performed on their own.

Table 3.14. Prediction of the percent correct with method results.

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	-.87**(.02)	.88**(.02)	.88**(.02)	.85**(.03)	.81**(.02)
Expert review	-.10**(.02)	-.09**(.02)	-.09**(.02)		
QAS	.02(.02)			-.05**(.02)	
Cognitive interviewing	-.28**(.12)	-.24**(.11)			-.26*(.14)
R-squared	.41	.41	.36	.10	.07
**p<.05, *p<.10					

I conducted generalized F testing to compare full and reduced regression models to determine if reducing to the model to the results from a single method significantly reduces the fit of the model (i.e. the complementary methods hypothesis). I then followed up by testing which combinations could be used in comparison to the full model. The F test comparing model 1 and model 2 result in a value of F_{47}^{1} = .74, which is less than the critical value of F at the .05 level (4.05). This means that dropping QAS from the model does not result in a significant reduction in the R-squared value. Next, I tested whether I could reduce the model further to include only the expert review results by testing model 2 against model 3. This results in a value of F_{47}^{1} = 4.78, which is greater that the critical value of F at the .05 level (4.04). This means that reducing to only the expert review results would lead to significant reduction in explanatory power. This lends support for the complementary methods hypothesis that it is better to use a combination of methods to better understand data quality. The ordering of the R-squared values clearly suggests that the expert review results were the best predictor of the percent correct. This is contrary to the test environment hypothesis.

Discussion

This chapter provided some evidence regarding the nature of the conclusions drawn about survey questions from some of the newer computer based methods such as QUAID and SQP. In support of the model-based method hypothesis, these methods do tend to find different problems than traditional methods such as expert review, QAS, and cognitive interviewing. As with the traditional methods, future research needs to investigate whether the problems identified by these methods are likely to cause issues in the field.

This chapter investigated some potential explanations for the typically low level of agreement between the findings from different question evaluation methods. One contributor to the level of agreement between methods is the nature of the problems that the methods detect (Yan, Kreuter, Tourangeau, 2012). Consistent with problem nature hypothesis, the findings from this study show that traditional methods like expert review, QAS, and cognitive interviewing had the strongest agreement on problems related to comprehension or interpretation of questions. This is somewhat reassuring given that these seem to be consistently the most prevalent problems found by question evaluation methods such as expert review, QAS, and cognitive interviewing (Presser and Blair, 1994; Rothgeb, Willis, and Forsythe, 2001; Willis, Schechter, and Whitaker, 1999).

However, agreement on these types of problems is still best characterized as moderate. QUAID and QAS tend to identify different types of problems than the other methods. If the goal in question evaluation is to cast a wide net, forms appraisal or QUAID in combination with another method may be the best option. This may ultimately lead to needing to filter through a number of mild problems though. The QAS is likely to lead to the most problems that the researcher will need to filter through when evaluating a question. It is likely that not all of the problems identified by QAS will be significant enough to warrant a change in question wording.

The remaining analyses in this chapter began to address the critical issue of whether the methods can differentiate between good and bad questions. QUAID, expert review, and cognitive interviewing did tend to find fewer problems with a revised set of survey questions compared to their original wording. Only the expert review found significantly fewer problems with the revised questions though. The QAS actually found more problems with the revised set of questions, perhaps calling into question the validity of the QAS results. However, it is difficult to tell from this data whether the problems found by expert review are relatively easier to find and fix than problems found by other methods. External data sources are needed to answer this question more thoroughly.

The final set of analyses did take advantage of external data sources regarding the accuracy of the survey questions. The results show that a combination of expert review and cognitive interviewing provides the best prediction of the percent of correct answers. These results are limited by the types of problems that are found with the record check questions in this chapter though. Many of the questions in this study were subject to social desirability bias due to the sensitive nature of their content. Future studies should consider whether this applies to a more diverse set of questions.

CHAPTER 4: AN ASSESSMENT OF THE LINK BETWEEN PROBLEM IDENTIFICATION AND DATA QUALITY: A CONFIRMATORY APPROACH USING INDIRECT DATA QUALITY INDICATORS

Introduction

It is a complex undertaking to understand the quality of the data produced by a survey question. Survey designers regularly rely on expert review and laboratory methods to assess the quality of a survey question. Several exploratory studies have examined the amount of agreement between different methods. These studies have often found that different methods of question evaluation lead to different conclusions about the quality of survey questions (Presser and Blair, 1994; Rothgeb, Willis, and Forsyth, 2001; Willis, Schechter, and Whitaker, 1999; Yan, Kreuter, and Tourangeau, 2012). Therefore, it is important to undertake confirmatory research to understand which methods produce results that are predictive of the quality of the data collected in the field. However, there is currently a dearth of research about how the conclusions from these methods relate to data quality in the field. There are only a few examples of studies that have explored this question (e.g. Blair et al., 2012; Forsyth, Rothgeb, and Willis, 2004; Willis and Schechter, 1997). So far these studies have not provided a clear picture of how close the link is between method results and what happens in the field. In addition, there have been some methodological shortcomings that have left gaps in the literature. The goal of this chapter is to further clarify the relationship between results from ex-ante and laboratory methods and how questions perform in the field.

This chapter undertakes a confirmatory approach to the method evaluation (Forsyth, Rothgeb, and Willis, 2004). The chapter investigates three hypotheses. The first hypothesis stems from the inconsistent agreement between methods in the literature. A frequent recommendation is that this inconsistent agreement indicates that the methods are best thought of as complementary and therefore it is better to use multiple methods together (Presser et al., 2004; Yan, Kreuter, and Tourangeau, 2012). I am calling this hypothesis the "complementary methods hypothesis."

The second hypothesis tested in this chapter is the "test environment hypothesis" which suggests an ordering of the ex-ante and laboratory methods according to how effective they should be at detecting data quality in the field. Methods that more closely observe the response process should have an advantage over methods that do not observe the process as closely. It has been argued that the survey response process is set within a sociocultural context and that laboratory techniques, such as cognitive interviewing, allow the researcher to observe this process in the context of the respondents' life circumstances (Gerber and Wellens, 1999; Miller, 2011). Experts cannot directly observe the process, but rather use their experience with previous research to predict which questions respondents might have difficulty with. The existing computer based systems have the least capability to account for sociocultural context.

The final hypothesis being investigated will provide insight about the circumstances under which ex-ante and laboratory methods provide useful results. There is much evidence that respondents with lower levels of cognitive ability, such as older respondents and those with lower levels of education, tend to have the most difficulty with survey questions (e.g. Krosnick, 1991; Knauper et al., 1997). Hence, it would be particularly helpful if pretesting methods could identify questions that are most likely to cause problems for these types of respondents. The “respondent and question problem interaction hypothesis” predicts that respondents with lower levels of cognitive ability will have more difficulty with questions identified as problematic by ex-ante and laboratory methods than respondents with higher levels of cognitive ability.

The next section reviews the methods that were used to evaluate the quality of the data from the field. Ideally, one would want to directly measure the reliability or validity of the survey questions; however, research designs for this type of assessment of data quality are often too expensive or infeasible for other reasons. Hence, researchers often use indirect measures to assess the quality of the questions in the field. This chapter utilizes behavior coding, response latency, and item nonresponse to measure data quality in the field.

Methods for evaluating questions in the field

In contrast to ex-ante and laboratory methods, other methods assess questions in a realistic field environment. Perhaps the ideal measurement of data quality in the field involves either a reinterview or record check study that can be used to assess the reliability or validity of survey questions. Research designs to directly assess reliability and validity can be cost prohibitive or impractical in some situations. Instead, it is often more feasible to collect proxy information about data quality. This chapter will examine the relationship between results from ex-ante methods or laboratory methods and three different proxy indicators of data quality in a field setting: behavior coding, item nonresponse, and response latency. I will now discuss their relevance to data quality.

Behavior coding is a method that has been use to understand the quality of the interaction between the interviewer and the respondent in the actual survey interview. It involves the coding of key observable behaviors that indicate a breakdown or potential issue with the question-answer process. The two most common indicators of the quality of survey questions with this method are the percentage of respondents who provide an adequate answer and the percentage of respondents who request clarification of the survey question. Although these behaviors are not always direct indicators of problems with questions, frequent deviations from the ideal question-answer process indicate the potential for problems with data quality. Hess, Singer, and Bushery (1999) found that behavior codes were significant predictors of the reliability of questions. Hence there certainly is evidence that frequent aberrant behavior in the survey interview indicates the potential for poor data quality.

Another potentially useful indicator of data quality is response latency or a measure of how long it takes respondents to answer the question. Like behavior coding, response latencies provide a quantitative assessment of the amount of difficulty that respondents are having with a question. One assumption behind the use of response latencies is that response latency is an indicator of the amount of information processing required to answer a question. This includes the amount of time that it takes a respondent to comprehend a question, retrieve information from memory, integrate that information into a summary judgment, and select a response option. A second assumption is that problems with a question lead to slower response times, because resolution of the problem requires processing time (Basilli and Scott, 1996). Draisma and Dijkstra (2004) provide evidence that longer response latencies are more likely to result in inaccurate responses.

Finally, item nonresponse is one of the most widely used indicators of data quality in the field. There is a plethora of research examining the causes of item nonresponse in the field. One key aspect of a question that affects the level of item nonresponse is the explicitness of the don't know filter (Schuman and Presser 1981). For example, in interviewer administered surveys, item noresponse rates will be higher when the don't know category is read as one of the response options. Beatty and Hermann (2002) proposed a model for item nonresponse that is driven by the cognitive state (i.e. how much the respondent knows) and communicative intent (i.e. what the respondent wants to reveal about herself) of the respondent. Research suggests that question sensitivity and the cognitive effort needed to answer survey questions seem to be two of the most important determinants of item nonresponse (Pickery and Loosveldt, 2001; Shoemaker et al 2002). Respondent characteristics also seem to be related to item nonresponse. Krosnick's theory of survey satisficing suggests that survey respondents are cognitive misers and may take shortcuts, such as don't know responding, when given the opportunity (Krosnick, 1991). The theory posits that those with less cognitive ability are more likely to take these shortcuts. For example, less educated respondents are also more prone to item nonresponse (Schuman and Presser, 1981). There is also evidence that item nonresponse does appear to be higher for the elderly population (Colsher and Wallace, 1989).

Analysis

Descriptive Statistics

Table 4.1 summarizes the primary dependent variables in the analyses to follow. The primary focus of the models is to understand how well the results from the ex-ante (e.g. QUAID, SQP, Expert Review, and QAS) and laboratory (e.g. cognitive interviewing) methods predict the results in the field, as measured by behavior coding, item nonresponse, and response latency measures. Descriptive statistics for these dependent variables are shown in Table 4.1.

Table 4.1. Mean and standard deviation for dependent variables in the models.

Variable	n	Mean	Standard Deviation	Minimum	Maximum
Adequate answers	17,666	.78	.41	0	1
Requests for clarification	17,666	.06	.25	0	1
Item nonresponse	34,955	.03	.18	0	1
Log response latency (milliseconds)	4,815	7.36	1.29	.69	12.19

Many of the independent variables in the model refer to the number of times that the methods discovered different types of problems from the Presser-Blair coding scheme (see Table 2.1) across the 88 questions that were administered in the final practicum questionnaire. Descriptive statistics for each type of problem are shown in Table 4.2.

Table 4.2. Mean and standard deviation for primary predictor variables in the models (n = 88 questions).

Variable	Mean	Standard Deviation	Minimum	Maximum
Semantic I problems: question structure				
Overall	1.70	1.63	0	7
QUAID	.57	.56	0	2
Expert Review	.35	.68	0	3
QAS	.67	.96	0	4
Cognitive Interviewing	.11	.38	0	2
Semantic II problems: meaning				
Overall	7.56	4.04	0	17
QUAID	1.70	1.28	0	5
Expert Review	.76	.87	0	3
QAS	3.45	1.91	0	8
Cognitive Interviewing	1.64	1.75	0	8
Respondent task I problems: recall				
Overall	2.86	2.89	0	13
QUAID	0	0	0	0
Expert Review	.38	.57	0	2
QAS	1.70	1.69	0	6
Cognitive Interviewing	.75	1.52	0	7
Respondent task II problems: response categories				
Overall	1.16	1.63	0	5
QUAID	.18	.47	0	3
Expert Review	.09	.33	0	2
QAS	.65	.98	0	3
Cognitive Interviewing	.24	.64	0	3
Respondent task III problems: sensitivity				
Overall	1.11	1.16	0	5
QUAID	0	0	0	0
Expert Review	.09	.33	0	2
QAS	.65	.98	0	3
Cognitive Interviewing	.24	.64	0	3
Analysis problems				
Overall	.13	.42	0	3
Interviewer problems				
Overall	.05	.21	0	1
Other methods				
SQP Total Quality Score	.57	.06	.46	.71

Finally, the models to follow will include control variables at the question and respondent level. Although previous method evaluations involve a sampling of questions, these studies have generally not paid attention to the characteristics of the questions. Including these variables in the models allows for an understanding of how robust the findings are to question characteristics. Respondent characteristics are included for the same reason, but also to explore whether the method results differ for respondents with different characteristics. Descriptive statistics for the question and respondent characteristics are shown in Table 4.3.

Table 4.3 Distribution of question and respondent characteristics.

Question Characteristic (n=88)	n	Percent
Reading Grade Level		
< 8	25	28.4
8 – 9.9	21	23.9
10 – 11.9	18	20.5
12 +	24	27.3
Question Type		
Factual	29	33.0
Subjective	52	59.1
Behavioral Frequency	7	7.9
Response Format		
Numerical	8	9.1
Yes/No	16	18.2
Verbal Label	64	72.7
Respondent Characteristic (n=709)		
Education		
High school or less	233	32.5
More than high school	506	68.5
Age		
Under 60	446	60.3
60 or older	293	39.7

Question Level Model Results

This chapter utilizes a few different dependent variables as indicators of data quality in the field. The ex-ante and laboratory methods can detect the different types of problems shown in Table 4.2. I chose to examine the relationship between different types of problems and the results in the field rather than combining all of these different types of problems together into a summary measure for each method. This decision is driven by the results from the literature and from chapter three of this dissertation suggesting that the methods have different abilities to detect certain types of problems. Hence, using the more specific categories of problems within methods provides the best opportunity to elucidate any differences between the methods in terms of their ability to detect problems that arise in the field. In addition, this approach is consistent with research on behavior coding by Holbrook, Cho, and Johnson (2006) that found that different behavior codes are better at capturing problems with different parts of the response process. Some behavior codes such as requests for clarification are better at capturing comprehension problems and other behavior codes such as inadequate answers are better capturing problems with mapping.

I began with some expectations about how the different classes of problems would map on to different indicators of data quality in the field. I model two different types of behavior codes: requests for clarification and adequate answers. The semantic I problems (question structure) and semantic II problems (question meaning) should be indications of a respondent’s ability to determine the focus of a question or determine the meaning of the words in a question. Therefore, I expect questions with more frequent problems of this nature to lead to a higher percentage of requests for clarification. In contrast, I expected the respondent task problems to be more predictive of adequate answers. Respondent task I problems (recall) occur when the respondent does not have enough information to answer the question. Therefore, the respondent may provide an initial response that is not adequate to answer the question. Respondent task II problems (response categories) occur when the response categories are inadequate or overlapping. Therefore the respondent may have a difficult time selecting among response categories. Finally, respondent task III problems (sensitivity) are likely to result in refusals to answer questions that the respondent might deem sensitive. I expect respondent task problems to also be predictive of item nonresponse for the same reasons that these types of problems are predictive of adequate answers.

Finally, I expected both semantic and respondent task problems to be more predictive of response latencies. I initially expected all types of semantic problems are likely to lead to confusion for the respondent, which in turn will lead to long pauses or requests for clarification, which will result in longer response times. I expect recall problems (respondent task I) to result in longer response latencies as the respondent searches for relevant information to answer the question.

I ran some preliminary regression models at the question level to evaluate which types of problems are predictive of the different types of results from the field. The Pearson correlations between the variables used in the analysis are shown in Table 4.4.

Table 4.4. Pearson correlations between variables in question level analysis (n=88).

					Semantic problems		Respondent task problems						Response format			Question type		
	Log % ad. answ.	Log % req. clar.	Log % item NR	Log response latency	Sem. 1 Structure	Sem. II Meaning	Task I Recall	Task II Resp. Cat.	Task III Sensitivity	Int. Problems	Analysis problems	Read level	Numerical	Yes/no	Verbal	Factual	Subjective	Beh.Freq
Log % ad. answ.	1	-.50*	-.46*	-.54*	.05	.03	-.60*	-.13	-.05	.16	.02	-.12	-.17	.26*	-.12	.53*	-.31*	-.36*
Log % req. clar.		1	.28*	.67*	.03	.28*	.41*	.09	-.01	-.18	.04	-.04	.23*	-.25*	.07	-.38*	.21*	.28*
Log % item NR			1	.48*	-.15	.13	.41*	.23*	.00	-.11	.10	.07	-.13	.04	.05	-.27*	.34*	-.17
Log resp. Lat.				1	.02	.20	.48*	.17	-.16	-.16	.14	.07	.14	-.24*	.12	-.34*	.20	.21*
Sem. I					1	.43*	.20	.02	-.11	.07	.15	.51*	-.26*	-.15	.30*	.13	-.12	.00
Sem. II						1	.25*	.04	.03	-.06	.15	.25	-.36*	.24*	.02	.04	.05	-.16
Task I							1	.21	-.05	-.08	.19	.14	.04	-.08	.04	-.37*	.11	.44*
Task II								1	.08	-.02	-.08	-.06	-.23*	-.30*	.41*	-.29*	.23*	.10
Task III									1	-.02	.02	-.17	-.10	.29*	-.18	.01	-.08	.12
Int, prob										1	-.06	.13	-.07	.18	-.11	.20	-.15	-.06
Analysis prob.											1	.17	-.09	.14	-.06	.02	.03	-.09
Read level												1	-.28*	.03	.15	.15	-.09	-.08
Numerical													1	-.15	-.52*	.11	-.38*	.49*
Yes/no														1	-.77*	.48*	-.39*	-.14
Verbal															1	-.49*	.58*	-.20
Factual																1	-.84*	-.21
Subjective																	1	-.35*
Beh.Freq.																		1
*p < .05																		

The goal of these preliminary models was to test my initial expectations about how the different types of problems identified by QUAID, expert review, QAS, and cognitive interviewing would be related to each dependent variable. The dependent variables in the analysis have been log transformed to create more normal distributions. For each dependent variable in Table 4.5 I show the full model including all of the different types of problems with controls for question characteristics and I also show a reduced model with the problem types that are the best predictors of each dependent variable with the control variables.

The results from Table 4.5 are largely consistent with my expectations with a few exceptions. Respondent task I problems associated with recall and respondent task II problems associated with sensitivity are predictive of the percentage of adequate answers. However, problems with response categories are not related to the percentage of adequate answers. Only semantic II problems associated with the meaning of terms or concepts are significant predictors of requests for clarification. Semantic I issues associated with question structure are not significantly related to requests for clarification. Respondent task I problems associated with recall and respondent task II problems associated with response categories are significant predictors of item nonresponse. Respondent task III problems related to sensitivity were not predictive of item nonresponse. Semantic II problems related to the meaning of words and concepts and respondent task I problems related to recall were significant predictors of response latencies. Generalized F tests comparing the full and reduced models for each dependent variable were not significant indicating that the models could be reduced to reduced set of predictors without losing significant predictive power.

Table 4.5. Prediction of field results with different types of problems at the question level (n=88).

	Dependent variable			
Predictor	Log % adequate answers	Log % requests for clarification	Log % item nonresponse	Log response latency
Full model				
Intercept	4.70*(.070)	.04(.36)	-.21(.50)	6.75(.37)
Semantic I: quest. structure	.017(.012)	-.060(.062)	-.165(.084)	-.09(.06)
Semantic II: meaning	.003(.004)	.092*(.022)	-.007(.030)	.053*(.023)
Respondent task I: recall	-.033*(.006)	.05(.03)	.222*(.045)	.087*(.033)
Respondent task II: resp. cat.	.0055(.0096)	-.015(.050)	.138*(.068)	.036(.050)
Respondent task III: sensitivity	-.029*(.013)	.055(.067)	.058(.091)	-.051(.068)
Interviewer problems	.048(.067)	-.04(.35)	-.40(.48)	-.147(.354)
Analysis problems	.05(.03)	.03(.18)	-.099(.24)	.165(.178)
Reading grade level	-.015*(.004)	.008(.022)	.050(.030)	.024(.022)
Factual questions	Ref.	Ref.	Ref.	Ref.
Subjective questions	-.15*(.04)	.45*(.22)	.39(.29)	.23(.22)
Behavioral frequency questions	-.09(.07)	.45(.39)	-1.80*(.53)	.12(.39)
Yes/no response format	Ref.	Ref.	Ref.	Ref.
Numerical response format	-.18*(.07)	1.48(.38)	.23(.51)	.97*(.38)
Verbal response format	-.03(.05)	.45(.28)	-.41(.38)	.35(.28)
R-squared	.61	.44	.45	.38
Reduced model				
Intercept	4.72*(.070)	.19*(.30)	-.10(.37)	6.76*(.31)
Semantic I: quest. structure				
Semantic II: meaning		.093*(.022)		.041*(.020)
Respondent task I: recall	-.027*(.006)		.190*(.041)	.094*(.031)
Respondent task II: resp. cat.			.161*(.068)	
Respondent task III: sensitivity	-.027*(.013)			
Interviewer problems				
Analysis problems				
Reading grade level	-.012*(.004)	.004(.019)	.020(.028)	.013(.02)
Factual questions	Ref.	Ref.	Ref.	Ref.
Subjective questions	-.18*(.04)	.65*(.18)	.66*(.27)	.34(.20)
Behavioral frequency questions	-.12(.07)	.82*(.30)	-1.57*(.50)	.06(.37)
Yes/no response format	Ref.	Ref.	Ref.	Ref.
Numerical response format	-.22*(.07)	1.35*(.34)	.28(.45)	1.00*(.35)
Verbal response format	-.01(.04)	.24(.21)	-.70*(.31)	.27(.22)
R-squared	.57	.41	.41	.34
F-test (Full Model-Reduced)	n.s.	n.s.	n.s.	n.s.
*p<.05				

I also conducted question level analyses to determine how successful the individual methods were at predicting questions that were among the top 25% most problematic questions according to the data quality indicators used in this chapter. Table 4.6 shows the results with problematic behaviors as the dependent variables.

The findings in Table 4.6 differ slightly depending on which behavior coding results are being predicted. First, I will discuss the results predicting the questions with the highest percentages of inadequate answers. I initially began modeling this percentage by using the respondent task I and respondent task III problems as suggested by the previous analyses in Table 4.5. However, the respondent task III problems related to the sensitivity of the questions were not predictive of the most problematic questions for any of the methods. Hence I dropped these problems from the models in Table 4.6. The results from the table show that respondent task I problems related to recall found by the QAS and cognitive interviewing were predictive of the questions with the highest amount of inadequate answers. Model 5 uses the results from both the QAS and cognitive interviewing to predict this outcome. Likelihood ratio tests with one degree of freedom comparing the deviance statistics from Model 5 to Model 2 (QAS results only) and Model 3 (cognitive interviewing results only) reveals that using the cognitive interviewing results alone does not result in a significant reduction in fit compared to using the results from both methods.[10] This provides counter evidence to the complementary methods hypothesis and some support to the test environment hypothesis.

[10] Chi-square$_{.05}$, 1 DF = 3.84

Table 4.6. Prediction of most problematic questions according to behavior coding results (n=88).

Adequate Answers						
	Model 1	Model 2	Model 3	Model 4	Model 5	
Intercept	-4.94**(1.77)	-4.51**(1.59)	-3.11*(1.20)	-3.71(3.65)	-4.25**(1.57)	
Respondent task I problems: recall						
Expert review	.91(.63)					
QAS		.31*(.18)			.15(.21)	
Cognitive interviewing			.60**(.23)		.53**(.26)	
SQP Total Quality				-.01(.07)		
Deviance	75.81	74.89	71.48	77.93	69.79	
AIC	89.81	88.89	83.48	91.93	86.79	
Requests for clarification						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	-4.22**(1.33)	-5.58**(1.59)	-6.39**(1.79)	-4.27**(1.30)	-8.67**(3.84)	-6.56(1.94)
Semantic II problems: meaning						
QUAID	.28(.22)					
Expert review		.93**(.36)				.48(.42)
QAS			.49**(.20)			.27(.22)
Cognitive interviewing				.49**(.17)		.37*(.19)
SQP Total Quality					.09(.06)	
Deviance	84.01	78.32	78.73	76.24	83.52	72.14
AIC	98.01	92.32	92.73	90.24	97.52	90.14

*p<.10, **p<.05

Note. All models include controls for reading level, question type, and response format.

There is a slightly different story when examining the prediction of the most problematic questions according the percentage of requests for clarification. The bottom panel of Table 4.6 shows that semantic II problems related to question meaning identified by expert review, QAS, and cognitive interviewing are predictive of the most problematic questions according to the percentage of requests for clarification. Notice that the AIC values for models 2-4 are very similar. This makes sense, because the correlations between the results for these methods are in the .4 to .5 range (see Table 4.10). This collinearity is demonstrated in model 6 that includes all of these methods. Only the cognitive interviewing results are marginally significant when all three methods are in the same model. Likelihood ratio tests with two degrees of freedom suggest that the cognitive interviewing results alone could be used to predict the most problematic questions, but the results from expert review and the QAS fall just outside the critical value of 5.99. Given all of the information in Table 4.6 though, it is hard to conclude that any one of the methods is better at predicting the questions with the highest percentage of requests for clarification. In addition, the sizeable correlations between the methods suggest that the methods are probably substitutable for one another. Hence, these findings provide counter evidence for the complementary methods hypothesis and weak support for the test environment hypothesis.

Prediction of the most problematic questions according to item nonresponse is shown in the top panel of Table 4.7. The table shows that respondent task I problems related to recall identified by the QAS and cognitive interviewing are predictive of the most problematic questions. Similar to the results for inadequate answers, comparison of the deviance statistics from model 5 including the results for both of these methods with model 2 including the results for only the QAS and model 3 including only the results for cognitive interviewing leads one to conclude that using the cognitive interviewing results alone does not significantly decrease the fit of the model.

Table 4.7. Prediction of most problematic questions according to item nonresponse and response latency (n= 88).

Item nonresponse						
	Model 1	Model 2	Model 3	Model 4	Model 5	
Intercept	-3.30**(1.50)	-3.85**(1.59)	-2.46*(1.63)	-9.56(3.82)	-2.75**(1.68)	
Respondent task I problems: recall						
Expert review	-.11(.63)					
QAS		.49*(.18)			.28(.20)	
Cognitive interviewing			1.28**(.47)		1.07**(.48)	
SQP Total Quality				-.01(.07)		
Deviance	80.84	72.51	64.77	77.06	62.89	
AIC	94.84	86.51	78.77	91.06	78.89	
Response latency						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	-2.67**(1.15)	-3.84**(1.36)	-3.43**(1.50)	-2.73**(1.14)	-10.6**(4.1)	-3.17(1.57)
Semantic II problems: meaning						
QUAID	.12(.22)					
Expert review		.83**(.35)				.21(.47)
QAS			.07(.19)			-.14(.22)
Cognitive interviewing				.39**(.16)		.40*(.22)
Respondent task I problems: recall						
Expert review		-.32(.59)				.24(.74)
QAS			.55**(.19)			.52**(.22)
Cognitive interviewing				.16(.19)		.00(.22)
SQP Total Quality					.14**(.07)	
Deviance	90.62	84.74	79.87	83.09	85.20	74.72
AIC	104.62	100.74	95.87	99.09	99.20	98.72

*p<.10, **p<.05

Note. All models include controls for reading level, question type, and response format.

Hence the results for item nonresponse also provide counter evidence to the complementary methods hypothesis and support for the test environment hypothesis.

Finally, the bottom panel of Table 4.7 illustrates how well the methods predict the most problematic questions according to response latency. Semantic II problems related to question meaning identified by expert review and cognitive interviewing are predictive of questions with the longest response latencies. Respondent task I problems related to recall identified by the QAS are also predictive of questions with the longest response latencies. I compared model 6 with models 2-4 separately using a likelihood ratio statistic with four degrees of freedom. This test compares the results for expert review alone, QAS alone, and cognitive interviewing alone to the results using all of the methods combined. The results of the likelihood ratio conclude that reducing to either the QAS or cognitive interviewing results do not significantly reduce the fit of the model. I then conducted some follow-up tests since the results from the expert review are not significant in the full model. The results of these follow-up tests are shown in Table 4.8.

Table 4.8. Follow-up tests predicting questions with longest response latencies (n = 88).

	Model 1	Model 2	Model 3
Intercept	-2.80**(1.14)	-3.15**(1.25)	-3.33**(1.25)
Semantic II problems: meaning			
Cognitive interviewing	.40**(.16)		.37**(.18)
Respondent task I problems: recall			
QAS		.57**(.19)	.53**(.19)
Deviance	83.79	79.99	75.32
AIC	97.79	93.99	91.32

**p<.05
Note. All models include controls for reading level, question type, and response format.

The likelihood ratio statistics with 1 degree of freedom comparing model 3 including the results from both the QAS and cognitive interviewing with either model 1 including only the cognitive interviewing results for question meaning problems or model 2 including only the QAS results for recall problems show that there is a significant reduction of fit when either results are not included in the model. These findings support the complementary methods hypothesis. However, the AIC values (lower values being better) in Table 4.7 and Table 4.8 do not support the test environment hypothesis. The AIC values suggest that the QAS results are more predictive than the cognitive interviewing results.

The above regression models provide insight about which categories of problems are predictive of indicators of data quality in the field. In the next section of this chapter I use multilevel models to test three hypotheses in this dissertation. Although the question level regression models are a useful starting point there are some limitations to this type of analysis that can be addressed by multilevel models. First, the distributions of the dependent variables from Table 4.5 are skewed. Hence, the models are improved by transforming the dependent variable with the log transformation to create more normally distributed dependent variables. Second, the threshold analysis in Tables 4.6 – 4.8 give the reader an idea of which methods may predict the most seriously flawed questions. However, the choice of these thresholds is somewhat arbitrary. In reality, the results from the field methods are the result of specific question-respondent exchanges. For example, for behavior coding and item nonresponse, one can construct a 0 or 1 indicator for whether or not a problem occurred at a specific question-respondent exchange. By using multilevel models that model responses rather than summaries of responses, I can use logistic regression techniques that are appropriate for dichotomous dependent variables such as these. No transformation of the dependent variable is needed. In addition, the multivevel model framework enables testing of cross-level interactions. This is important for testing the question problem and respondent interaction hypothesis.

Multilevel Model Results

This section presents the results of the multilevel models. The results will be presented first for the prediction of behavior coding results, next for item nonresponse, and last for response latency. I show the correlations between these variables in Table 4.9. There is a substantial correlation between adequate answers and item nonresponse. There are weaker correlations between the other variables. Although there are correlations between these measures, as we would expect if they are all measuring data quality, most of the correlations are far from perfect. This indicates that these measures probably tap different dimensions of data quality. I include item nonresponse in this dissertation because it is a commonly used and available measure of data quality. However, these correlations suggest that the results should be similar for adequate answers and item nonresponse.

Table 4.9. Correlation between dependent variables.

	Adequate answer	Request for clarification	Item nonresponse	Log response latency
Adequate answer	1	-.12*	-.33*	-.56*
Request for clarification		1	.06*	.43*
Item nonresponse			1	.19*
Log response latency				1
*p < .05				

Table 4.10. Pearson correlations between question level predictor (n = 88 questions).

	Semantic II problems: Meaning				Respondent task I problems: Recall			Respondent task II problems: response categories				Respondent task II problems: sensitivity			SQP		Response format			Question type		
	QUAID	ER	QAS	CI	ER	QAS	CI	QD	ER	QAS	CI	ER	QAS	CI	Tot. Qual.	Read level	Num.	Yes/no	Verbal	Factual	Subjective	Beh.Freq.
Sem II QD	1	.09	.14	.04	.23*	.08	.19	-.14	.15	-.17	-.04	-.10	-.09	.23*	.03	.18	.04	.04	-.06	.14	-.23	.17
Sem II Er		1	.48*	.48*	.00	.40*	-.10	.11	-.13	.18	-.02	.00	.05	-.04	.13	-.05	-.23*	.23*	-.05	.00	.14	-.26*
Sem II QAS			1	.43*	.13	.24*	-.02	.07	-.23*	-.07	-.05	.04	.03	.03	.19	.28*	-.43*	.32*	.00	.14	-.04	-.16
Sem II CI				1	-.26*	.31*	.10	.23*	-.18	.37*	-.05	-.02	-.01	.03	.12	.18	-.27*	.06	.12	-.16	.25*	-.18
RT I ER					1	.13	.23*	-.09	.00	-.35*	-.09	-.06	.10	-.03	.10	.18	.28*	.00	-.18	.05	-.34*	.54*
RT I QAS						1	.43*	.17	-.05	.30*	.14	-.20	.03	-.14	.34	.14	-.06	-.02	.06	-.34*	.25*	.13
RT I CI							1	.15	.29*	.00	.21	-.12	.00	.08	.08	.04	.05	-.13	.08	-.35*	.06	.49*
RT II QD								1	.27*	.42*	.12	.12	.15	-.05	.02	-.06	-.12	-.18	.24*	-.27*	.13	.25*
RT II ER									1	.07	.33*	.14	.11	.02	-.08	-.23	-.09	-.13	.17	.03	-.19	.31*
RT II QAS										1	.17	-.04	.19	-.22*	.13	.00	-.21	-.28*	.38*	-.27*	.34*	-.15
RT II CI											1	-.10	.04	.08	.13	.01	-.12	-.13	.19	-.15	.06	.15
RT III ER												1	.29*	-.10	-.39*	-.24*	-.09	-.13	.17	-.20	.23*	-.08
RT III QAS													1	.21	-.29*	-.13	-.13	.11	-.01	-.08	.04	.07
RT III CI														1	-.32*	-.02	.03	.51*	-.45*	.26*	-.35*	.19
SQP															1	.36*	.15	-.03	-.07	.34*	-.32*	-.01
Read																1	-.28*	.03	.15	.15	-.09	-.08
Num.																	1	-.15	-.52*	.11	-.38*	.49*
Yes/no																		1	-.77*	.48*	-.39*	-.14
Verbal																			1	-.49*	.58*	-.20
Factual																				1	-.84*	-.21
Subjective																					1	-.35*
Beh.Freq.																						1
*P < .05																						

The correlations between the question characteristics that are included as predictor variables are presented in Table 4.10. This includes the method results and question characteristics. The correlation between the respondent level predictors of age and education is .11.

Next, I present several models for each dependent variable. Each table of results includes sections for fixed effects, random effects, and model fit statistics. I will discuss these sections of the tables to evaluate the hypotheses in this dissertation.

Behavior coding results

Table 4.11 presents eight models to summarize how closely the ex-ante and laboratory methods predict the probability of adequate answers. The first set of models presented assist in understanding how much of the variability in adequate answers is explained by the method results. Model 1 in Table 4.11 represents the null or empty model which provides baseline information for the amount of variance attributable to the questions and respondents. Model 2 introduces fixed effects for the results for the number of respondent task problems detected by each ex-ante and laboratory method. Model 3 introduces fixed effects for question characteristics. This model demonstrates the marginal contribution of each method toward the predication of adequate answers when the methods are used together. Model 4 introduces fixed effects for respondent characteristics. Controlling for different question and respondent characteristics provides the reader with a sense for how robust the findings are to the composition of the questions and the respondents who answered the questions. One can also compare the random effects across these four models to understand the contribution of each block of predictor variables towards explaining the variability in adequate answers.

Table 4.11. Prediction of adequate answers (0 = not adequate; 1 = adequate).

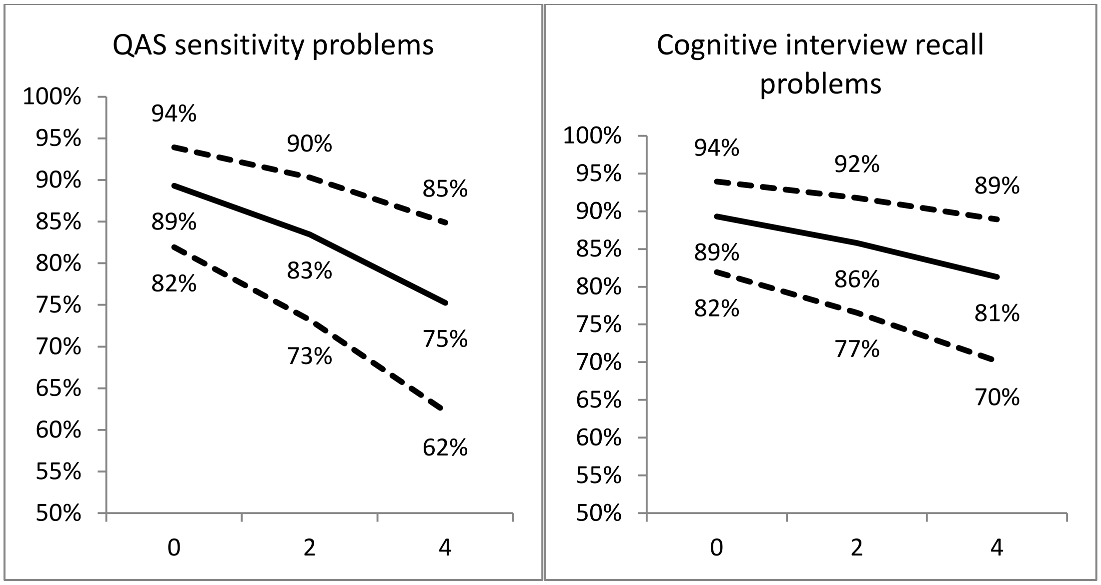
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Effect	Null	Methods	Question characteristics	Question and respondent characteristics	Expert review only	QAS only	Cognitive Interviewing only	SQP only
Fixed Effects								
Intercept	1.66* (.11)	.40(.91)	3.45*(.26)	4.29*(.95)	3.54*(.27)	3.88*(.27)	3.38*(.27)	5.03*(.76)
Respondent Task I problems: recall								
Expert Review		-.12(.13)	-.14(.13)	-.14(.13)	-.25(.14)			
QAS		-.15*(.05)	-.03(.05)	-.03(.05)		-.12*(.04)		
Cog.Int		-.24*(.05)	-.16*(.05)	-.16*(.05)			-.17*(.05)	
Respondent Task III problems: sensitivity								
Expert Review		.18(.27)	.20(.22)	.20(.22)	.22(.22)			
QAS		-.21*(.10)	-.27*(.08)	-.27*(.08)		-.20*(.08)		
Cog.Int		.56*(.15)	.19(.15)	.19(.15)			.20(.14)	
SQP Total Quality		.031*(.015)	-.010(.010)	-.011(.016)				-.03*(.01)
Grade level			-.05*(.02)	-.05*(.02)	-.05*(.02)	-.06*(.02)	-.06*(.02)	-.05*(.02)
Yes/No			Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Numeric			-.76*(.30)	-.76*(.30)	-.54(.30)	-.81*(.30)	-.58(.31)	-.41(.31)
Verbal			-.07(.20)	-.07(.20)	-.11(.20)	-.24(.20)	.03(.21)	-.01(.20)
Factual			Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Subjective			-1.17*(.20)	-1.17*(.20)	-1.37*(.17)	-1.11*(.18)	-1.16*(.17)	-1.45*(.18)
Beh.Freq			-.84*(.34)	-.84*(.34)	-1.40*(.32)	-1.31*(.28)	-1.02*(.33)	-1.79*(.29)
>H.S.				Ref.	Ref.	Ref.	Ref.	Ref.
H.S.				-.32*(.11)	-.32*(.11)	-.32*(.11)	-.32*(.11)	-.32*(.11)
Under age 65				Ref.	Ref.	Ref.	Ref.	Ref.
65 or older				-.37*(.11)	-.37*(.11)	-.36*(.11)	-.37*(.11)	-.37*(.11)
Random Effects								
Question level	.76*(.12)	.39*(.07)	.22*(.04)	.22*(.05)	.31*(.05)	.27*(.05)	.27*(.05)	.31*(.12)
Respondent level	.89*(.09)	.89*(.08)	.89*(.08)	.83*(.08)	.83*(.08)	.83*(.08)	.83*(.08)	.83*(.09)
Model Fit								
Deviance	15898.00	15843.70	15798.85	15776.61	15801.80	15791.88	15791.42	15801.22
AIC	15904.00	15863.70	15828.85	15810.61	15825.80	15815.88	15815.42	15823.22

* p < .05. n = 17666.

We begin by examining Model 2. The fixed effects demonstrate that QAS and cognitive interviewing results are predictive of the likelihood of obtaining an adequate answer. We can compare the random effects between model 2 and model 1 to understand how much of the variability in adequate answers is explained by the method results. The random effects from Model 2 demonstrate that approximately 49% ((.76-.39)/.76) of the variance in adequate answers at the question level is explained by the introduction of fixed effects for the ex-ante and laboratory results. Model 3 introduces question characteristics. After controlling for question characteristics, the number of times that recall problems are discovered by cognitive interviewing is still predictive of the likelihood of an adequate answer; however, the QAS problems are no longer predictive of adequate answers. The correlations in Table 4.10 show that recall problems are less likely to be identified by QAS on factual questions and more likely on subjective questions. There is also a significant positive correlation between the number of times that recall problems are identified by QAS and the number of times that these same problems are identified by cognitive interviewing. This pattern of relationships explains a significant component of the relationship between recall problems identified by QAS and adequate answers. Overall, this provides evidence that the cognitive interviews provide the most robust findings with respect to recall problems.

The relationship between the number of times that sensitivity problems are identified by cognitive interviewing and adequate answers is explained in large part by question type. The cognitive interviews were more likely find problems with factual questions than other types of questions. More specifically, six out of the eleven questions that the cognitive interviews identified as sensitive were related to the respondents' reports about their own health conditions. In contrast, the QAS found sensitivity problems at nearly the same rate across all types of questions. Looking at the respondent characteristics, as expected, respondents with lower education and older respondents are less likely to provide adequate answers.

The complementary methods hypothesis and the test environment hypothesis can be tested more formally by comparing the fit of models that include only the results from each individual method. The last four models of Table 4.11 include these results. Model 6 and Model 7 demonstrate that the results for cognitive interviewing and the QAS are predictive of adequate answers. The models suggest that the cognitive interviews provide the best assessment of recall problems. In contrast, the models clearly suggest that the QAS provided the most robust evaluation of the sensitivity of these survey questions. Figure 4.1 plots the predicted probability of an adequate answer given the number of times these different types of problems are detected.


QAS sensitivity problems

Cognitive interview recall problems

Figure 4.1. Predicted probability of adequate answers given the number of times that problems are detected (solid line). Dashed lines represent 95% confidence intervals around the estimates.

Model 8 from Table 4.11 shows that the total quality scores from SQP is related to the percentage of adequate answers; however, the relationship is in the opposite from expected direction. This is consistent with findings van der Zouwen and Smit (2004) that the SQP tends to find different types of problems than are found by behavior coding.

In order to formally test the complementary methods hypothesis we can compare the deviance statistics from the model including all of the methods (model 4 in Table 4.11) with the deviance statistic from each of the models including the individual methods (models 5-8 in Table 4.11). A likelihood ratio test can then be performed to understand whether using only one of the methods to predict adequate answers (i.e. removing all other method results from the model) results in a significant difference in model fits. In other words, a significant likelihood ratio test implies that the reduced model results in a significantly worse fit compared to the more complete model. Table 4.12 shows the results.

Table 4.12. Test of difference in model fit (Deviance) between the full model (D_0) and models including individual methods or combinations of methods (D_1) predicting adequate answers.

Method	Difference (D_1-D_0)	Critical Value	Significant difference?
SQP	24.61	11.07	Yes
Expert review	25.19	11.07	Yes
QAS	15.27	11.07	Yes
Cognitive interviewing	14.81	11.07	Yes
QAS and cognitive interviewing	2.26	7.81	No

The results of the likelihood ratio tests reveal that using any of the methods individually will significantly reduce the fit of the model. A combination of QAS and cognitive interviewing does significantly predict the likelihood of an adequate answer. This is shown by the last row of Table 4.12. The pattern of the coefficients suggests that the two methods have particular strengths that work well in combination when predicting adequate answers. The QAS is a better indicator of problems with sensitivity and cognitive interviewing provides better indications of problems with recall. That is, in this case it would be better to use the combination of QAS and cognitive interviewing to predict the likelihood of adequate answers in the field than any single method.

One can also see a general trend in favor of the test environment hypothesis in Table 4.11. However, a different measure of model fit must be used in order to examine the hypothesis more closely. Although the deviance statistic is appropriate for testing nested models, it is not appropriate for testing non-nested models. Fortunately the AIC model fit statistic can be used for this purpose. The AIC is a model fit statistic that is an estimate of the predictive accuracy for a model. Similar to the deviance statistic, a smaller value for the AIC represents a better model. An important property of the AIC is that it penalizes the model for including additional predictors. In fact, the AIC is equal to the deviance for a model plus two times the number of parameters estimated. Although one cannot conduct significance tests with the AIC, it can be used to rank order models in terms of their distance from the “true” model. Anderson (2008) provides guidelines for differences in AIC values that represent appreciable differences. He suggests that differences of 4.0 in AIC values between models represent a strong difference and greater than 8.0 are considered very strong differences. Rank ordering the individual methods in Table 4.11 provides a pattern that is somewhat consistent with Hypothesis 4. There are somewhat large differences in values of the AIC between cognitive interviewing and the remainder of the methods. However, there is almost no difference between expert review and the SQP.

The results of models predicting requests for clarification are shown in Table 4.13. The fixed effects for Models 2-4 in the table illustrate that the results from expert review, QAS, and cognitive interviewing predict more frequent requests for clarification. All together, the method results explain roughly 14% of the variability in requests for clarification at the question level. Hence, compared to adequate answers, less of the variability in requests for clarification is explained by the methods overall. The method results and question characteristics together explain another 54% of the variability in requests for clarification. Questions with numeric response formats were more likely to elicit requests for clarification than questions with verbal labels or yes/no response options. Subjective questions and behavioral frequency questions were also more likely to elicit requests for clarification compared to factual questions.

Table 4.13 Prediction of requests for clarification (0=no request for clarification, 1=request for clarification).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Effect	Null	Methods	Question characteristics	Question and respondent characteristics	QUAID only	Expert review only	QAS only	Cognitive interviewing only
Fixed Effects								
Intercept	-3.15*(.11)	-3.52*(.22)	-4.98*(.36)	-4.95(.37)	-4.31*(.34)	-4.68*(.34)	-1.89*(.38)	-4.28*(.32)
Semantic problems								
QUAID		.11(.07)	.09(.06)	.09(.06)	.13(.07)			
Expert Review		.10(.13)	.17(.11)	.17(.11)		.36*(.10)		
QAS		-.04(.06)	.09(.05)	.09(.05)			.18*(.05)	
Cog.Int.		.15*(.06)	.10*(.05)	.10*(.05)				.17*(.05)
Grade level			.02(.02)	.02(.02)	.02(.02)	.04(.022)	.02(.02)	.02(.02)
Yes/No			Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Numeric			1.56*(.38)	1.56*(.38)	.92*(.37)	1.26*(.37)	1.62*(.37)	1.12*(.36)
Verbal			.28(.24)	.28(.24)	-.03(.25)	.25(.25)	.23(.25)	.11(.25)
Factual			Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Subjective			.83*(.21)	.83*(.21)	1.00*(.22)	.81*(.21)	.97*(.21)	.73*(.22)
Beh.Freq			1.02*(.31)	1.02*(.31)	.98*(.35)	1.21*(.33)	.98*(.33)	1.06*(.33)
H.S.				Ref.	Ref.	Ref.	Ref.	Ref.
<H.S.				-.11(.11)	-.11(.11)	-.11(.11)	-.11(.11)	-.11(.11)
Under age 65				Ref.	Ref.	Ref.	Ref.	Ref.
65 or older				.03(.11)	.03(.11)	.03(.11)	.03(.11)	.03(.11)
Random Effects								
Question level	.71*(.13)	.61*(.12)	.33*(.07)	.33*(.06)	.43*(.09)	.38*(.08)	.39*(.08)	.39*(.12)
Respondent level	.54*(.07)	.54*(.07)	.54*(.07)	.54*(.07)	.54*(.07)	.54*(.07)	.54*(.07)	.54*(.07)
Model Fit								
Deviance	7882.91	7871.17	7829.32	7828.25	7848.97	7838.79	7839.90	7839.40
AIC	7888.91	7885.17	7853.32	7856.25	7870.97	7860.79	7861.90	7861.40

* p < .05. n = 17,666.

Figure 4.2 illustrates the predicted probability of requests for clarification given the number of times that each method finds problems with the meaning of terms in a question. It is easy to see that requests for clarification are somewhat rare. Therefore, there is relatively little variability for the methods to explain on this measure.

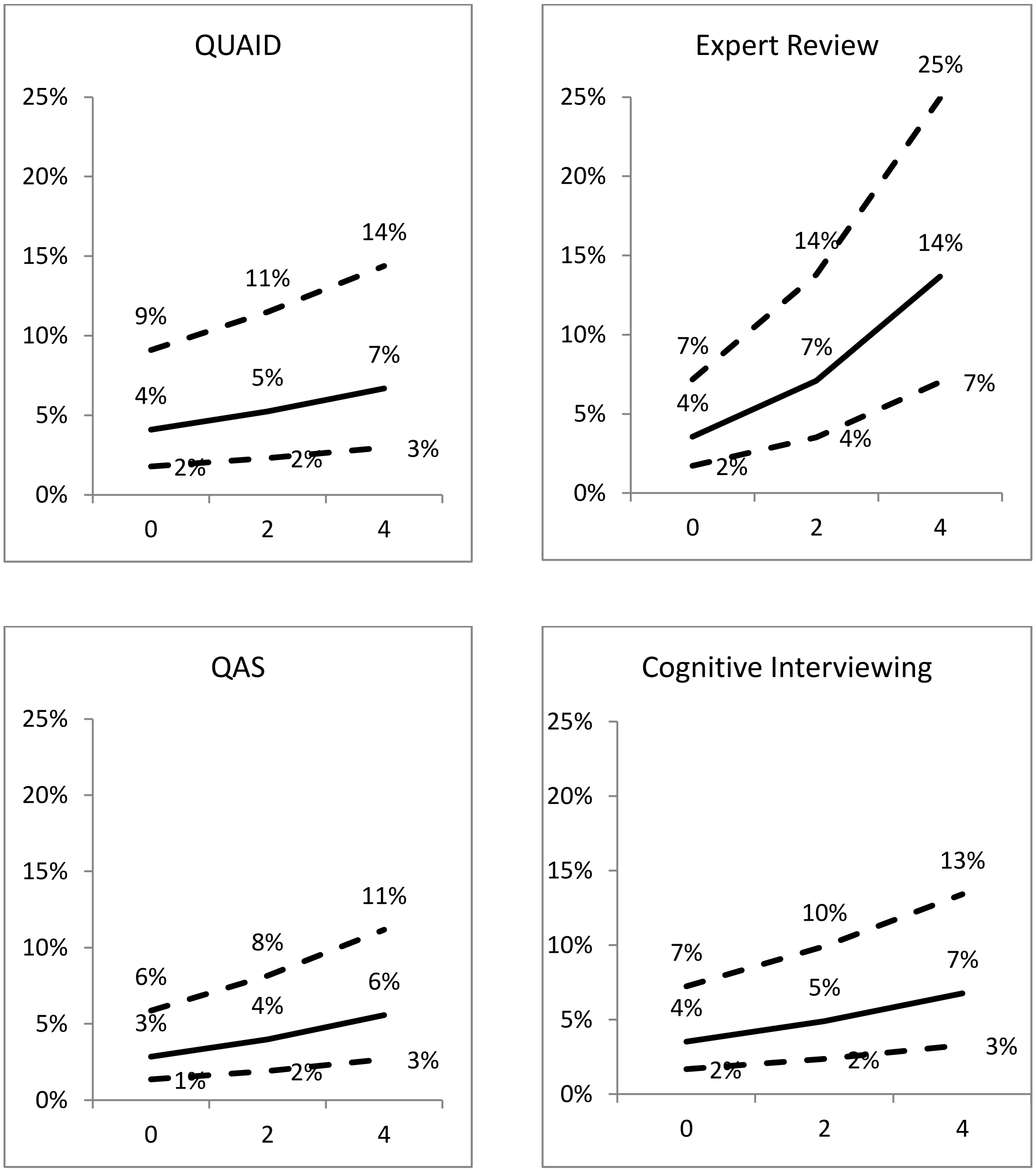


Figure 4.2 Predicted probabilities of requests for clarification given the number of times that semantic problems are detected by QUAID, expert review, cognitive interviewing and QAS.

The strongest predictor of requests for clarification appears to be expert review; however, there is once again considerable variability in the estimated slope coefficient for expert review. The results for QAS and cognitive interviewing are nearly identical. The results for QUAID are actually quite similar to QAS and cognitive interviewing. However, there was slightly more variability associated with the QUAID results so that the coefficient for QUAID is not significant.

I next test the complementary methods hypothesis and the test environment hypothesis. As shown in Table 4.14, all of the likelihood ratio tests show a significant difference between the full model and the models reduced to the use of a single method. In fact, even models that reduced to two methods demonstrate a significant reduction in fit compared to the full model. This suggests that researchers are better served by incorporating a multiple method strategy when trying to detect problems with the meaning of survey questions. Therefore, the complementary methods hypothesis is supported with respect to requests for clarification.

Table 4.14. Test of difference in model fit (Deviance) between the full model (D_0) and models including individual methods (D_1) predicting requests for clarification.

Method	Difference (D_0-D_1)	Critical Value	Significant difference?
QUAID	20.72	7.81	Yes
Expert review	10.54	7.81	Yes
QAS	11.65	7.81	Yes
Cognitive interviewing	11.15	7.81	Yes

An examination of the AIC values provides no support for the test environment hypothesis. Although there is a large difference between QUAID and all other methods with respect to the AIC, there are almost no differences between expert review, QUAID, and cognitive interviewing.

I also conducted analyses on a combined behavior coding dependent variable. The variable was coded 0, 1, or 2 depending on how many behavior codes were assigned to an exchange. For example, an exchange was given a score of 2 if the answer was inadequate and the respondent requested clarification. I included the results for expert review, QAS, and cognitive interviewing as predictor variables in this analysis. I then ran cumulative logit models that assume an ordinal dependent variable. The cognitive interview results were the only results that were significant. This is true in the full model that included the results for expert review and QAS and it is also true when looking at models including the methods individually. Although the cognitive interview results are significant, the hypothesis tests using the model fit statistics really suggest that none of the methods are very good predictors of the overall behavior coding results. Overall, this could suggest that mapping the results from the individual methods on to specific behavior codes works better than predicting overall results. The weak correlation between adequate answers and requests for clarification (-.29) supports this claim.

Item nonresponse results

Table 4.15 shows the results of the models predicting item nonresponse. The table shows that the number of times that QAS and cognitive interviewing found recall problems is a significant predictor of item nonresponse. The number of times that expert review and QAS found problems with the response categories is also a significant predictor of item nonresponse. Models 2-4 show that response category problems identified QAS are significant when used in combination with other methods, but the same results for expert review are not significant when used with other methods. All together, the methods explain approximately 36% of the question level variability in item nonresponse. The method results and question characteristics together explain 61% of the question level variability in item nonresponse. Item nonresponse is more likely for subjective questions and less likely for behavioral frequency questions compared to factual questions. This indicates that although initial answers for behavioral frequency questions are often inadequate, the interviewer is eventually able to work with the respondent and record a valid response. Model 4 includes the respondent level characteristics and shows item nonresponse is higher for older respondents and lower educated respondents.

Table 4.15. Prediction of item nonresponse (0=valid answer, 1=item nonresponse).

Effect	Model 1 Null	Model 2 Methods	Model 3 Question characteristics	Model 4 Question and respondent characteristics	Model 5 QUAID	Model 6 Expert review only	Model 7 QAS only	Model 8 Cognitive Interviewing only	Model 9 SQP only
Fixed Effects									
Intercept	-5.38*(.20)	-5.86*(.27)	-6.65*(.52)	-7.13*(.52)	-6.71*(.56)	-7.38*(.59)	-6.91*(.55)	-6.69*(.49)	-10.01*(1.67)
Respondent Task I problems: recall									
Expert Review		-.92*(.30)	-.04(.30)	-.04(.30)		-.17(.32)			
QAS		.26*(.10)	.09(.09)	.09(.09)			.25*(.09)		
Cog.Int.		.33*(.11)	.41*(.10)	.42*(.10)				.48*(.09)	
Respondent Task II problems: resp.cat.									
QUAID					-.42(.38)				
Expert Review		-.46*(.48)	.84(.47)	.84(.47)		1.51*(.52)			
QAS		.20(.16)	.30*(.14)	.30*(.14)			.25(.16)		
Cog.Int.		.07(.23)	.20(.19)	.20(.19)				.36(.19)	
SQP Total Quality									.06*(.03)
Grade level			.07(.04)	.07(.04)	.05(.04)	.11*(.04)	.04(.04)	.04(.04)	.02(.04)
Yes/No			Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Numeric			1.18*(.59)	1.19*(.59)	.27(.69)	.93(.69)	.54(.67)	.95(.60)	-.04(.71)
Verbal			-1.26*(.42)	-1.26*(.42)	-.78(.44)	-1.34*(.45)	-.80(.45)	-.87*(.38)	-1.03*(.44)
Factual			Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Subjective			1.51*(.40)	1.51*(.41)	1.91*(.39)	2.35*(.41)	1.30*(.41)	1.44*(.34)	2.06*(.39)
Beh.Freq			-2.92*(.75)	-2.93*(.74)	-.79*(.73)	-1.40*(.80)	-1.46*(.71)	-2.87*(.70)	-.69*(.71)
H.S.				Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
<H.S.				.93*(.15)	.93*(.15)	.93*(.15)	.93*(.15)	.93*(.15)	.93*(.15)
Under age 60				Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
60 or older				.43*(.14)	.43*(.14)	.43*(.14)	.43*(.14)	.43*(.14)	.43*(.14)
Random Effects									
Question level	2.26*(.43)	1.44*(.28)	.88*(.19)	.89*(.19)	1.41*(.29)	1.30*(.27)	1.30*(.26)	.98*(.21)	1.40*(.29)
Respondent level	2.24*(.21)	2.26*(.22)	2.27*(.22)	2.05*(.20)	2.03*(.20)	2.04*(.20)	2.04*(.20)	2.05*(.20)	2.04*(.20)
Model Fit									
Deviance	7641.07	7604.03	7575.44	7521.37	7559.00	7551.52	7547.22	7532.34	7555.87
AIC	7647.07	7622.03	7603.44	7553.37	7581.00	7575.52	7571.22	7556.34	7577.87

* p < .05. n = 34,955.

Table 4.16 tests the complementary methods hypothesis. The full model consists of the results for expert review, QAS, and cognitive interviewing. I have excluded the results from QUAID and SQP from the full model because they are in the opposite from expected direction. The results of the likelihood ratio tests reveal that reducing down to any single method significantly reduces the fit of the model. The use of QAS and cognitive interviewing together without the expert review results does not significantly reduce the fit of the model. Hence, the use of cognitive interviewing to uncover recall problems and the QAS to uncover problems with response categories works well with predicting item nonresponse. Next, we examine the AIC values in Table 4.15 to understand if the methods are ordered as suggested by the test environment hypothesis. The hypothesis is supported by the data. There are large differences in the AIC between cognitive interviewing and all other methods. The QAS has a lower value of the AIC compared to expert review, QUAID, and SQP. Expert review has a lower AIC than QUAID and SQP.

Table 4.16. Test of difference in model fit (Deviance) between the full model (D_0) and models including individual methods (D_1) predicting item nonresponse.

Method	Difference (D_0-D_1)	Critical Value	Significant difference?
Expert review	30.15	9.49	Yes
QAS	25.85	9.49	Yes
Cognitive interviewing	10.97	9.49	Yes
QAS and cognitive interviewing	3.17	5.99	No
Expert review and cognitive interviewing	7.82	5.99	Yes

Figure 4.3 plots the predicted probability of item nonresponse given the number of times that the QAS and cognitive interviewing found respondent task problems. It is clear from the figure that there is a lot of variability around the lines. Also, there is very little overall

variability in item nonresponse between questions. The average level of item nonresponse across items is only around 3% since a “don’t know” or refused option was not offered to respondents during the interview. The figure also depicts a slightly stronger relationship between cognitive interviewing recall problems and item nonresponse than between QAS response category problems and item nonresponse.

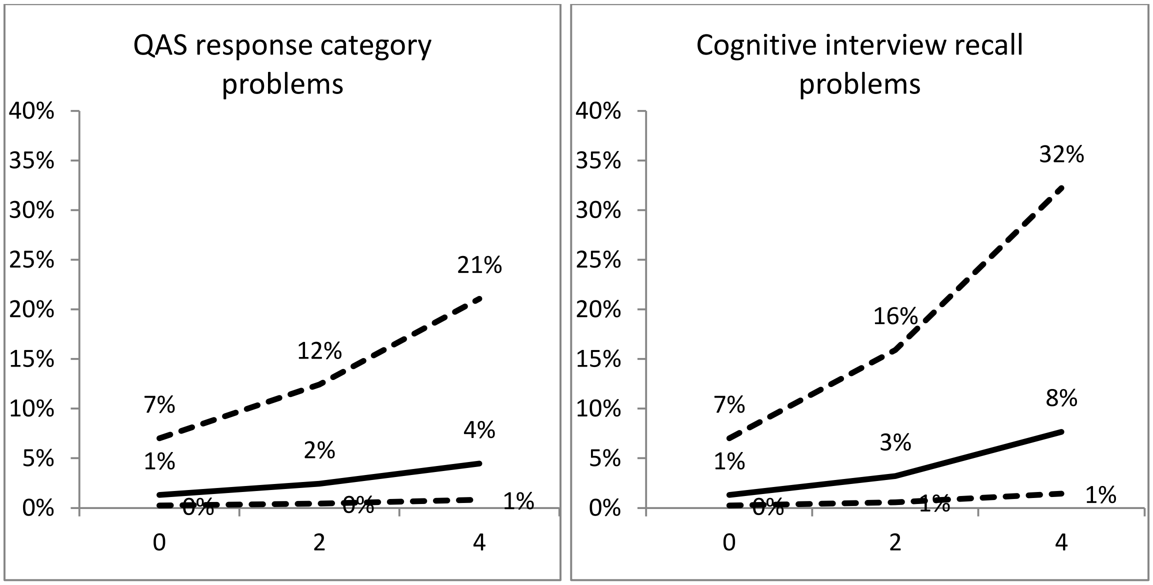


Figure 4.3. Predicted probabilities of item nonresponse given the number of times that response category problems are detected by QAS and recall problems are detected by cognitive interviewing.

Response latency results

It was hypothesized that recall problems would reflect uncertainty and the possibility of producing nonsubstantive responses. In addition, Draisma and Dijkstra's (2004) show that nonsubstantive responses produce the longest response times. The question level analyses and previously presented multilevel models showed that semantic problems related to question meaning are predictive of requests for clarification, which will in turn cause the respondents to take longer to answer questions.

Table 4.17 does show that some of the methods are systematically related to response latencies. Recall problems identified by the QAS and cognitive interviewing lead to significantly longer response latencies according to Model 2. There is also evidence in the table that semantic II problems related to question meaning found by expert review and cognitive interviewing are also related to longer response latencies. However, the cognitive interviewing results are only marginally significant (p=.07). All together the methods explained roughly 36% of the question level variability ((.25-.16)/.25). I have excluded the SQP results from the models involving all methods since the SQP results are in the opposite from expected direction. Question characteristics and the method results together explain approximately 53% of the question level variability ((.25-.12)/.25.) Questions requiring numeric answers have significantly longer response latencies. Subjective questions and behavioral frequency questions have longer response latencies compared to factual questions.

Table 4.17. Prediction of log transformed response latency.

Effect	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
	Null	Methods	Question characteristics	Question and respondent characteristics	QUAID	Expert review	QAS	Cognitive Interviewing	SQP only
Fixed Effects									
Intercept	7.30*(.07)	7.09*(.12)	6.56*(.21)	6.48*(.21)	6.50*(.19)	6.42*(.20)	6.53*(.20)	6.57*(.18)	4.90*(.51)
Semantic II problems: meaning									
QUAID		.05(.04)	.06(.04)	.06(.04)	.08*(.04)				
Expert Review		-.03(.07)	.02(.07)	.02(.07)		.14*(.06)			
QAS		-.08*(.03)	-.01(.03)	-.01(.03)			.00(.03)		
Cog. Int.		.058(.034)	.03(.03)	.03(.03)				.05(.03)	
Respondent task I problems: recall									
Expert Review		.15(.09)	.00(.11)	.00(.11)		-.03(.10)			
QAS		.13*(.04)	.09*(.03)	.09*(.03)			.11*(.03)		
Cog. Int.		.04(.04)	.02(.04)	.02(.04)				.07(.04)	
SQP Total Quality									.03*(.01)
Grade level			.00(.01)	.00(.01)	.01(.01)	.02(.01)	.01(.01)	.01(.01)	.00(.01)
Yes/No			Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Numeric			.66*(.22)	.66*(.22)	.57*(.21)	.68*(.21)	.65*(.22)	.68*(.21)	.66*(.22)
Verbal			.11(.13)	.11(.13)	.00(.13)	.09(.14)	.12(.14)	.05(.14)	.11(.13)
Factual			Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Subjective			.31*(.13)	.31*(.15)	.54*(.12)	.46*(.12)	.31*(.13)	.40*(.12)	.31*(.15)
Beh.Freq			.50*(.25)	.50*(.25)	.72*(.20)	.86*(.23)	.55*(.20)	.53*(.23)	.50*(.25)
> H.S.				Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
H.S.				.13(.09)	.13(.09)	.13(.09)	.13(.09)	.13(.09)	.13(.09)
Under age 65				Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
65 or older				.14(.09)	.14(.09)	.14(.09)	.14(.09)	.14(.09)	.14(.09)
Random Effects									
Question level	.25*(.04)	.16*(.03)	.12*(.02)	.12*(.02)	.12*(.02)	.15*(.01)	.13*(.02)	.14*(.03)	.14*(.02)
Respondent level	.19*(.03)	.19*(.03)	.19*(.03)	.19*(.03)	.19*(.03)	.19*(.01)	.19*(.03)	.19*(.03)	.18*(.03)
	1.19*(.02)	1.19*(.02)	1.19*(.02)	1.19*(.02)	1.19*(.02)	1.19*(.01)	1.19*(.02)	1.19*(.02)	1.19*(.02)
Model Fit									
Deviance	14953.90	14917.70	14898.50	14894.30	14910.30	14908.90	14899.30	14907.30	14902.30
AIC	14961.90	14939.70	14930.50	14930.50	14934.30	14934.90	14925.30	14933.30	14926.30

* p < .05. n = 4,815.

Table 4.18 compares the model fit for the models using the individual methods versus the model that includes all methods. The results suggest that the use of QUAID, expert review, and cognitive interviewing individually result in a significant reduction in model fit. The use of the QAS does not result in a significant reduction in model fit compared to using all methods together. As mentioned, before the SQP results are in the opposite from expected direction. Hence, these results do not support the complementary hypothesis. The test environment hypothesis is also not supported by the AIC values in Table 4.18. QAS is the best predictor of response latency followed by cognitive interviewing and then expert review and QUAID. Once again, the AIC value for SQP is not very meaningful since the relationship is in the opposite from expected direction.

Table 4.18. Test of difference in model fit (Deviance) between the full model (D_0) and models including individual methods (D_1) predicting response latency.

Method	Difference (D_0-D_1)	Critical Value	Significant difference?
QUAID	16.00	12.59	Yes
Expert review	14.60	11.07	Yes
QAS	5.00	11.07	No
Cognitive interviewing	13.00	11.07	Yes

The response latencies from model 7 for QAS are shown in figure 4.4. The figure gives the reader and idea of how much response latency increases by the number of times that recall problems are found.

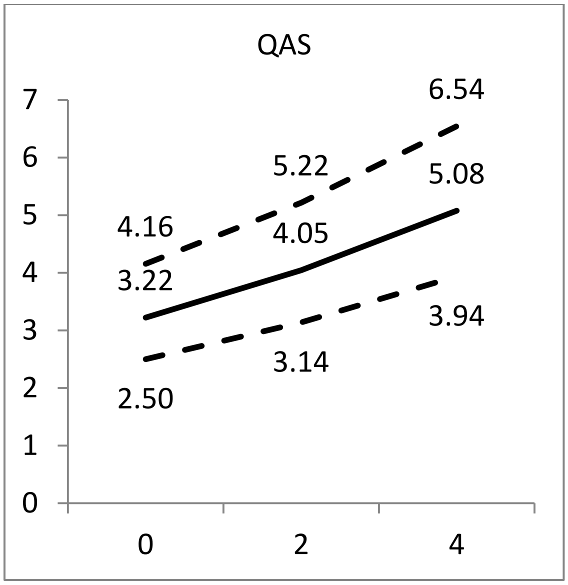


Figure 4.4. Predicted response latency (in seconds) given the number of times that recall problems are detected by QAS.

Interactions with respondent characteristics

Finally, interactions between the method results and respondent characteristics were investigated. The results for the ex-ante and laboratory methods were combined together for this analysis. In other words, one summary measure for the number of respondent task problems was calculated across all results for expert review, QAS, and cognitive interviewing. This summarized variable was then interacted with education and age to test for interactions. A similar summarization was done for semantic meaning such that the results across QUAID, expert review, QAS, and cognitive interviewing were added together. Two significant interactions were detected. Both involved interactions between the method results and the education level of the respondent. This interaction was detected for two dependent variables: adequate answers and item nonresponse. The nature of the interaction with adequate answers is shown in figure 4.5. It is evident from the figure that the slope of the line for respondents with a high school education or less is steeper than the slope of the line for respondents with more than a high school education. A similar figure could be produced for item nonresponse. Hence there is support for the hypothesis that problems identified by ex-ante and laboratory methods predict more significant problems for those with lower levels of education.

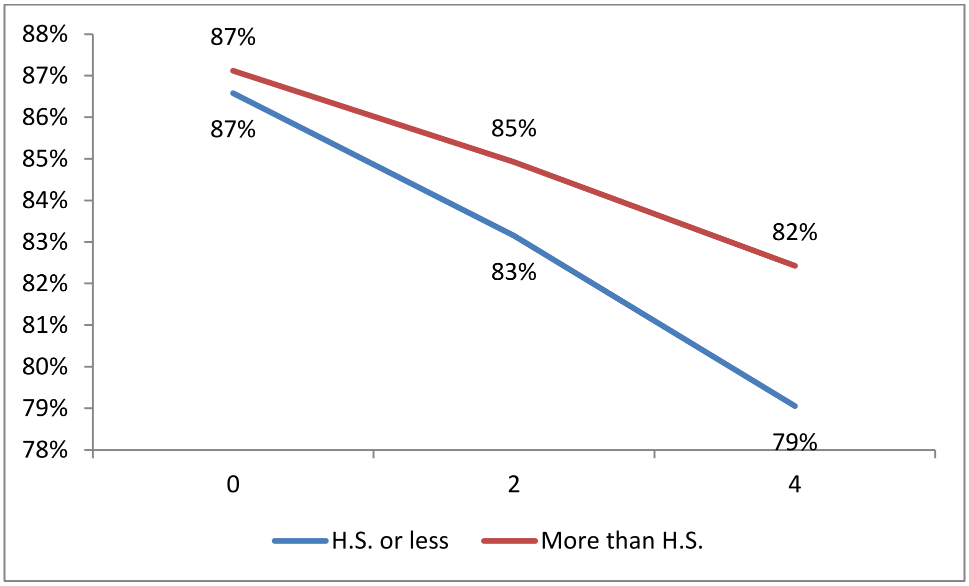


Figure 4.5. Predicted probabilities of adequate answers given the number of times that response task problems are detected by ex-ante and laboratory methods for respondents with a high school education or less and respondents with more than a high school education.

Discussion

This chapter took a confirmatory approach to determine if problems identified by ex-ante and laboratory methods cause problems in the field. It examined three main hypotheses. A summary of the first two complementary methods hypothesis and test environment hypothesis are shown in Table 4.19. The problems that are found by these methods do tend to predict quality in the field. At least two of the methods were predictive of each of the indirect data measures of data quality in the field used in this chapter. The exact combination of methods varied between the dependent variable examined. All together the methods explained between 14% and 49% of the question level variability in these indirect measures of data quality. It is often suggested that because of the low level of agreement between methods that it is best to use as many methods together in combination as feasible. This hypothesis at least partially supported for three of the four measures of data quality in the field. However, it was often the case that cognitive interviewing alone did just as well as using a combination of methods for predicting the most problematic questions. QAS alone did just as well as using a combination of methods to predict response latency.

Table 4.19. Summary of evidence for the complementary methods hypothesis and test environment hypothesis.

Measure	Complementary methods hypothesis		Test environment hypothesis	
	Conclusion	Results	Conclusion	Results
Adequate answers	Partially supported	Cognitive interview recall problems and QAS sensitivity problems predict adequate answers the best in multilevel models. Cognitive interviewing alone best predictor of most problematic questions.	Partially supported	Cognitive interviewing better predictor than expert review and SQP; Cognitive interviewing similar to QAS
Requests for clarification	Partially supported	Multilevel model fit reduced unless cognitive interviewing, Expert review, and QAS used together. Cognitive interviewing alone best predictor of most problematic questions.	Not supported	No difference between expert review, QAS, and cognitive interviewing, but all three better than QUAID
Item nonresponse	Partially supported	QAS response category problems and cognitive interviewing recall problems predict item nonresponse. Use of cognitive interviewing recall problems alone most problematic questions.	Supported	Cognitive interviewing better predictor than expert review, QAS, better than QUAID, and SQP
Response latency	Partially supported	Model fit not reduced when using only QAS. Cognitive interviewing semantic problems and QAS recall problems predict questions with longest latencies.	Not supported	QAS better predictor than cognitive interviewing which is better than expert review

In general, this chapter suggests that cognitive interviewing had an advantage over the other methods in assessing the quality of the information that respondents processed to answer survey questions. This is evident from the models predicting adequate answers and item nonresponse. This is not entirely surprising given that observing respondents struggle with answering certain questions in the lab should be related to how well respondents are able to answer in the field. Experts and forms appraisal do appear to be better at detecting problems with response categories or problems with item sensitivity that might cause problems in the field.

Much of the question evaluation literature focuses on issues of meaning with survey questions. This chapter found that expert review, QAS, and cognitive interviewing results are predictive of requests for clarification in the field. In contrast to response task problems, the data did not show that any one of these methods was necessarily any better than the other. Hence, at least with respect to overt evidence of question misunderstanding, this chapter does not provide clear guidance about which method predicts more serious problems. Future research should be guided towards understanding this more thoroughly since these are the bulk of the problems found by ex-ante and laboratory methods.

This chapter also found evidence that the method results apply differently to different types of respondents. This supports the current practice of attempting to recruit respondents with diverse backgrounds. This finding highlights one advantage to the analytical approach of this dissertation. Cross-classified multilevel models are an important tool for identifying how both respondent and question level characteristics influence data quality. In addition, these types of models not only allow us to understand more about fixed or systematic relationships between these characteristics and data quality, but also allow us to understand more about how much of the random variability is explained. Hence, this dissertation adds to a growing literature using these types of models for methodological research.

CHAPTER 5: THE REALTIONSHIP BETWEEN METHOD RESULTS AND DIRECT INDICATORS OF DATA QUALITY

Introduction

The results from chapter 4 provide evidence that some of the ex-ante and laboratory methods are predictive of the quality of questions in the field. However, that chapter utilized field-based methods as indirect indicators of data quality. More research is needed that examines the relationship between method results and direct indicators of data quality such as reliability or validity. Fortunately, the survey practicum included a reinterview design that enables such an examination. This chapter explores the relationship between ex-ante, laboratory, and field method results and the reliability of survey questions. The methods will be used to predict discrepancies in the answers to survey questions between the original interview and the reinterview.

This chapter also undertakes a confirmatory approach to the method evaluation (Forsyth, Rothgeb, and Willis, 2004) and explores hypotheses similar to those of chapter 4. However, these hypotheses are investigated with respect to the prediction of the reliability of survey questions. First, this chapter will test the complementary methods hypothesis that states that it is better to use the findings from question evaluation methods together since there is considerable disagreement between the methods (Presser et al., 2004; Yan, Kreuter, and Tourangeau, 2012).

Second, this chapter tests the test environment hypothesis. This hypothesis tests an ordering of the methods according to how well they predict discrepant answers. According to the review in chapter one, behavior coding is the only method that has been shown to predict the reliability of survey questions (Hess, Singer, and Bushery, 1999).

This perhaps reflects the importance of observing how the questions perform in a realistic survey setting. Expert review and cognitive interviewing are the only other methods in this dissertation to be studied with respect to reliability and no relationship was found between the results from those methods and reliability (Yan, Kreuter, and Tourangeau, 2012).

The analysis begins with a brief presentation of the descriptive statistics for the variables used in the analysis. Next, I present the results of cross-classified multilevel models predicting the likelihood of discrepant answers between the original interview and the reinterview. These models build on O'Muircheartaigh (1991), who used regression models to understand the effect of proxy reporting on gross discrepancy rates in the Current Population Survey reinterview. This chapter uses multilevel models in order to explain the variability in discrepancies at both the question and respondent level. As illustrated in chapter 4, these models are powerful tools for this type of analysis because they enable the researcher to test a variety of hypothesis regarding combinations of variables and their utility at predicting a dependent variable.

Analysis

Descriptive Statistics

Table 5.1 shows the descriptive statistics for the variables that will be used in the models. The predictor variables in these models include the method results at the question level and some results that refer to the number of times that the respondents visited the questions (i.e. respondents and questions are repeated in the data set). The first set of results includes the number of response task problems found by expert review, QAS, and cognitive interviewing. These types of problems may interfere with the respondent's ability to recall information and form an accurate answer. The second set of problems refers to problems identified with the response categories of the questions. For example, the methods might identify overlapping response categories that make it difficult to choose between categories. Also included in the models are the total quality score from SQP, response latency, and behavior coding results. The percent of initial exchanges resulting in adequate answers and the percent of question answer exchanges where the respondent paused for longer than one second or used speech fillers. The behavior coding results will be modeled at both the question and respondent level. Finally, whether or not respondents gave a discrepant answer will serve as the dependent variable in the models.

Table 5.1. Mean and standard deviation for variables in the models.

Variable	N	Mean	Standard Deviation	Minimum	Maximum
Question level results					
Recall problems					
Expert Review	53	.51	.64	0	2
QAS	53	2.42	1.57	0	6
Cognitive Interviewing	53	1.19	1.82	0	7
Response category problems					
QUAID	53	.23	.54	0	3
Expert Review	53	.08	.27	0	1
QAS	53	.62	.92	0	3
Cognitive Interviewing	53	.26	.68	0	3
Other methods					
SQP Total Quality	53	58.17	5.06	47.60	66.40
Response latency	53	4.10	2.86	.50	14.90
% Adequate answers (question level)	53	76.60	14.62	40.00	97.16
% Pauses or fillers (question level)	53	11.70	8.16	1.11	41.97
Question visit results					
Adequate answer (respondent level)	5426	.77	.43	0	1
Pauses or filler (respondent level)	5426	.12	.32	0	1
Discrepant answers	10523	.21	.41	0	1

The correlations between the question level predictor variables in Table 5.1 are shown in Table 5.2.

Table 5.2. Correlations between question level predictor variables (n=53).

		Recall problems			Response category problems								
		ER	QAS	CI	QUAID	ER	QAS	CI	SQP	Resp. Lat.	Ad. Answ.	Pauses	IOI
Recall problems	ER	1	.03	.18	-.12	.11	-.42*	-.09	.38*	.20	-.19	.36*	-.22
	QAS		1	.33*	.09	.02	.35*	.09	.28*	.47*	-.56*	.49*	.49*
	CI			1	.13	.56*	-.06	.22	.02	.26	-.57*	.34*	.31*
Response category problems	QUAID				1	.41*	.44*	.15	-.06	-.15	-.03	-.22	.34*
	ER					1	.04	.42*	-.03	-.11	-.12	-.16	.19
	QAS						1	.04	-.00	-.02	-.17	-.12	.29*
	CI							1	.14	-.03	-.17	-.01	.09
	SQP								1	.17	-.13	.08	.05
	Resp.Lat.									1	-.71*	.72*	.07
	Ad. Answ.										1	-.69*	-.21
	Pauses											1	.11
	IOI												1

*p< .05

Model Results

Table 5.3 presents the results of ten models. Model 1 represents the null model and illustrates the total amount of variation at the question and respondent levels respectively. One can compare the random effects of the other models to the null model to understand how much of the variation in discrepant answers is explained by the predictor variables. Model 2 is the full model that includes the results for all ex-ante, laboratory, and field based methods. The large drop in question level variability suggests that the methods are able to explain a significant amount of the variability in discrepant answers. Almost two thirds (63%) of the question level variance in discrepant answers is explained by the methods collectively. It is also noteworthy that, in contrast to the random effects in chapter 4, most of the variability appears to be between questions rather than between respondents.

Three methods in model 2 have a significant relationship with discrepant answers: recall problems found by QAS, problems with categories found by QUAID, and the percentage of exchanges that involved pauses and fillers as found by behavior coding. An examination of the question level random effects from model 10 and model 1 reveals that the results from these three methods together explains more than half of the question level variability in discrepant answers ((1.04-.49)/1.04). Models 3-9 investigate the effect of the individual methods on discrepant answers. Among these individual models, the smallest estimate of question level variability occurs in the model that includes the behavior coding estimates. Behavior coding explains roughly 36% of the variability in discrepant answers ((1.04-.66/1.04)). This provides further evidence regarding the strength of behavior coding at predicting reliability. These individual models also suggest that recall problems from QAS, adequate answers from behavior coding, and longer response latencies are related to discrepant answers. Hence, even though some of the methods are not significant predictors of discrepancies in the full model, there does at least appear to be a bivariate relationship between the results for these methods and the likelihood of discrepant answers. One might notice that the results for QUAID and pauses and fillers have a stronger relationship with discrepant answers when they are included in models with other variables. This suggests that these methods are best utilized in combination with other methods rather than individually.

Table 5.3. Prediction of discrepant answers between wave 1 and wave 2 by different methods with behavior codes modeled at the question level (n=10,523).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Intercept	-1.65*(.15)	-2.07(1.83)	-1.41(1.70)	-1.76*(.15)	-1.70*(.19)	-2.56*(.23)	-1.90(.16)	-.22*(1.02)	-2.37*(.23)	-2.95*(.22)
SQP										
Total Quality		-.02(.03)	.00(.03)							
Recall problems										
Expert Review		-.13(.65)			-.04(.29)					
QAS		.20*(.08)				.37*(.08)				.21*(.08)
Cog. Int.		.04(.08)					.27*(.07)			
Problem with categories										
QUAID		.45*(.21)		.50(.26)						.64*(.20)
Expert Review		.94(.57)			.85(.53)					
QAS		.04(.15)				.02(.14)				
Cog. Int.		-.32(.17)					-.24(.20)			
Field methods										
% Adequate answer		.00(.01)						-.02*(.01)		
% Pauses/fillers		.05*(.02)						.037(.020)		.06*(.01)
Response latency		.06(.05)							.18*(.04)	
Random effects										
Question	1.04	.38	1.04	.95	.99	.70	.81	.66	.79	.49
Respondent	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14
Model fit										
Deviance	10001.15	9954.07	10001.13	9997.51	9998.67	9980.92	9988.90	9979.45	9987.05	9964.75
AIC	10007.15	9982.07	10009.13	10005.51	10008.67	9990.92	9998.90	9989.45	9995.05	9976.75

Figure 5.1 plots the predicted probability of a discrepant answer given the number of times that specific problems were detected by QUAID, QAS, and cognitive interviewing. The model plots the probabilities from model 4 (QUAID), model 6 (QAS), and model 7 (cognitive interviewing) from Table 5.3. The figure gives the reader a sense for how much the probability of discrepant answers increase given the number of times that each problem is found by a specific method. The dotted lines represent the 95% confidence interval around the estimated values. It is noticeable that there is a wider prediction interval for the QUAID results representing more variability in the estimate.

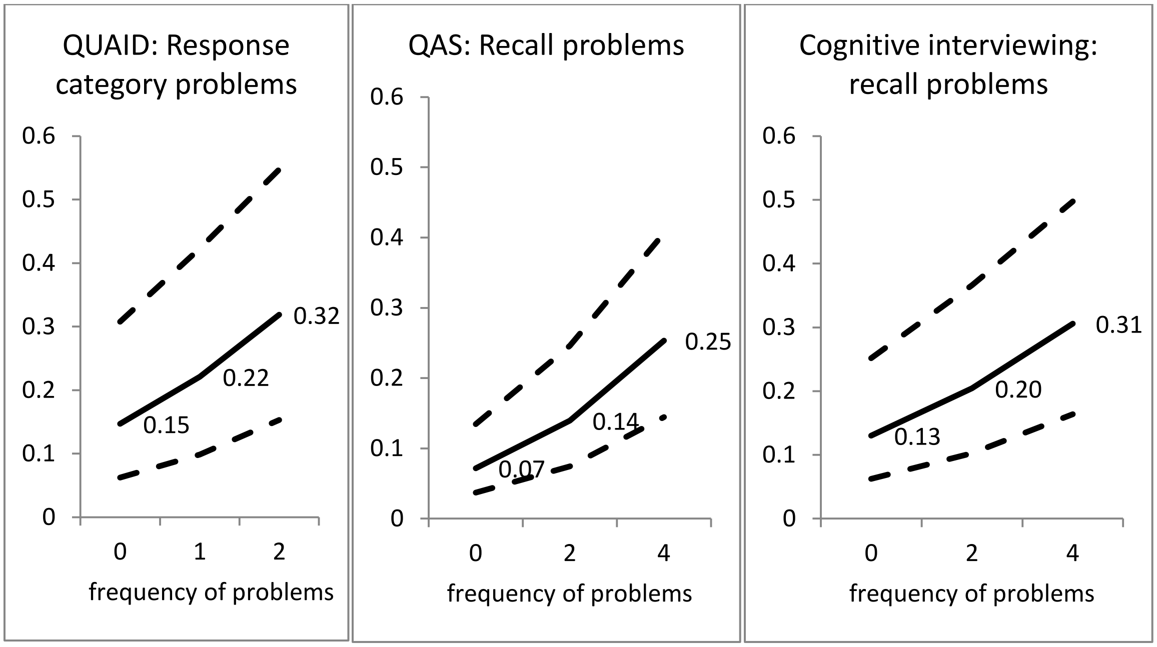


Figure 5.1. Effect of problem detection by QUAID, QAS, and cognitive interviewing on the proportion of discrepant answers.

The effect of the behavior coding results and response latencies on the predicted probability of discrepant answers is plotted in Figure 5.2.

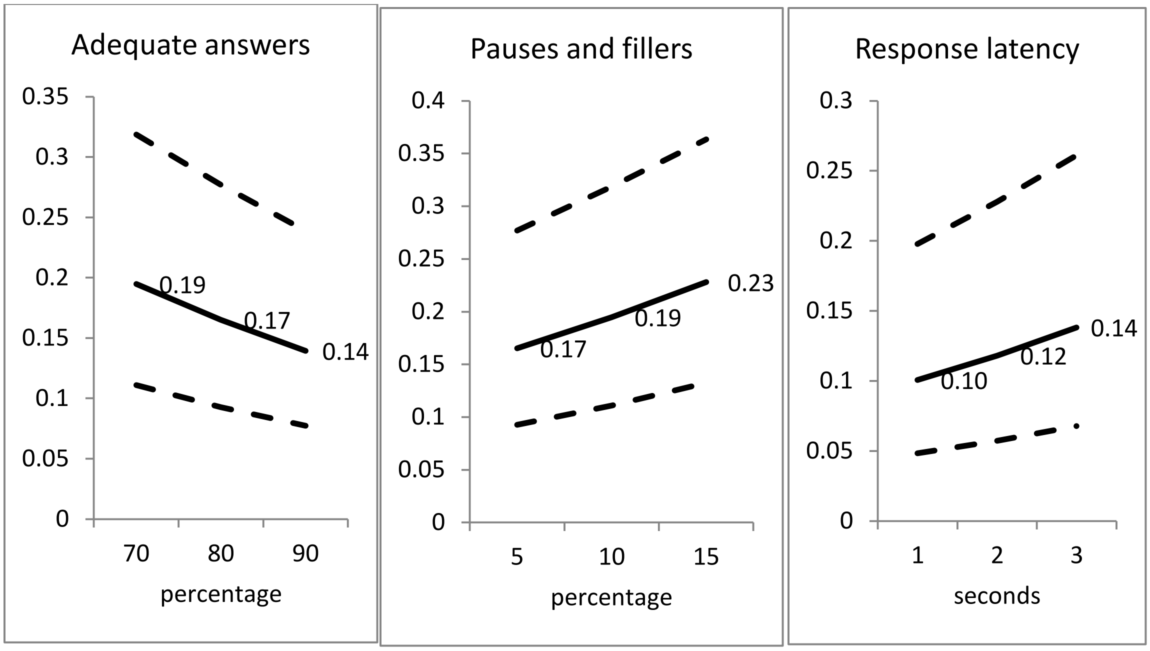


Figure 5.2. The effect of the behavior coding results and response latencies on the predicted probability of discrepant answers.

Table 5.4 provides the relevant tests for the complementary methods hypothesis. None of the individual methods alone provide an adequate fit to the data compared to the full model. Therefore, the complementary methods hypothesis is supported. Model 10 which includes a combination of results from QAS, QUAID, and behavior coding to predict discrepancies does provide an adequate fit to the data compared to the full model using all methods. This suggests that although no individual method is adequate, it is possible to do well at predicting discrepancies using a smaller subset of the methods.

Table 5.4. Test of difference in model fit (Deviance) between the full model (D_0) and reduced methods (D1) predicting discrepant answers between the original interview and the reinterview.

Method	Difference (D_0-D_1)	Critical Value	Significant difference?
SQP	47.06	18.31	Yes
QUAID	43.44	18.31	Yes
Expert Review	44.60	16.92	Yes
QAS	26.13	16.92	Yes
Cognitive interviewing	34.83	16.92	Yes
Behavior coding	25.83	16.92	Yes
Response latency	32.98	18.31	Yes
QAS,QUAID, Behavior coding	10.68	15.51	No

Note. Full model includes QUAID, Expert Review, QAS, Cognitive Interviewing, Response Latency, Behavior coding

Ordering of the methods by their AIC values in Table 5.3 examines the test environment hypothesis. Behavior coding is better than nearly all of the methods at predicting discrepant answers. However, the behavior coding and QAS results provide the best fit to the data according to the AIC. Their AIC values are very similar. This is followed by response latency and cognitive interviewing, which are better fits to the data than QUAID, expert review, and SQP. However, QUAID is a better fit than expert review and SQP. With the exception of QAS, there is partial support for the ordering of the methods such that field methods are better than laboratory methods and laboratory methods are better than expert methods and computer-based systems.

Unlike other studies such as Hess, Singer, and Bushery (1999), the behavior coding results from this study can be modeled at either the question or respondent level of the multilevel models. This is because the behavior coding was performed on the original interview rather than an independent sample from reinterview study. Inclusion of the behavior coding results at the question level examines the overall relationship between behavior coding results and discrepant answers. Inclusion of the behavior coding results at the respondent level specifically assesses the impact of the behavior by a specific respondent in the original interview on discrepant answers between time 1 and time 2. Table 5.5 illustrates the results of modeling the behavior codes at the respondent level. Model 1 and model 2 are repeated from Table 5.3 These models are estimated once again because the models are being fit to a subset of the data, which will produce different model fit statistics. An important goal of this analysis is to understand whether modeling the behavior codes at the respondent level improves the fit of the model compared to modeling the overall results at the question level. Comparison of the AIC values between model 2 and model 3 does provide evidence that the model fit is much improved by modeling the behavior codes at the respondent level.

Table 5.5. Prediction of discrepant answers between wave 1 and wave 2 by different methods with behavior codes modeled at the respondent level (n=5,426).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Intercept	-1.65*(.14)	-1.92(1.84)	-.78(1.43)	-2.03*(.24)	-2.03*(.25)	-1.85*(.23)	-1.83*(.22)
Question level							
SQP							
Total Quality		-.01(.03)	-.02(.03)				
Response difficulty							
Expert Review		-.18(.29)	-.08(.29)				
QAS		.17*(.08)	.20*(.08)	.17*(.07)	.20*(.08)	.24*(.07)	.28*(.07)
Cog. Int.		.04(.08)	.05(.07)	.09(.06)		.10(.06)	
Problem with categories							
QUAID		.46*(.21)	.42(.22)	.54*(.19)	.57*(.20)	.45*(.19)	.49*(.20)
Expert Review		.98(.57)	.76(.56)				
QAS		.05(.15)	.02(.15)				
Cog. Int.		-.35*(.17)	-.34(.18)				
Field methods							
% Adequate answer		.00(.01)					
% Pauses/fillers		.04*(.02)					
Response latency		.06(.05)	.07(.04)	.077(.041)	.09*(.04)		
Respondent level							
W1 Adequate ans.			-.87*(.08)	-.87*(.08)	-.87*(.08)	-.87*(.08)	-.88*(.08)
W1 Pause/Filler			.50*(.10)	.50*(.10)	.50*(.10)	.51*(.10)	.51*(.10)
Random effects							
Question	.94	.35	.39	.46	.48	.49	.52
Respondent	.15	.16	.10	.10	.10	.10	.10
Model fit							
Deviance	5204.25	5159.99	5022.43	5028.75	5030.90	5032.28	5035.15
AIC	5210.25	5187.99	5050.43	5046.75	5046.90	5048.28	5049.15

The results from Table 5.5 are consistent with the results in Table 5.3 in some respects. As shown in Table 5.6, the likelihood ratio test comparing the deviance statistics between model 3 and model 7 fails to reject the null hypothesis of no difference. This means that one could still use the results from QUAID, QAS, and behavior coding without a significant loss of model fit.

Table 5.6. Test of difference in model fit (Deviance) between the full model (D_0) and reduced methods (D_1) predicting discrepant answers between the original interview and the reinterview.

Method	Difference (D_1-D_0)	Critical Value	Significant difference?
QUAID, QAS, Cognitive Interviewing, Response Latency, Behavior coding	6.32	11.07	No
QUAID, QAS, Response Latency, Behavior coding	8.47	12.59	No
QUAID, QAS, Cognitive interviewing, Behavior coding	9.85	12.59	No
QUAID, QAS, Behavior coding	12.72	14.07	No

Note. Full model includes QUAID, Expert Review, QAS, Cognitive Interviewing, Response Latency, Behavior coding

Some differences are also present between the two tables. One difference is that adequate answers in the initial exchange of the original interview are predictive of discrepant answers between the original interview and reinterview. In contrast, the percentage of adequate answers at the question level was not significant in the full model. Table 5.6 is also less clear about which combination of methods are the best. There is a significant reduction in model fit when comparing model 4 with model 7. Model 4 includes the results for both cognitive interviewing and response latency. As was demonstrated earlier in Table 5.3, these methods do seem to have a marginal effect on the prediction of discrepant answers. The results from this analysis suggest that their results may actually complement the other methods and they should be considered together when understanding the reliability of survey questions.

Question level analysis

I supplemented the results from the multilevel models with a question level analysis of reliability. The following analysis supplements the previous analysis in a few different ways. First, a question level analysis allows me to use the Index of inconsistency (IOI) as a measure of reliability. Hess, Singer, and Bushery (1999) show that the IOI is equal to 1 – kappa. As explained in chapter one, kappa corrects for the probability that two measures can agree by chance. In addition, since some of the variables in my analysis are ordinal variables with more than two categories, I can calculate different measures of reliability that penalize disagreement less when there is a difference one category between waves and penalize disagreement more when there is a difference of more than one category. I can then see how this affects the conclusions drawn from the models. Finally, as in the last chapter, I can look at how the method results predict the IOI using different thresholds. For example, I can see whether behavior coding is a better predictor of the IOI when 80 percent of the answers for a question are adequate versus 85 percent of the answers.

I begin by showing how the results from this study compare to a previous study from the literature. Hess, Singer, and Bushery (1999) used behavior coding to predict the IOI. They found that a threshold of 85% of adequate answers for a question was the most predictive of the IOI. The results from this dissertation also show that behavior coding results are predictive of the IOI. However, there were a couple of differences. First, the R-squared values from the models were lower for this dissertation compared to Hess et al. Second, the results from this study, shown in Table 5.7, show that there was little difference in predictive power between thresholds of 80 percent and 85 percent. In fact, a threshold of 80 percent was somewhat more predictive of the IOI than a threshold of 85 percent. It is difficult to tell what might be causing the differences. The main difference might be that Hess, Singer, and Bushery study covered the topic of food security, whereas the study for this dissertation covered current events. There were significantly more attitudinal items included in this dissertation, which may have played a role in the differences. Even with some of these differences, this lends further support for the use of behavior coding in understanding data quality.

Table 5.7. Prediction of question level reliability (n=53).

	Threshold		
Variables	80 percent	85 percent	90 percent
Hess, Singer, and Bushery (1999)			
Intercept	47.60*(16.24)	50.85*(10.96)	50.85*(13.11)
Percent exact/slight change	15.89(17.03)	8.33(11.62)	.34(13.66)
Percent adequate answers	-22.32*(6.15)	-22.87*(5.48)	-12.28*(7.29)
R-square	.29	.36	.09
Survey practicum results			
Intercept	28.09*(9.80)	28.09*(10.01)	41.28*(4.55)
Percent exact/slight change	14.48 (10.15)	12.75 (10.30)	-2.86 (5.22)
Percent adequate answers	-12.55*(3.90)	-11.80*(4.20)	-11.48*(5.04)
R-square	.18	.15	.12
*p < .05			

Next, I extend the analyses by Hess, Singer, and Bushery by incorporating other methods. Table 5.8 shows how the other methods predict the IOI. Model 1 includes the behavior coding results for adequate answers only. As shown in the previous table, the behavior coding results are predictive of the IOI. The R-squared for this model is .15. There are a few other methods that predict the IOI in Table 5.8. First, problems with response categories identified by QUAID are significant predictors of the IOI. Respondent task problems found by the QAS and cognitive interviewing are also significant predictors of the IOI. The respondent task problems identified by the QAS are actually the best predictors of the IOI across all methods. The R-squared for the model with the QAS problems is .25. Table 5.9 shows what happens when the IOI is calculated with weights that penalize more severe disagreements on ordinal variables. The table shows that the QUAID and cognitive interviewing results are no longer significant. However, there is relatively little effect on the QAS or behavior coding results.

Table 5.8. Prediction of the unweighted IOI (n=53).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Intercept	41.06*(2.56)	34.44*(2.13)	38.55*(2.60)	25.05*(3.36)	33.42*(2.45)	27.74(24.25)	34.97*(3.65)
Behavior Coding							
Adequate answer >= 80%	-11.59*(3.88)						
Response difficulty							
Expert Review			-5.70(3.18)				
QAS				4.18*(1.24)			
Cog. Int.					2.56*(1.14)		
Problem with categories							
QUAID		9.45*(3.65)					
Expert Review			12.35(7.62)				
QAS				2.31(2.12)			
Cog. Int.					.43(.89)		
SQP							
Total Quality						.15(.42)	
Other methods							
Response latency							.00(.00)
R-Square	.15	.12	.09	.25	.10	.00	.01

Table 5.9. Prediction of the weighted IOI (n=53).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Intercept	37.40*(2.25)	32.54*(1.92)	35.13*(2.32)	22.55*(2.80)	31.69*(2.16)	7.88(20.66)	31.11*(3.14)
Behavior Coding							
Adequate answer >= 80%	-8.77*(3.42)						
Response difficulty							
Expert Review			-3.52(2.84)				
QAS				4.13*(1.04)			
Cog. Int.					1.74(1.01)		
Problem with categories							
QUAID		4.62(3.29)					
Expert Review			3.39(6.81)				
QAS				1.70(1.76)			
Cog. Int.					-.66(2.67)		
SQP							
Total Quality						.44(.35)	
Other methods							
Response latency							.00(.00)
R-Square	.11	.04	.03	.31	.06	.03	.02

Discussion

This chapter utilized multilevel models to understand the relationship between pretest method results and the reliability of survey questions. This chapter contributes a multiple methods evaluation of whether a variety of ex-ante, laboratory, and field methods are predictive of discrepant answers to survey questions over time. The findings from this chapter suggest that a number of the methods are related to the reliability of survey questions.

This chapter investigated two hypotheses. First, the chapter investigated the hypothesis that the methods need to be used together in order to understand data quality (Presser et al., 2004, Yan, Kreuter, and Tourangeau, 2012). The findings from this chapter strongly suggest that most of the methods used in this dissertation are complementary and are best used in combination with each other. Depending on the specific model, a combination of 3-5 methods worked best for predicting discrepant answers between the original interview and reinterview. QUAID seems to provide the most reliable way to identify problems with vague or overlapping response categories.

This chapter also tested the hypothesis that field methods would be the best predictors of discrepant answers. The evidence suggests that field methods such as behavior coding are important to understanding the reliability of survey data. In fact, this chapter confirms the findings from Hess, Singer, and Bushery (1999) that behavior coding is predictive of the reliability of survey questions. Similar to Hess, Singer, and Bushery (1999), whether or not the initial exchange results in an adequate answer is predictive of reliability. However, the overall percentage of initial exchanges that result in an adequate answer is also correlated with the results from other methods such as the findings from the QAS or cognitive interviewing. The size of this correlation for these data is approximately -.60. Hence, in models including other methods, the relationship between adequate answers and discrepant answers at the question level is partially explained by the other methods. This means that much can be learned through ex-ante or laboratory methods regarding the adequacy of the initial exchange. This is important since these methods can be conducted at a lower cost.

Other behavior coding information is not as easily explained by other methods. Long pauses and the use of speech fillers were some of the most reliable predictors of discrepant answers in this study. This is consistent with research by Draisma and Dijkstra (2004) that long pauses lead to more inaccurate answers. It is also consistent with research by Schaeffer and Maynard (2002) that behaviors such as hesitations, reports, and feedback are more effective indications of problems with questions than are behaviors such as explicit requests for clarification. A concern with this type of behavior code is that it may be harder to code than the traditional behavior codes. This was definitely the case in this study as the associated kappa statistic of this behavior code was approximately .42 compared to .80 for most other behavior codes.

CHAPTER 6: CONCLUSION

This dissertation sought to provide a better understanding of how the results from question evaluation methods relate to data quality and how to use the methods together. The results have provided some clues with respect to both of these goals. I first review the results in light of the hypotheses examined in this dissertation.

> **Hypothesis 1: (Model-based method hypothesis) There will be higher levels of agreement between the traditional methods (e.g. expert review, QAS, cognitive interviewing) than between the model-based methods (e.g. QUAID and SQP) and traditional methods.**

There have been some new methods added to the survey designer's toolkit in recent years. Computer-based tools such as QUAID and SQP may help survey designers focus on different types of problems than they have in the past. It is clear from the results in this dissertation and from past research (e.g. Yan, Kreuter, and Tourangeau, 2012a; van der Zouwen and Smit, 2004) that SQP tends to offer different conclusions from traditional question evaluation methods. Chapter 1 showed that this is also true for QUAID. There are negative correlations between problems found by QUAID and traditional methods. There is still much research that needs to be done to understand how these methods relate to data quality. The findings in this dissertation are somewhat limited by the type of questions that were examined. For example, the questions used in the field study had already been submitted to question evaluation that may have already fixed many of the problems that QUAID might have identified in the first place. It could be that a questionnaire at a less advanced stage could see a greater benefit from QUAID than was shown in this dissertation. The same could be said for SQP. It is possible that question designers would benefit from the focus on the form of the question by SQP at an earlier stage of question development.

Hypothesis 2: (Problem nature hypothesis) The rate of agreement between qualitative methods will vary by type of problem.

The results in this dissertation do support the hypothesis put forth by Yan, Kreuter, and Tourangeau (2012a) that the nature of the problem does affect the level of agreement between methods. Among the traditional methods such as expert review, QAS, and cognitive interviewing, there was much more substantial agreement on semantic problems related to the meaning of words or concepts. These correlations are in the .5-.6 range. The correlations on other types of problems across methods are much weaker or often negative. This provides evidence that the nature of the problem is important to consider when comparing methods and determining which methods should be used together. Although the results in this dissertation show agreement varies across methods similar to Yan, Kreuter, and Tourangeau (2012a), their results showed that there were higher levels of agreement on problems with recall. It is unclear what the source is for these different findings. Differences between the types of questions between studies and the procedures for coding problems with questions make it difficult to determine the source of disagreement between the studies. In any event, both studies present evidence that the techniques tend to focus on different aspects of question evaluation. Future research should continue to focus on the nature of the problems on which the methods are more likely to agree or disagree. This research can be used to either hone the evaluation process across methods so that the methods agree more closely or to provide guidelines about which methods are better at detecting each type of problem.

Hypothesis 3: (Complementary method hypothesis) Using multiple methods together will be better at predicting data quality in the field than using individual methods.

This dissertation has taught a great deal about how question evaluation methods relate to data quality. Table 6.1 provides a review of these findings from the dissertation.

Table 6.1. Summary of findings.

Field Result	Chapter	Best predictors
Percent correct	3	Expert review and cognitive interviewing problems with sensitivity
Behavior coding: Adequate answers	4	QAS problems with sensitivity and cognitive interviewing recall problems
Behavior coding: Requests for clarification	4	Expert review, QAS, and Cognitive interviewing problems with the meaning of terms or concepts
Item nonresponse	4	QAS problems with response categories and cognitive interviewing problems with recall
Response latency	4	QAS recall problems best predictor overall; Cognitive interviewing semantic problems also predict questions with longest response latencies.
Reliability	5	QUAID problems with response categories, QAS problems with recall, behavior coding adequate answers, behavior coding pauses and fillers

The results in Table 6.1 generally support the complementary method hypothesis. Multiple methods were better at predicting the results in the field than single methods.

This was true for all of indicators of data quality in the field. These results highlight that, although there is some overlap in problem detection across methods, the methods tend to make significant unique contributions that facilitate a question evaluation process that involves multiple methods rather than any single method. The results for behavior coding and item nonresponse suggest that cognitive interviewing does a better job at identifying questions where respondents may have difficulty recalling relevant information to answer a question. In contrast, the QAS was better for identifying problems with response categories. In combination, these two did the best job at predicting the level of adequate answers and item nonresponse than any individual method. This was also true in cases where two or more methods identified the same class of problems that were predictive of data quality in the field. For example, problems with sensitivity identified by QAS and cognitive interviewing were better predictors of the percent correct for questions from record check studies than the use of either method's individual results. The same is true for the prediction of requests for clarification. Problems with the meaning of terms or concepts identified by expert review, QAS, and cognitive interviewing were better than using any individual method. Prediction of reliability probably benefited the most from a multiple method evaluation. Various combinations of 3-5 methods worked best for predicting discrepant answers.

Hypothesis 4: (Test environment hypothesis) Methods that are implemented in a more realistic survey setting will be most closely related to data quality.

I only found partial support for this hypothesis. In the case of item nonresponse, the model fit for the cognitive interviewing results was better than expert review, QAS, QUAID, and SQP. However, the picture was much less clear for the other dependent variables in the dissertation. For example, when predicting adequate answers the model fit for QAS was similar to cognitive interviewing. The model fit statistics for expert review, QAS, and cognitive interviewing were similar for predicting requests for clarification. In the case of reliability, the model fit for QAS was as good as or even better than most laboratory or field methods. Hence, it is not necessarily the case that observation of the response process is necessary to identify significant problems. It ultimately depends upon the type of problems that a question is prone to and what method detects that type of problem.

Hypothesis 5: (Respondent and question problem interaction hypothesis) Respondents with lower levels of cognitive ability will have more difficulty with questions identified as problematic by ex-ante and laboratory methods than respondents with higher levels of cognitive ability.

There was also partial support for this hypothesis. Questions where the methods identified recall problems are more likely to lead to lower levels of adequate answers for those with lower levels of education compared to those with higher levels of education. A similar result was found for item nonresponse. Those with lower levels of education demonstrated higher levels of item nonresponse on questions identified as having problems with recall compared to the higher educated counterparts. These interactions suggest that some problems might have a differential impact on different types of respondents. Future research should examine this result more carefully since the goals of question evaluation are often focused on different groups of respondents such as those with lower levels of cognitive ability.

In addition to providing additional evidence about the circumstances under which question evaluation results relate to data quality, this dissertation has also shown a way forward when trying to assess the effectiveness of question evaluation methods. The growing evidence in the literature is that the methods are complementary. This requires more studies such as the present one that compare multiple methods of evaluation. The results in Table 6.1 provide some recommendations for potentially useful combinations of methods. Future comparisons should involve traditional methods such as expert review and cognitive interviewing and newer methods such as QUAID and SQP. This will help us to learn how the extra information added by these new computer-based methods can aid question designers in their task.

Finally, statistical models that are flexible enough to test different combinations of variables should preferred when comparing different methods. Multilevel models using a regression framework offer this flexibility. The model fit statistics utilized in this dissertation enable the analyst to test a multitude of hypotheses using different combinations of methods. The use of these or similar models in a confirmatory framework will over time lead to a better understanding of how the methods work together.

As with any research, this study has its limitations. One limitation refers to the mix of the methods that were used. While the intention of the dissertation was to evaluate some new and some traditional methods simultaneously, one may argue that there are other important methods that could have been evaluated. For example, there are a number of methods in Table 1.2 that could have been studied such as interviewer and respondent debriefing studies. These are often inexpensive alternatives to behavior coding that can be used to supplement conventional pretests. Future research, should examine how these studies compare with some of the methods used in this dissertation.

Another potential limitation is that many of the methods used in this dissertation are used specifically to evaluate questions for interviewer administered surveys. In particular, this was a survey administered by telephone. Although telephone surveys are becoming perhaps less popular over time, they still are widely used in the field of survey research. Nonetheless, the choice of this type of survey as focus for this dissertation may cause results to differ from other studies. For example, it is possible that results from behavior coding might be even more important in field studies where interviewer behavior is less standardized. The advent of computer audio recorded interviewing makes the possibility of observing field behavior much easier and future studies should investigate whether the mode of the interview affects the various combinations of methods that are effective. In addition, new technology has made it possible to conduct eye-tracking studies on self-administered instruments. Future studies, should examine how the results from eye-tracking on self-administered instruments compares with the traditional approaches such as cognitive interviewing.

Several decisions were made regarding the implementation of the methods in this dissertation that may affect the generalizeability of the findings. For example, students were used to conduct the cognitive interviews rather than Ph.D. level experienced survey researchers. There is some disagreement about the level of experience that is necessary for effective cognitive interviewing (Willis, Schechter, and Whitaker, 1999). In addition, a relatively high number of cognitive interviews (~52) compared to what might be common practice were conducted as part of this project. Although the literature suggests that more interviews are necessary, common practice is probably closer to 20-30 interviews for a project (Blair and Conrad, 2011). Finally, relatively unstructured individual expert reviews were also chosen for this project instead of panels of experts as in other studies (e.g. Presser and Blair, 2004). The goal of this dissertation was to compare a variety of traditional and new methods. Hence, I have tried to choose sensible approaches to each method rather than teasing out the differences in how to specifically implement the methods. As with any research, researchers need to carefully evaluate their own circumstances and determine how the findings in this dissertation apply to their specific situation.

Overall, this dissertation supports the growing body of evidence that multiple method approaches to question evaluation should be pursued (Yan, Kreuter, and Tourangeau, 2012a; Presser et al., 2004). This recommended approach is could be considered a best practice for different reasons. First, past research has shown that some of the most common methods that we use are inherently unreliable. Research should continue to identify sources of unreliability in cognitive interview data and the data resulting from other methods in order to implement the methods in a manner that leads to more reliable findings (e.g., Conrad and Blair 2004;2009). A second reason that multiple methods may be a wise choice is that the methods do have different strengths. This dissertation suggests that this is the case and has provided some guidance about how to combine the methods together. Future studies should continue to provide evidence about the best combinations of methods to use together to produce better data. This research program would add a great deal of confidence to the recommendations and guidelines that exist in the survey methodological literature on question design and evaluation.

APPENDIX A: 2006 SURVEY PRACTICUM PRETEST QUESTIONNAIRE

PART I: WAR IN IRAQ

Context 1: Costs

1. Do you think the war in Iraq has helped or hurt the image of the United States in the world?
1 HELPED
2 HURT
8 DON'T KNOW
9 REFUSED

2. Do you think Iraq will turn out to be another Vietnam, or do you think the United States will accomplish its goals in Iraq?
1 LIKE VIETNAM
2 US WILL ACCOMPLISH ITS GOALS
8 DON'T KNOW
9 REFUSED

3. Do you think that removing Saddam Hussein from power was or was not worth the lives of the American soldiers who have died in the war?
1 WAS
2 WAS NOT
8 DON'T KNOW
9 REFUSED

4. Over the next year, do you think that the U.S. military in Iraq will suffer more casualties or fewer casualties than it did in the last year?
1 MORE
2 FEWER
3 THE SAME (IF VOLUNTEERED)
8 DON'T KNOW
9 REFUSED

5. How much longer do you think the United States will have a significant number of troops in Iraq?
1 Less than a year
2 One to 3 years, or
3 More than 3 years
8 DON'T KNOW
9 REFUSED

SKIP TO Q11

Context 2: Terrorism

6. The U.S. government has been trying to prevent terrorist attack in the United States. Do you think it is doing too much to try to prevent such attacks, not enough, or the right amount?

1 TOO MUCH
2 NOT ENOUGH
3 THE RIGHT AMOUNT
8 DON'T KNOW
9 REFUSED

7. Just your best guess, do you think Osama bin Laden is currently planning a terrorist attack against the United States?

1 YES
2 NO
8 DON'T KNOW
9 REFUSED

8. Do you believe the U.S. will win the war against terrorism?

1 YES
2 NO
8 DON'T KNOW
9 REFUSED

9. How worried are you that there will be another terrorist attack on the United States in the next few months? Would you say you are…

1 Very worried
2 Somewhat worried
3 Not very worried, or
4 Not worried at all
8 DON'T KNOW
9 REFUSED

10. How worried are you that you or someone in your family will become a victim of terrorism?

1 Very worried
2 Somewhat worried
3 Not very worried, or
4 Not worried at all
8 DON'T KNOW

9 REFUSED

Target questions

11. How strongly do you favor or oppose the United States war with Iraq? Would you say you…

1 Strongly favor the war
2 Somewhat favor it
3 Neither favor nor oppose it
4 Somewhat oppose the war, or
5 Strongly oppose it?
8 DON’T KNOW
9 REFUSED

12a. Did that answer come to you immediately or did you have to think for a moment before you answered?

1 IMMEDIATELY
2 HAD TO THINK FOR A MOMENT
8 DON’T KNOW
9 REFUSED

IF Q12a=2, ASK: 12b. What thoughts or feelings came to mind as you decided how to answer?

____

13. Do you think the United States made a mistake in sending troops to Iraq, or not?

1 MISTAKE
2 NOT A MISTAKE
8 DON’T KNOW
9 REFUSED

14a. Did that answer come to you immediately or did you have to think for a moment before you answered?

1 IMMEDIATELY
2 HAD TO THINK FOR A MOMENT
8 DON’T KNOW
9 REFUSED

IF Q14a=2, ASK: 14b. What thoughts or feelings came to mind as you decided how to answer?

____

15. How important is the Iraq war to you personally—very important, somewhat important, not too important, or not important at all?

1 VERY IMPORTANT
2 SOMEWHAT IMPORTANT
3 NOT TOO IMPORTANT
4 NOT IMPORTANT AT ALL
8 DON’T KNOW
9 REFUSED

16. Would you say that you are strongly on one side or the other on the Iraq war or would you say your feelings are mixed?

1 STRONGLY ON ONE SIDE
2 MIXED
8 DON’T KNOW
9 REFUSED

PART II: WIRETAPPING

Context 1: Security

17. How important do you think wiretapping and other covert intelligence gathering efforts are in maintaining the security of the United States?
1 Very important
2 Somewhat important
3 Not very important, or
4 Not at all important
8 DON'T KNOW
9 REFUSED

18. Do you approve or disapprove of the government's monitoring of suspicious telephone calls in the United States as a way to reduce the threat of terrorism?

1 APPROVE
2 DISAPPROVE
8 DON'T KNOW
9 REFUSED

SKIP TO Q21

Context 2: Privacy

19. How important is the right to privacy to you personally
1 Very important
2 Somewhat important
3 Not very important, or
4 Not at all important
8 DON'T KNOW
9 REFUSED

20. How concerned are you about losing some of your civil liberties as a result of the steps taken by the Bush Administration to fight terrorism?
1 Very concerned
2 Somewhat concerned
3 Not very concerned, or
4 Not at all concerned
8 DON'T KNOW
9 REFUSED

Target question

21. Do you feel the President is justified in authorizing wiretaps without prior court approval?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

22a. Did that answer come to you immediately or did you have to think for a moment before you answered?
1 IMMEDIATELY
2 HAD TO THINK FOR A MOMENT
8 DON'T KNOW
9 REFUSED

IF Q22a=2, ASK: 22b. What thoughts or feelings came to mind as you decided how to answer?

____

23. How important is the wiretapping issue to you personally—very important, somewhat important, not too important, or not important at all?

1 VERY IMPORTANT
2 SOMEWHAT IMPORTANT
3 NOT TOO IMPORTANT
4 NOT IMPORTANT AT ALL
8 DON'T KNOW
9 REFUSED

24. Would you say that you are strongly on one side or the other on the wiretapping issue or would you say your feelings are mixed?

1 STRONGLY ON ONE SIDE
2 MIXED
8 DON'T KNOW
9 REFUSED

PART III: EDUCATION

Context 1: Mathematics

25. How important do you think mathematics training in our schools is to the economic competitiveness of the United States?

1 Very important
2 Somewhat important
3 Not very important, or
4 Not at all important
8 DON'T KNOW
9 REFUSED

26. When recent school graduates look for jobs, how important do you think their mathematics skills are?

1 Very important
2 Somewhat important
3 Not very important, or
4 Not at all important
8 DON'T KNOW
9 REFUSED

27. How much do you think increases in the quality of life in America depend on people having a high level of training in mathematics and science? Would you say increases in the quality of life depend…

1 A lot on mathematics and science training
2 Somewhat
3 Not much, or
4 Not at all
8 DON'T KNOW
9 REFUSED

28. Now I'd like you to compare the mathematics skills of American students to those of students in other developed countries. Would you say American students are as skilled, more skilled, or less skilled in mathematics than students in other developed countries?

1 AS SKILLED
2 MORE SKILLED
3 LESS SKILLED
8 DON'T KNOW
9 REFUSED

SKIP TO Q33

Context 2: Reading and language

29. How important do you think reading and language training in our schools is to the economic competitiveness of the United States?

1 Very important
2 Somewhat important
3 Not very important, or
4 Not at all important
8 DON'T KNOW
9 REFUSED

30. When recent school graduates look for jobs, how important do you think their reading and language skills are?

1 Very important
2 Somewhat important
3 Not very important, or
4 Not at all important
8 DON'T KNOW
9 REFUSED

31. How much do you think increases in the quality of life in America depend on people having a high level of training in reading and language? Would you say increases in the quality of life depend…

1 A lot on reading and language training
2 Somewhat
3 Not much, or
4 Not at all
8 DON'T KNOW
9 REFUSED

32. Now I'd like you to compare the reading and language skills of American students to those of students in other developed countries. Would you say American students are as skilled, more skilled, or less skilled in reading and language than students in other developed countries?

1 AS SKILLED
2 MORE SKILLED
3 LESS SKILLED
8 DON'T KNOW
9 REFUSED

Target question

33. Some educators have proposed two new programs for fourth-grade students in the United States. One program would require schools to add two 60 minute weekly practice sessions for improving mathematics skills. The other program would require schools to add two 60 minute weekly practice sessions for improving reading and language skills.

If there were only resources for one of these programs, which would you prefer – the mathematics program or the reading and language program?

1 MATHEMATICS
2 READING AND LANGUAGE
8 DON'T KNOW
9 REFUSED

34a. Did that answer come to you immediately or did you have to think for a moment before you answered?

1 IMMEDIATELY
2 HAD TO THINK FOR A MOMENT
8 DON'T KNOW
9 REFUSED

IF Q34a=2, ASK: 34b. What thoughts or feelings came to mind as you decided how to answer?

____

35. How important is this issue to you personally—very important, somewhat important, not too important, or not important at all?

1 VERY IMPORTANT
2 SOMEWHAT IMPORTANT
3 NOT TOO IMPORTANT
4 NOT IMPORTANT AT ALL
8 DON'T KNOW
9 REFUSED

36. Would you say that you are strongly on one side or the other on this issue or would you say your feelings about it are mixed?

1 STRONGLY ON ONE SIDE
2 MIXED
8 DON'T KNOW
9 REFUSED

PART IV: HEALTH

Now, I would like to ask you a few questions about how you feel these days.

Context 1. Sickness (conditions from NHIS)

37. Have you ever been told by a doctor or other health professional that you had arthritis, also called rheumatism?
1 YES
2 NO
8 DON'T KNOW 9 REFUSED

38. Have you ever been told by a doctor or other health professional that you had a heart problem?
1 YES
2 NO
8 DON'T KNOW 9 REFUSED

39. Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?
1 YES
2 NO
8 DON'T KNOW 9 REFUSED

40. Have you ever been told by a doctor or other health professional that you had diabetes?
1 YES
2 NO
8 DON'T KNOW 9 REFUSED

41. Have you ever been told by a doctor or other health professional that you had a kidney, bladder, or renal problem?
1 YES
2 NO
8 DON'T KNOW 9 REFUSED

42. Have you ever been told by a doctor or other health professional that you had Multiple Sclerosis (MS), or Muscular Dystrophy (MD)?
1 YES
2 NO
8 DON'T KNOW 9 REFUSED

SKIP TO Q44

Context 2. Neutral

43. How satisfied are you currently with your life as a whole? Would you say you are…

1 Very satisfied
2 Somewhat satisfied
3 Neither satisfied nor dissatisfied
4 Somewhat dissatisfied, or
5 Very dissatisfied
8 DON'T KNOW
9 REFUSED

Target question

44. Would you say that in general your health is …

1 Excellent
2 Very good
3 Good
4 Fair, or
5 Poor
8 DON'T KNOW
9 REFUSED

45a. Did that answer come to you immediately or did you have to think for a moment before you answered?

1 IMMEDIATELY
2 HAD TO THINK FOR A MOMENT
8 DON'T KNOW
9 REFUSED

IF Q45a=2, ASK: 45b. What thoughts or feelings came to mind as you decided how to answer?

____

PART V: DOCTOR VISITS

The next few questions are about some other aspects of your life.

<u>Context 1: Rates</u>

46. In the last 12 months, about how often did you go to the movies? Would you say…

1 At least once a week
2 A few times a month
3 A few times a year
4 Once or twice, or
5 Never
8 DON'T KNOW
9 REFUSED

47. In the last 12 months, about how often did you eat in a restaurant? Would you say…

1 At least once a week
2 A few times a month
3 A few times a year
4 Once or twice, or
5 Never
8 DON'T KNOW
9 REFUSED

48. In the last 12 months, about how often did you exercise? Would you say…

1 At least once a week
2 A few times a month
3 A few times a year
4 Once or twice, or
5 Never
8 DON'T KNOW
9 REFUSED

SKIP TO Q48

Context 2: Counts

49. In the last 30 days, how many times did you go to the movies?

NUMBER OF TIMES: _______ 8 DON’T KNOW 9 REFUSED

50. In the last 30 days, how many times did you eat in a restaurant?

NUMBER OF TIMES: _______ 8 DON’T KNOW 9 REFUSED

51. In the last 30 days, how many times did you exercise?

NUMBER OF TIMES: _______ 8 DON’T KNOW 9 REFUSED

Target question

52. During the past 6 months, how many times have you seen a doctor or other health care professional about your own health at a doctor's office, a clinic, or some other place?

NUMBER OF TIMES: _______ 8 DON’T KNOW 9 REFUSED

53. How did you arrive at your answer?

____

PART VI: DEMOGRAPHICS

Finally I have a few questions about your background.

54. In what year were you born?

YEAR: __________   8 DON'T KNOW   9 REFUSED

55. What is the highest level of education that you have completed?
1 NONE, OR GRADE 1-8
2 HIGH SCHOOL INCOMPLETE (GRADES 9-11)
3 HIGH SCHOOL GRADUATE (GRADE 12 OR GED)
4 BUSINESS, TECHNICAL, OR VOCATIONAL SCHOOL NOT INCL.UDING HIGH SCHOOL
5 SOME COLLEGE, NO 4-YEAR DEGREE
6 COLLEGE GRADUATE, (B.S., B.A., OTHER 4-YR. DEGREE)
7 POST-GRADUATE TRAINING OR PROFESSIONAL SCHOOLING AFTER COLLEGE (E.G., TOWARD A MASTER'S DEGREE OR PH.D.; LAW OR MEDICAL SCHOOL)
8 DON'T KNOW
9 REFUSED

56. [ASK ONLY IF NOT OBVIOUS] Are you male or female?
1. MALE
2. FEMALE
8 DON'T KNOW
9 REFUSED

57. Are you Spanish, Hispanic, or Latino?
1 YES – SPANISH, HISPANIC, OR LATINO (FOR EXAMPLE: CHICANO, CUBAN, MEXICAN, MEXICAN-AMERICAN, PUERTO RICAN, ETC.)
2 NO – NONE OF THESE CATEGORIES APPLY
8 DON'T KNOW
9 REFUSED

58. What is your race? Would you say you are…
[CODE ALL THAT APPLY – READ EXAMPLES IN ITALICS IF NECESSARY]
1 White,
2 Black or African-American,
3 Asian, *(includes: Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese)*
4 Pacific Islander, *(includes: Native Hawaiian, Guamanian, Samoan)*
5 American Indian or Alaska Native, or
6 Some other race? (SPECIFY)__________________________________
8 DON'T KNOW
9 REFUSED

APPENDIX B: 2006 SURVEY PRACTICUM FIELDED COMBINED QUESTIONNAIRE FOR WAVES 1 AND 2

NOTES FOR THE INTERVIEWER: RESPONSE OPTIONS THAT SHOULD NOT BE READ TO THE RESPONDENT ARE IN CAPITAL LETTERS.

NOTES FOR THE PROGRAMMER.

PROGRAMMING NOTES ARE IN SQUARE BRACKETS.
RANDOM VARIABLES FOR EXPERIMENTAL MANIPULATIONS ARE NAMED RAND1, RAND2 etc.
PLEASE PUT A TIME STAMP AFTER EACH ITEM.

WAVE 2 RANDOMIZATIONS:

- VARIABLES RAND 1, 3, 5, 6, 9, 11, AND 13 HAVE THE SAME VALUES IN BOTH WAVES (I.E. A RESPONDENT GETS THE SAME VALUE IN BOTH WAVES)
- VARIABLES RAND 2, 4, 7, 8, 10, AND 12 SYSTEMATICALLY CHANGE VALUES ACROSS WAVES FOR SOME RESPONDENTS; SEE MORE DETAILS NEXT TO EACH OF THE VARIABLES.

SCREENER – WAVE 1 ONLY

SINTRO_1
Hello, my name is _____ and I'm calling for a University of Maryland research study about people's opinions on current social issues.

RESIDENTIAL
Are you a member of this household and at least 18 years old?

1. YES (GO TO SINTRO_3)
2. NO (ASK TO SPEAK WITH HHM 18+)
3. PROBABLE BUSINESS (GO TO SINTRO_3)

AM. ANSWERING MACHINE
RT. RETRY AUTODIALER
NW. NONWORKING, DISCONNECTED, CHANGED
GT. GO TO RESULT

SINTRO_3
Is this phone number used for…

4. Home use,
5. Home and business use, or
6. Business use only? [READ: "Thank you, but we are only interviewing in private residences. Good-bye." CODE NR RESULT]

GT. GO TO RESULT

GETNAME

Your participation is voluntary and all of your answers will be kept completely confidential.

We would like to interview the adult member of your household who had the most recent birthday. Would you please give me this person' first name so I know who to ask for should I need to call back?

[IF FIRST NAME REFUSED OR DON'T KNOW, ASK FOR INITIALS, AGE/SEX, RELATION, OR OTHER IDENTIFYING INFORMATION.]

	[X BY RESP.]
FIRST NAME: _______________	( )

<FENCEPOST – END SCREENER>

IF WAVE 1, GO TO INTRO_W1
IF WAVE 2, GO TO INTRO_W2

INTRO_W1
The questions usually take less than 15 minutes. Participation in this survey is voluntary, and all of your answers will be kept completely confidential. If we come to a question at any time that you do not want to answer, please just tell me and we will go on to the next question.
[GO TO REC_PERM]

INTRO_W2
Hello, my name is _____ and I'm calling for a University of Maryland research study about people's opinions on current social issues. I would like to speak with [INSERT RESPONDENT'S NAME].

We spoke to you a couple of weeks ago. Just to remind you, participation in this survey is voluntary, and all of your answers will be kept completely confidential. If we come to a question at any time that you do not want to answer, please just tell me and we will go on to the next question.

IF THE RESPONDENT HESITATES:
[IF **RAND13** WAS A, INPUT SENTENCE A. IF RAND13 WAS B, INPUT SENTENCE B.]

A. Unfortunately, the information you've already provided to us will be much less valuable unless you complete the second interview.

B. The information you've already provided to us will be a lot more valuable if you complete the second interview.

REC_PERM
This interview will be recorded for quality control and training purposes.
[IF NEEDED: You may ask me to stop the tape at any time during the interview.]

1. RECORD INTERVIEW
2. DO NOT RECORD INTERVIEW [READ: OK, that's fine. Let's continue.]

[IF RESPONDENT REQUESTS LATER TO HAVE RECORDING TURNED OFF, INTERVIEWER WILL DO SO AND READ FOLLOWING SCRIPT:

"This interview will not be recorded"

IF RECORDING IS TURNED OFF, OR IF RESPONDENT DOES NOT AGREE TO INITIAL STATEMENT, VARIABLE "RECRDINT" WILL EQUAL 2; ELSE IF RECORDING IS AGREED TO AND NEVER TURNED OFF, RECRDINT=1]

[NOTE – RECORDING STATEMENT IS READ EACH TIME RESPONDENT IS CONTACTED FOR ANY PORTION OF THE INTERVIEW]

[**RAND1.** FOR RANDOM HALF OF THE RESPONDENTS, SWITCH PART ‘A: WAR IN IRAQ’ WITH PART ‘D: EDUCATION’]

[**WAVE 2 RAND1:** ASSIGN THE RESPONDENTS TO THE SAME GROUPS AS IN WAVE 1]

PART A: WAR IN IRAQ

The {first/next} questions are about issues that have been in the news.
[USE ‘first’ IF STARTING WITH SECTION A, USE ‘next’ IF STARTED WITH SECTION D]

[**RAND2.** A RANDOM HALF OF THE RESPONDENTS SHOULD GET QUESTIONS Q1-Q4, AND ANOTHER HALF Q6-Q9]

[**WAVE 2 RAND2:** A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q1-Q4 IN WAVE 1 SHOULD NOW GET QUESTIONS Q6-Q9; A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q6-Q9 IN WAVE 1 SHOULD NOW GET QUESTIONS Q1-Q4; THE OTHERS SHOULD GET THE SAME QUESTIONS AS IN WAVE 1]

Context 1: Costs

1. Do you think the war in Iraq has helped, hurt, or had no effect on the image of the United States in the world?
1 HELPED
2 HURT
3 HAD NO EFFECT
-8 DON’T KNOW
-7 REFUSED

2. Do you think the Iraq war will turn out to be another Vietnam?
1 YES
2 NO
-8 DON’T KNOW
-7 REFUSED

3. Over the next year, do you think that the U.S. military in Iraq will suffer more casualties or fewer casualties than it did in the last year?
1 MORE
2 FEWER
3 THE SAME (IF VOLUNTEERED)
-8 DON’T KNOW
-7 REFUSED

4. When do you think the United States will withdraw all of its troops from Iraq? Would you say…
1 in less than a year,
2 one to 3 years from now, or
3 more than 3 years from now?
-8 DON'T KNOW
-7 REFUSED

[SKIP TO Q10]

Context 2: Terrorism

6. Do you think Osama bin Laden is currently planning an attack against the United States?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

7. Do you believe the U.S. and its allies will defeat the Al Qaeda terrorist network?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

8. How worried are you that there will be another terrorist attack on the United States? Would you say you are…
1 very worried,
2 somewhat worried,
3 not very worried, or
4 not worried at all?
-8 DON'T KNOW
-7 REFUSED

9. How worried are you that you or someone in your family will become a victim of terrorism in the United States? Would you say you are…
1 very worried,
2 somewhat worried,
3 not very worried, or
4 not worried at all?
-8 DON'T KNOW
-7 REFUSED

Target questions

[**RAND3.** A RANDOM HALF OF THE RESPONDENTS SHOULD GET Q10A, AND ANOTHER HALF Q10B.]

[**WAVE 2 RAND3.** ASSIGN THE RESPONDENTS TO THE SAME GROUPS AS IN WAVE 1]

<u>NOTE TO THE INTERVIEWER</u>: WRITE DOWN EVERYTHING (IF ANYTHING) THE RESPONDENT SAYS WHILE FORMULATING THE ANSWER TO THE FOLLOWING QUESTION ON YOUR HARDCOPY NOTES FORM.

10a. How do you now feel about continued U.S. military involvement in the Iraq war?
Do you…
1 strongly favor it,
2 somewhat favor it,
3 somewhat oppose it, or
4 strongly oppose it?
-8 DON'T KNOW
-7 REFUSED

10b. How do you now feel about continued U.S. military involvement in the Iraq war?
Do you…
1 favor, or
2 oppose it?
-8 DON'T KNOW
-7 REFUSED

Q10AB
[DID YOU WRITE DOWN ANYTHING R SAID IN FORMULATING A RESPONSE?]

1. YES
2. NO (SKIP TO Q11B)

[IF R SAID SOMETHING BUT REFUSED TO ANSWER (10A/10B=REFUSED) – IF WAVE 1 SKIP TO Q12, IF WAVE 2 SKIP TO NEXT SECTION.]

[NOTE: NOTES WRITTEN BY INTERVIEWERS DURING TARGET QUESTIONS ARE KEY-ENTERED AFTER RESPONDENT IS OFF PHONE – SEE END OF QUESTIONNAIRE FOR INFORMATION RE: VARIABLES FOR EACH ITEM.]

11a. [IF R SAID SOMETHING WHILE FORMULATING THE ANSWER TO Q10a/b] You said: (READ FROM HARDCOPY). As you decided how to answer the last question, did you have any additional thoughts or feelings?

1. YES
2. NO (SKIP TO 12; IF WAVE 2 SKIP TO NEXT SECTION)

(RECORD VERBATIM)

____

____

[A11TXT1 – A11TXT6 USED TO STORE TEXT FROM EITHER A11A OR A11B]

11b. [IF R SAID NOTHING WHILE FORMULATING THE ANSWER TO Q10a/b] As you decided how to answer the last question, what thoughts or feelings went through your mind?

(RECORD VERBATIM AND READ BACK TO THE RESPONDENT)
[IF NONE, RECORD 'NONE'.]

____

____

[IF WAVE 2 SKIP TO NEXT SECTION]

12. How important is the Iraq war to <u>you</u>? Would you say…
1 very important,
2 somewhat important,
3 not too important, or
4 not important at all?
-8 DON'T KNOW
-7 REFUSED

13. Would you say your views on the Iraq war are mainly on one side of the issue, or are your views about this issue mixed?
1 MAINLY ON ONE SIDE
2 MIXED
-8 DON'T KNOW
-7 REFUSED

14. Do you have any family members or close friends who are serving or did serve in Iraq?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

<FENCEPOST – END SECTION A>

PART B: WIRETAPPING

[**RAND4.** A RANDOM HALF OF THE RESPONDENTS SHOULD GET QUESTIONS Q15-Q16, AND ANOTHER HALF Q17-Q18]
[**WAVE 2 RAND4.** A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q15-Q16 IN WAVE 1 SHOULD NOW GET QUESTIONS Q17-Q18; A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q17-Q18 IN WAVE 1 SHOULD NOW GET QUESTIONS Q15-Q16; THE OTHERS SHOULD GET THE SAME QUESTIONS AS IN WAVE 1]

Context 1: Security
Now I'd like to ask you a few questions about national security.

15. How important do you think wiretapping is in maintaining the security of the United States? Would you say…
1 very important,
2 somewhat important,
3 not very important, or
4 not at all important?
-8 DON'T KNOW
-7 REFUSED

16. How do you feel about the government's monitoring of telephone calls in the United States as a way to reduce the threat of terrorism? Would you say that you…
1 strongly approve it,
2 somewhat approve it,
3 neither approve nor disapprove,
4 somewhat disapprove, or
5 strongly disapprove it?
-8 DON'T KNOW
-7 REFUSED

[SKIP TO Q19]

Context 2: Privacy
Now I'd like to ask you a few questions about privacy.

17. How important is it to you that the government protects Americans' right to privacy? Is it…
1 very important,
2 somewhat important,
3 not very important, or
4 not at all important?
-8 DON'T KNOW
-7 REFUSED

18. How concerned are you about losing your right to privacy as a result of the steps taken by the government to fight terrorism? Are you…
1 very concerned,
2 somewhat concerned,
3 not very concerned, or
4 not at all concerned?
-8 DON'T KNOW
-7 REFUSED

Target questions

[**RAND5.** A RANDOM HALF OF THE RESPONDENTS SHOULD GET Q19A, AND ANOTHER HALF Q19B]

[**WAVE 2 RAND5.** ASSIGN THE RESPONDENTS TO THE SAME GROUPS AS IN WAVE 1]

<u>NOTE TO THE INTERVIEWER</u>: WRITE DOWN VERBATIM EVERYTHING (IF ANYTHING) THE RESPONDENT SAYS WHILE FORMULATING THE ANSWER TO THE FOLLOWING QUESTION ON YOUR HARDCOPY NOTES FORM.

19a. How much do you favor or oppose the President authorizing wiretaps of Americans without prior court approval? Would you say you…
1 strongly favor it,
2 somewhat favor,
3 somewhat oppose, or
4 strongly oppose it?
-8 DON'T KNOW
-7 REFUSED

19b. Do you favor or oppose the President authorizing wiretaps of Americans without prior court approval?
1 FAVOR
2 OPPOSE
3 NEITHER FAVOR NOR OPPOSE (IF VOLUNTEERED)
-8 DON'T KNOW
-7 REFUSED

Q19AB
[DID YOU WRITE DOWN ANYTHING R SAID IN FORMULATING A RESPONSE?]

1. YES
2. NO (SKIP TO Q20B)

[IF R SAID SOMETHING BUT REFUSED TO ANSWER (19A/19B REFUSED) – SKIP TO Q21; IF WAVE 2 GO TO NEXT SECTION.]

20a. IF R SAID SOMETHING WHILE FORMULATING THE ANSWER TO Q19a/b: You said: (READ FROM HARDCOPY). As you decided how to answer the last question, did you have any additional thoughts or feelings?

1. YES
2. NO (SKIP TO 21; IF WAVE 2 GO TO NEXT SECTION)

(RECORD VERBATIM)

____

____

[B20TXT1 – B20TXT6 USED TO STORE TEXT FROM EITHER B20A OR B20B]

20b. IF R SAID NOTHING WHILE FORMULATING THE ANSWER TO Q19a/b: As you decided how to answer the last question, what thoughts or feelings went through your mind?
(RECORD VERBATIM AND READ BACK TO THE RESPONDENT)
[IF NONE, RECORD 'NONE'.]

____

____

[IF WAVE 2, GO TO NEXT SECTION]

21. How important is the wiretapping issue to <u>you</u>? Is it…
1 very important,
2 somewhat important,
3 not too important, OR
4 not important at all?
-8 DON'T KNOW
-7 REFUSED

22. Would you say your views on the wiretapping issue are mainly on one side of the issue, or are your views about this issue mixed?
1 MAINLY ON ONE SIDE
2 MIXED
-8 DON'T KNOW
-7 REFUSED

<FENCEPOST – END SECTION B>

PART C: POLITICAL INVOLVEMENT (NES)

The next few questions are about community involvement.

[**RAND6.** ONE THIRD SHOULD BE ASKED Q23-Q26; ONE THIRD Q27-Q30; AND ONE THIRD Q31-Q34.]

[**WAVE 2 RAND6.** ASSIGN THE RESPONDENTS TO THE SAME GROUPS AS IN WAVE 1]

Version 1: Old questions

23. During the last two years, did you work as a volunteer for a political candidate running for national, state, or local office and got no pay at all or only a very small amount of pay for your work?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

24. During the last two years, did you contribute money to a political candidate, a political party, a political action committee, or any other organization that supported political candidates?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

25. During the last two years, did you work with others in your community or neighborhood to deal with some issue or problem?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

26. During the last two years, did you contact a government official in person, by phone, or by letter about a problem or issue?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

[SKIP TO INTRO BEFORE Q35]

Version 2: New questions

27. During the last two years, did you ever work as a volunteer for a political candidate running for national, state, or local office and got no pay at all or only a very small amount of pay for your work, or did you never do this?
1 DID
2 NEVER DID
-8 DON'T KNOW
-7 REFUSED

28. During the last two years, did you ever contribute money to a political candidate, a political party, a political action committee, or any other organization that supported political candidates, or did you never do this?
1 DID
2 NEVER DID
-8 DON'T KNOW
-7 REFUSED

29. During the last two years, did you ever work with others in your community or neighborhood to deal with some issue or problem, or did you never do this?
1 DID
2 NEVER DID
-8 DON'T KNOW
-7 REFUSED

30. During the last two years, did you ever contact a government official in person, by phone, or by letter about a problem or issue, or did you never do this?
1 DID
2 NEVER DID
-8 DON'T KNOW
-7 REFUSED

[SKIP TO INTRO BEFORE Q35]

Version 3: Modified new questions

31. During the last two years, did you or did you not work as a volunteer for a political candidate running for national, state, or local office and got no pay at all or only a very small amount of pay for your work?
1 DID
2 DID NOT
-8 DON'T KNOW
-7 REFUSED

32. During the last two years, did you or did you not contribute money to a political candidate, a political party, a political action committee, or any other organization that supported political candidates?
1 DID
2 DID NOT
-8 DON'T KNOW
-7 REFUSED

33. During the last two years, did you or did you not work with others in your community or neighborhood to deal with some issue or problem?
1 DID
2 DID NOT
-8 DON'T KNOW
-7 REFUSED

34. During the last two years, did you or did you not contact a government official in person, by phone, or by letter about a problem or issue?
1 DID
2 DID NOT
-8 DON'T KNOW
-7 REFUSED

<FENCEPOST – END SECTION C>

PART D: EDUCATION

[IF STARTED WITH SECTION A] Now I'd like to ask you some questions about schooling.
[IF STARTED WITH SECTION D] First, I'd like to ask you some questions about schooling.

[**RAND7.** A RANDOM HALF OF THE RESPONDENTS SHOULD GET QUESTIONS Q35-Q38, AND ANOTHER HALF Q39-Q42]

[**WAVE 2 RAND7.** A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q35-Q38 IN WAVE 1 SHOULD NOW GET QUESTIONS Q39-Q42; A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q39-Q42 IN WAVE 1 SHOULD NOW GET QUESTIONS Q35-Q38; THE OTHERS SHOULD GET THE SAME QUESTIONS AS IN WAVE 1]

Context 1: Mathematics

35. How important do you think mathematics training in our elementary schools is to the economic success of the United States? Would you say…
1 very important,
2 somewhat important,
3 not very important, or
4 not at all important?
-8 DON'T KNOW
-7 REFUSED

36. Do you think having good mathematics skills has a positive effect, a negative effect, or no effect at all on the job opportunities available to a recent high school graduate?
1 POSITIVE EFFECT
2 NEGATIVE EFFECT
3 NO EFFECT
-8 DON'T KNOW
-7 REFUSED

37. How much do you think a person's standard of living in America depends on having good mathematics skills? Would you say a person's standard of living depends…
1 a lot on math skills,
2 somewhat,
3 not much, or
4 not at all on math skills?
-8 DON'T KNOW
-7 REFUSED

38. Do you think that the mathematics skills of American elementary school students are better, worse, or about the same as those of elementary school students in countries such as Singapore and Japan?
1 BETTER
2 WORSE
3 ABOUT THE SAME
-8 DON'T KNOW
-7 REFUSED

[SKIP TO INTRO BEFORE Q43]

Context 2: Reading and writing

39. How important do you think reading and writing training in our elementary schools is to the economic success of the United States? Would you say…
1 very important,
2 somewhat important,
3 not very important, or
4 not at all important?
-8 DON'T KNOW
-7 REFUSED

40. Do you think having good reading and writing skills has a positive effect, a negative effect, or no effect at all on the job opportunities available to a recent high school graduate?
1 POSITIVE EFFECT
2 NEGATIVE EFFECT
3 NO EFFECT
-8 DON'T KNOW
-7 REFUSED

41. How much do you think a person's standard of living in America depends on having good reading and writing skills? Would you say a person's standard of living depends…
1 a lot on reading and writing skills,
2 somewhat,
3 not much, or
4 not at all on reading and writing skills?
-8 DON'T KNOW
-7 REFUSED

42. Do you think that the reading and writing skills of American elementary school students are better, worse, or about the same as those of elementary school students in countries such as Singapore and Japan?
1 BETTER
2 WORSE
3 ABOUT THE SAME
-8 DON'T KNOW
-7 REFUSED

Target questions

Some educators have proposed new programs for fourth-grade students in the United States. One program would require schools to add practice sessions for improving mathematics skills. The other program would require schools to add practice sessions for improving reading and writing skills.

<u>NOTE TO THE INTERVIEWER</u>: WRITE DOWN VERBATIM EVERYTHING (IF ANYTHING) THE RESPONDENT SAYS WHILE FORMULATING THE ANSWER TO THE FOLLOWING QUESTION ON YOUR HARDCOPY NOTES FORM.

43. If there were only resources for <u>only one</u> of these programs, which would you prefer – the mathematics program or the reading and writing program?
1 MATHEMATICS
2 READING AND WRITING
-8 DON'T KNOW
-7 REFUSED

Q43A
[DID YOU WRITE DOWN ANYTHING R SAID IN FORMULATING A RESPONSE?]

1. YES
2. NO (SKIP TO Q44B)

[IF R SAID SOMETHING BUT REFUSED TO ANSWER (43=REFUSED) – SKIP TO Q45; IF WAVE 2 SKIP TO NEXT SECTION.]

44a. IF R SAID SOMETHING WHILE FORMULATING THE ANSWER TO Q43: You said: (READ FROM HARDCOPY). As you decided how to answer the last question, did you have any additional thoughts or feelings?

1. YES
2. NO (SKIP TO 45; IF WAVE 2 SKIP TO NEXT SECTION)

(RECORD VERBATIM)

____

____

[D44TXT1 – D44TXT6 USED TO STORE TEXT FROM EITHER D44A OR D44B]

44b. IF R SAID NOTHING WHILE FORMULATING THE ANSWER TO Q43:
As you decided how to answer the last question, what thoughts or feelings went through your mind?
(RECORD VERBATIM AND READ BACK TO THE RESPONDENT)
[IF NONE, RECORD 'NONE'.]

____

____

[IF WAVE 2, GO TO NEXT SECTION]

45. How important is this choice between mathematics versus reading and writing to <u>you</u>? Is it…
1 very important,
2 somewhat important,
3 not too important, or
4 not important at all?
-8 DON'T KNOW
-7 REFUSED

46. Would you say your views on the choice between more attention to mathematics versus reading and writing are mainly on one side of the issue, or are your views about this issue mixed?
1 MAINLY ON ONE SIDE
2 MIXED
-8 DON'T KNOW
-7 REFUSED

47. Are you or anyone in your household currently employed by a school or educational institution?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

<FENCEPOST – END SECTION D>

PART E: HEALTH

Now, let me turn to a different subject.

[**RAND8.** A RANDOM HALF OF THE RESPONDENTS SHOULD GET QUESTIONS Q48-Q53, AND ANOTHER HALF Q54]

[**WAVE 2 RAND8.** A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q48-Q53 IN WAVE 1 SHOULD NOW GET QUESTION Q54; A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTION Q54 IN WAVE 1 SHOULD NOW GET QUESTIONS Q48-Q53; THE OTHERS SHOULD GET THE SAME QUESTIONS AS IN WAVE 1]

Context 1. Sickness

48. Have you ever been told by a doctor or other health professional that you had arthritis, also called rheumatism?
1 YES
2 NO
-8 DON’T KNOW
-7 REFUSED

49. Have you ever been told by a doctor or other health professional that you had a heart problem?
1 YES
2 NO
-8 DON’T KNOW
-7 REFUSED

50. Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?
1 YES
2 NO
-8 DON’T KNOW
-7 REFUSED

51. Have you ever been told by a doctor or other health professional that you had diabetes?
1 YES
2 NO
-8 DON’T KNOW
-7 REFUSED

52. Have you ever been told by a doctor or other health professional that you had a kidney, bladder, or renal problem?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

53. Have you ever been told by a doctor or other health professional that you had Multiple Sclerosis (MS), or Muscular Dystrophy (MD)?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

[SKIP TO Q55]

Context 2. Neutral

54. How satisfied are you currently with your life as a whole? Would you say you are…
1 very satisfied,
2 somewhat satisfied,
3 neither satisfied nor dissatisfied,
4 somewhat dissatisfied, or
5 very dissatisfied?
-8 DON'T KNOW
-7 REFUSED

Target questions

[**RAND9.** TWO THIRDS OF THE RESPONDENTS SHOULD GET QUESTION Q55a, AND A THIRD Q55b]
[**WAVE 2 RAND9.** ASSIGN THE RESPONDENTS TO THE SAME GROUPS AS IN WAVE 1]

NOTE TO THE INTERVIEWER: WRITE DOWN VERBATIM EVERYTHING (IF ANYTHING) THE RESPONDENT SAYS WHILE FORMULATING THE ANSWER TO THE FOLLOWING QUESTION ON YOUR HARDCOPY NOTES FORM.

55a. Would you say that your physical health in general is …
1 excellent,
2 very good,
3 good,
4 fair, or
5 poor?
-8 DON'T KNOW
-7 REFUSED

55b. Would you say that your health in general is …
1 excellent,
2 very good,
3 good,
4 fair, or
5 poor?
-8 DON'T KNOW
-7 REFUSED

Q55AB
[DID YOU WRITE DOWN ANYTHING R SAID IN FORMULATING A RESPONSE?]

1. YES
2. NO (SKIP TO Q56B)

[IF R SAID SOMETHING BUT REFUSED TO ANSWER (55A/55B=REFUSED)– SKIP TO INTRO BEFORE 57.]

56a. IF R SAID SOMETHING WHILE FORMULATING THE ANSWER TO Q55a/b:
You said: (READ FROM HARDCOPY). As you decided how to answer the last question, did you have any additional thoughts or feelings?

1. YES
2. NO (SKIP TO INTRO BEFORE 57)

(RECORD VERBATIM)

____

____

[E56TXT1 – E56TXT6 USED TO STORE TEXT FROM EITHER E56A OR E56B]

[GO TO INTRO BEFORE 57]

56b. IF R SAID NOTHING WHILE FORMULATING THE ANSWER TO Q55a/b:
As you decided how to answer the last question, what thoughts or feelings went through your mind?
(RECORD VERBATIM AND READ BACK TO THE RESPONDENT)
[IF NONE, RECORD 'NONE'.]

____

____

<FENCEPOST – END SECTION E>

PART F: DOCTOR VISITS

The next few questions are about some other aspects of your life.

[**RAND10.** A RANDOM HALF OF THE RESPONDENTS SHOULD GET QUESTIONS Q57-Q59, AND ANOTHER HALF Q60-62]
[**WAVE 2 RAND10.** A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q57-Q59 IN WAVE 1 SHOULD NOW GET QUESTIONS Q60-62; A RANDOM HALF OF THE RESPONDENTS WHO GOT QUESTIONS Q60-62 IN WAVE 1 SHOULD NOW GET QUESTIONS Q57-Q59; THE OTHERS SHOULD GET THE SAME QUESTIONS AS IN WAVE 1]

Context 1: Rates
57. In the last 12 months, about how often did you go to a theater to see a movie? Would you say…
1 at least once a week,
2 a few times a month,
3 about once a month,
4 a few times a year,
5 once or twice a year, or
6 never?
-8 DON'T KNOW
-7 REFUSED

58. In the last 12 months, about how often did you eat in a restaurant, not including take-out? Would you say…
1 at least once a week,
2 a few times a month,
3 about once a month,
4 a few times a year,
5 once or twice a year, or
6 never?
-8 DON'T KNOW
-7 REFUSED

59. In the last 12 months, about how often did you exercise, including walking for fitness, gardening, or running? Would you say…
1 at least once a week,
2 a few times a month,
3 about once a month,
4 a few times a year,
5 once or twice a year, or
6 never?
-8 DON'T KNOW
-7 REFUSED
[SKIP TO Q63]

Context 2: Counts

60. In the last 30 days, how many times did you go to a theater to see a movie?

NUMBER OF TIMES: _______ -8 DON'T KNOW -7 REFUSED
[HARD RANGE 0-30]

61. In the last 30 days, how many times did you eat in a restaurant, not including take-out?

NUMBER OF TIMES: _______ -8 DON'T KNOW -7 REFUSED
[HARD RANGE 0-90, SOFT RANGE 0-30]

62. In the last 30 days, how many times did you exercise, including walking for fitness, gardening, or running?

NUMBER OF TIMES: _______ -8 DON'T KNOW -7 REFUSED
[HARD RANGE 0-90, SOFT RANGE 0-30]

Target question

NOTE TO THE INTERVIEWER: WRITE DOWN VERBATIM EVERYTHING (IF ANYTHING) THE RESPONDENT SAYS WHILE FORMULATING THE ANSWER TO THE FOLLOWING QUESTION ON YOUR HARDCOPY NOTES FORM.

63. Since January 2006, how many times have you seen a doctor, a dentist, or other health care professional about your own health at a doctor's office, a clinic, or some other place?

NUMBER OF TIMES: _______ -8 DON'T KNOW -7 REFUSED
[HARD RANGE 0-240,
SOFT RANGE 0-32]

Q63A
[DID YOU WRITE DOWN ANYTHING R SAID IN FORMULATING A RESPONSE?]

1. YES
2. NO (SKIP TO RAND11)

[IF R SAID SOMETHING BUT REFUSED TO ANSWER (63=REFUSED) SKIP TO RAND12; IF R REPORTED DON'T KNOW (63= DON'T KNOW) SKIP TO RAND11]

[**RAND11.** ONE FOURTH OF THE RESPONDENTS SHOULD GET QUESTIONS Q64A/B, AND THREE FOURTHS Q65; IF DK RESPONSE TO Q63 FOLLOW SPECIAL PATH HERE (1/4TH GO TO 64A/B THEN SKIP TO NEXT SECTION; 3/4TH SKIP TO NEXT SECTION)]

[**WAVE 2 RAND11.** ASSIGN THE RESPONDENTS TO THE SAME GROUPS AS IN WAVE 1]

64a. IF R SAID SOMETHING WHILE FORMULATING THE ANSWER TO Q63:
You said: (READ FROM HARDCOPY). As you decided how to answer the last question, did you have any additional thoughts or feelings?

1. YES
2. NO (SKIP TO 66)

(RECORD VERBATIM)

____

____

[F64TXT1 – F64TXT6 USED TO STORE TEXT FROM EITHER F64A OR F64B]

[SKIP TO Q66]

64b. IF R SAID NOTHING WHILE FORMULATING THE ANSWER TO Q63:
As you decided how to answer the last question, what thoughts or feelings went through your mind?
(RECORD VERBATIM AND READ BACK TO THE RESPONDENT)
[IF NONE, ENTER ‘NONE’.]

____

____

[SKIP TO Q66]

Q65. How did you arrive at your answer? Did you …
[RANDOMIZE ORDER OF RESPONSE OPTIONS]
1 recall each visit and count them,
2 estimate from how often you usually see a doctor, or
3 just guess?
-8 DON’T KNOW
-7 REFUSED

Q66. How certain are you that you have seen a doctor, dentist or other health care professional [INSERT ANSWER TO 63 OR “zero” IF 63=0] times since January 2006? Would you say you are…
1 very certain,
2 somewhat certain,
3 somewhat uncertain, or
4 very uncertain?
-8 DON’T KNOW
-7 REFUSED

[IF WAVE 1, GO TO RAND12
IF WAVE 2, GO TO WAVE2 RAND12]

[**RAND12.** 1/6 OF THE RESPONDENTS SHOULD GET QUESTIONS Q67a1 & Q67a2; 1/6 Q67b1 & Q67b2; 1/6 Q67c1 & Q67c2; 1/6 Q67d1 & Q67d2; 1/6 Q67e1 & Q67e2; AND 1/6 SHOULD JUST SKIP TO Q68.]

Now for a couple of related questions.

Q67a1. How likely is it that you will eat fatty foods in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON'T KNOW
-7 REFUSED

Q67a2. How likely is it that you will eat sweets in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON'T KNOW
-7 REFUSED

[GO TO Q68]

Q67b1. How likely is it that you will <u>not</u> eat fatty foods in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON'T KNOW
-7 REFUSED

Q67b2. How likely is it that you will <u>not</u> eat sweets in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON'T KNOW
-7 REFUSED

[GO TO Q68]

Q67c1. How likely is it that you will avoid eating fatty foods in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON'T KNOW
-7 REFUSED

Q67c2. How likely is it that you will avoid eating sweets in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON'T KNOW
-7 REFUSED

[GO TO Q68]

Q67d1. How likely is it that you will eat fresh fruit, such as apples, strawberries, watermelon, or bananas, in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON'T KNOW
-7 REFUSED

Q67d2. How likely is it that you will eat fresh vegetables, such as lettuce, tomatoes, peppers, or spinach, in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON'T KNOW
-7 REFUSED

[GO TO Q68]

Q67e1. How likely is it that you will eat fresh fruit in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON’T KNOW
-7 REFUSED

Q67e2. How likely is it that you will eat fresh vegetables in the next couple of weeks? Would you say…
1 very likely,
2 somewhat likely,
3 not very likely, or
4 not likely at all?
-8 DON’T KNOW
-7 REFUSED

[**WAVE 2 RAND12.** THE RESPONDENTS WHO RECEIVED QUESTIONS Q67a-c IN WAVE 1 SHOULD BE ASKED Q67.1&2; THE RESPONDENTS WHO RECEIVED QUESTIONS Q67d-e IN WAVE 1 SHOULD BE ASKED Q67.3&4; THE RESPONDENTS WHO DID NOT RECEIVE ANY QUESTIONS IN WAVE 1 SHOULD BE ASKED ALL FOUR QUESTIONS IN RANDOMIZED ORDER.]

Now for a couple of related questions.

Q67.1. Since we last spoke with you on {DATE OF INTERVIEW}, on how many days did you eat fatty foods?

Q67.2. Since we last spoke with you on {DATE OF INTERVIEW}, on how many days did you eat sweets?

Q67.3. Since we last spoke with you on {DATE OF INTERVIEW}, on how many days did you eat fresh fruit?

Q67.4. Since we last spoke with you on {DATE OF INTERVIEW}, on how many days did you eat fresh vegetables?

[RESPONSE OPTIONS FOR ALL QUESTIONS Q67.1-67.4]
NUMBER OF DAYS: _______ -8 DON’T KNOW -9 REFUSED

W2THANK
Thank you so much for your time – we really appreciate your help.
[END SURVEY WAVE 2]
<FENCEPOST – END SECTION F>

PART G: DEMOGRAPHICS

Finally, I have a few questions about your background.

68. In what year were you born?

YEAR: __________   -8 DON'T KNOW   -7 REFUSED
[HARD RANGE 1891-1988]

69. What is the highest level of education that you have completed?
1 NONE, OR GRADE 1-8
2 HIGH SCHOOL INCOMPLETE (GRADES 9-11)
3 HIGH SCHOOL GRADUATE (GRADE 12)
4 GED
5 BUSINESS, TECHNICAL, OR VOCATIONAL SCHOOL OTHER THAN HIGH SCHOOL
6 SOME COLLEGE, NO 4-YEAR DEGREE
7 COLLEGE GRADUATE, (B.S., B.A., OTHER 4-YR. DEGREE)
8 MASTER'S DEGREE, PH.D.; LAW MEDICAL OR OTHER PROFESSIONAL DEGREE
-8 DON'T KNOW
-7 REFUSED

70. Generally speaking, do you usually think of yourself as…
1 a Republican, ⇒ GO TO 70a
2 a Democrat, ⇒ GO TO 70b
3 an Independent, or ⇒ GO TO 70c
4 something else? ⇒ GO TO 70c
-8 DON'T KNOW ⇒ GO TO 70c
-7 REFUSED (SKIP TO Q71)

70a. Would you call yourself…
1 a strong Republican, or
2 a not very strong Republican?
-8 DON'T KNOW
-7 REFUSED
[SKIP TO Q71]

70b.Would you call yourself a...
1 strong Democrat, or
2 a not very strong Democrat?
-8 DON'T KNOW
-7 REFUSED
[SKIP TO Q71]

70c. Do you think of yourself as closer to…
1 the Republican party or
2 the Democratic party?
3 NEITHER (IF VOLUNTEERED)
-8 DON’T KNOW
-7 REFUSED

71. When it comes to politics, do you usually think of yourself as…
1 liberal, ⇒ GO TO 71a
2 middle of the road,
3 conservative, or ⇒ GO TO 71b
4 haven't you thought much about this?
-8 DON’T KNOW
-7 REFUSED

[RECODE VALUES AFTER ENTRY AS FOLLOWS:
1 LIBERAL = 2
2 MIDDLE OF THE ROAD = 4
3 CONSERVATIVE = 6
4 HAVEN'T YOU THOUGHT MUCH ABOUT THIS = 0]

71a. Would you say you are…
1 extremely liberal,
2 liberal, or
3 slightly liberal?
-8 DON’T KNOW
-7 REFUSED
[NO RECODE OF VALUES NEEDED FOR THIS VARIABLE, SKIP TO Q72]

71b. Would you say you are...
1 extremely conservative,
2 conservative, or
3 slightly conservative?
-8 DON’T KNOW
-7 REFUSED

[RECODE VALUES AFTER ENTRY AS FOLLOWS:
1 EXTREMELY CONSERVATIVE = 7
2 CONSERVATIVE = 6
3 SLIGHTLY CONSERVATIVE = 5]

<FENCEPOST – END SECTION G1>

72. (ASK ONLY IF NOT OBVIOUS) Are you male or female?
1. MALE
2. FEMALE
-8 DON’T KNOW
-7 REFUSED

73. Are you Spanish, Hispanic, or Latino?
1 YES – SPANISH, HISPANIC, OR LATINO (FOR EXAMPLE: CHICANO, CUBAN, MEXICAN, MEXICAN-AMERICAN, PUERTO RICAN, ETC.)
2 NO – NONE OF THESE CATEGORIES APPLY
-8 DON'T KNOW
-7 REFUSED

74. What is your race? Would you say you are…
[CODE ALL THAT APPLY – READ EXAMPLES IF NECESSARY]
1 White,
2 Black or African-American,
3 Asian, [includes: Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese]
4 Pacific Islander, [includes: Native Hawaiian, Guamanian, Samoan]
5 American Indian or Alaska Native, or
91 Some other race? (SPECIFY)___________________________________
-8 DON'T KNOW
-7 REFUSED

75. Are you now…
1 married,
2 living with a partner,
3 widowed,
4 divorced,
5 separated, or
6 never married?
-8 DON'T KNOW
-7 REFUSED

76. Including yourself, how many people live in your home?

NUMBER: ______
-8 DON'T KNOW
-7 REFUSED
[HARD RANGE 1-25]

77. (IF MORE THAN ONE PERSON IN THE HOUSEHOLD): How many of these people are age 18 and under?

NUMBER: ______
-8 DON'T KNOW
-7 REFUSED
[HARD RANGE 0-15
[EDIT – IF Q77 > Q76, "The number of people age 18 and under cannot be greater than the number of people living in the home."]

78. (IF THERE ARE 2+ PEOPLE 18 AND UNDER): How many of them are currently in school?
[IF NEEDED: Please include students who are currently on summer break.]
NUMBER: ______
-8 DON'T KNOW
-7 REFUSED
[HARD RANGE 0-15]
[EDIT – IF Q78 > Q77, "The number of people age 18 and under who are currently in school cannot be greater than the number of people age 18 and under living in the home."]

78a. (IF THERE IS 1 PERSON 18 AND UNDER): Is this person currently in school?
[IF NEEDED: Please include students who are currently on summer break.]

1 YES
2 NO
-8 DON'T' KNOW
-7 REFUSED
[IF YES(1) CODE 1 FOR Q78; IF NO (2) CODE 0 FOR Q78; ELSE CODE -7/-8]

<FENCEPOST – END SECTION G2>

[AFTER ANSWERING PHONE QUESTIONS, CODE COMPLETE]

79. We really appreciate the help you've given us today. We are interested in how people's views about the issues we discussed today change over time so it is important we talk to you again in a couple of weeks.

[**RAND13.** READ SENTENCE A. TO A RANDOM HALF OF THE RESPONDENTS, AND SENTENCE B. TO ANOTHER HALF]

A. Unfortunately, the information you've already provided to us will be much less valuable unless you complete the second interview.
B. The information you've already provided to us will be a lot more valuable if you complete the second interview.

1. AGREED TO CALLBACK APPOINTMENT
2. REFUSES CALLBACK APPOINTMENT (GO TO Q80)

[IF INTERVIEW IS BEING RECORDED, STOP RECORDING HERE AND STATE: "I have turned off the recording for these next questions."]

79a. When is a good time for us to call back and speak to you in about 2 weeks?

80. Thank you so much for your time. {IF AGREED TO CALLBACK: We look forward to talking to you again soon.}

<FENCEPOST – END SECTION G3>

APPENDIX C: RECORD CHECK QUESTIONS

Question 1. Here are some questions about registration and voting in [INSERT CITY WHERE RESPONDENT LIVES]. Have you been registered to vote in [CITY WHERE RESPONDENT LIVES] at any time since 2004?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 2. {IF YES OR DON'T KNOW TO QUESTION 1} Have you voted in any election in [INSERT CITY WHERE RESPONDENT LIVES] since 2004, either in person or by mailing an absentee ballot back to [CITY WHERE RESPONDENT LIVES] at any time since 2004.
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

{UNLESS NO TO QUESTION 1 OR 2} We know a lot of people aren't able to vote in every election. Do you know for certain whether or not you voted in any of these elections? First … (ELECTIONS, READ OFF ONE AT A TIME)

[RESPONSE OPTIONS FOR QUESTIONS 3a-3e]
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 3a. November 2004 Presidential election.

Question 3b. September 2004 Primary election.

Question 3c. November 2003 city charter election.

Question 3d. May 2003 Mayoralty election.

Question 3e. November 2006 Congressional election.

Question 4. Did you yourself happen to contribute or pledge any money to the United Way during its campaign last fall?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 5. Do you have a library card for the [INSERT CITY WHERE RESPONDENT LIVES] public library in your own name?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 6. Do you have your own [INSERT CITY WHERE RESPONDENT LIVES] Public Library card?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 7. Are you now a registered voter in the precinct where you live?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 8. Did you vote in the last primary election-the one that took place last (INSERT MONTH AND YEAR)?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 9. The next question is about the elections in November. In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. How about you - did you vote in the elections this November?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

The next few questions are about the judicial court system.

Question 10. *The judicial court system includes city, county, and federal courts. In general, do you feel that the courts are run efficiently?*
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 11. Do you think that the courts treat all citizens equally, or do they give some people better treatment than others?

1 TREAT ALL CITIZENS EQUALLY
2 GIVE SOME PEOPLE BETTER TREATMENT THAN OTHERS
8 DON'T KNOW
9 REFUSED

Have you ever been involved in a case in any of the following courts? ASK AND CODE FOR EACH.

[RESPONSE OPTIONS FOR QUESTIONS 12a-12d]
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 12a. Bankruptcy Court?

Question 12b. *Probate Court?*

Question 12c. *Divorce Court?*

Question 12d. *Small Claims Court?*

Question 13. The next question is about the elections in November. In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. We also sometimes find that people who thought that they had voted actually did not vote. Also, people who usually vote may have trouble saying for sure whether they voted in a particular election. In a moment, I'm going to ask you whether you voted on Tuesday, November 5th, which was ____ [time fill] ago. Before you answer, think of a number of different things that will likely come to mind if you actually did vote this past election day; things like whether you walked, drove, or were driven by another person to your polling place [pause], what the weather was like on the way [pause], the time of day that was [pause], and people you went with, saw, or met while there [pause]. After thinking about it, you may realize that you did not vote in this particular election. [pause]. Now that you've thought about it, which of these statements best describes you? [INTERVIEWER:READ STATEMENTS IN BOXES 1-4 to R]

1. I did not vote in the November 5th election.
2. I thought about voting this time but didn't.
3. I usually vote but didn't this time.
4. I am sure I voted in the November 5th election.
7. (VOLUNTEERED) I VOTED BY ABSENTEE BALLOT.

Question 14. During the time you were an undergraduate at the [INSERT REPONDENT'S UNIVERSITY], did you ever drop a class and receive a grade of "W"?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 15. Did you ever receive a grade of 'D' or 'F' for a class?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 16. Were you ever placed on academic warning or academic probation?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 17. What was your cumulative overall undergraduate grade point average or GPA at the time you received your undergraduate degree?

______ GPA
-8 DON'T KNOW
-7 REFUSED

Question 18. Did you graduate with cum laude, magna cum laude, or summa cum laude?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 19. Are you a dues-paying member of the [INSERT REPONDENT'S UNIVERSITY] Alumni Association?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 20. Since you graduated, have you ever donated financially to the [INSERT REPONDENT'S UNIVERSITY]?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 21. Did you make a donation to the [INSERT REPONDENT'S UNIVERSITY] in calendar year 2004?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 22. *Do you currently have health insurance coverage?*
1 YES
2 NO [SKIP QUESTIONS 23a-23c]
-8 DON'T KNOW
-7 REFUSED

During the past 2 months, since (date), have you had any of the following procedures done under your current insurance coverage?
[RESPONSE OPTIONS FOR QUESTIONS 23(4)a-23(4)c]
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 23a. Blood pressure reading?

Question 23b. Test of blood in your stool?

Question 23c. Had a new prescription filled at a pharmacy?

During the past 6 months, since (date), have you had any of the following procedures done under your current insurance coverage?
[RESPONSE OPTIONS FOR QUESTIONS 24a-24c]
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 24a. Blood pressure reading?

Question 24b. Test of blood in your stool?

Question 24c. Had a new prescription filled at a pharmacy?

Question 25. *Do you have any children under the age of 18 living in your household?*
1 YES
2 NO [SKIP QUESTION 26]
8 DON'T KNOW
9 REFUSED

Question 26. Is your child covered by Medicaid, a health insurance program for low income families?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 27. *When was your last visit to see doctor, a medical doctor or assistant at a doctor's office, a clinic, or some other place?*

RECORD DATE __ / __/ ____

Question 28. What was the reason for this visit? (Can you tell me more about that?)

(RECORD VERBATIM AND READ BACK TO THE RESPONDENT)

Question 29. I'm going to ask you a series of questions about different procedures you may have had done during your last visit to a medical doctor or assistant. This includes x-rays, lab tests, surgical procedures, and prescriptions. For each of these areas, I'll ask you whether or not it happened, and whether you paid any of your own money to cover the costs. First, during your last visit to a medical doctor or assistant, did you have an x-ray, CAT scan, MRI, or NMR?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 30. During your last visit to a medical doctor or assistant, did you have any lab tests done that required blood, urine, or other body fluids?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 31. During your last visit to a medical doctor or assistant, did you have any surgical procedures?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 32. I've asked you a number of questions about x-rays, lab tests, or surgical procedures that you have had done at your last visit. This is an important area for our research. Can you think of any other tests or procedures you had done at your last visit to a medical doctor or assistant that you have not already had a chance to tell me about?
1 YES
2 NO
-8 DON'T KNOW
-7 REFUSED

Question 33. *The next few questions are about any benefits provided you through an employer. Are you currently employed?*
1 YES
2 NO {SKIP QUESTIONS 33-45}
8 DON'T KNOW
9 REFUSED

Question 34. Is your current job covered by a Union Contract?
1 YES
2 NO {SKIP QUESTION 35}
8 DON'T KNOW
9 REFUSED

Question 35. Do you belong to that union?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 36. Do you have medical, surgical, or hospital insurance that covers any illness or injury that might happen to you when you are not at work?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 37. Do you receive sick days with full pay?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 38. Are dental benefits provided to you on your main job?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 39. Do you have life insurance that would cover a death occurring for reasons not connected with your job?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 40. Do you get paid vacation days?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 41. Do you have (maternity/paternity) leave that will allow you to go back to your old job or one that pays the same as your old job?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 42. How about (maternity/paternity) leave with pay. Is that available to you on your main job?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 43. Now I need to get some information about any pension or retirement plan you may be eligible for at your place of work. Not including Social Security or Railroad Retirement, are you covered by a pension or retirement plan on your present job?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 44. Have you worked under the main or basic plan long enough to earn the right of vesting?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 45. If you wished to retire earlier (than time needed to receive full benefits), could you receive part but not full benefits from this plan?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 46. Do you or your family rent, or own, the place where you live?
1 RENT
2 OWN
8 DON'T KNOW
9 REFUSED

Question 47. Is there a telephone in your home in your family's name?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 48. Do you have a [INSERT STATE WHERE RESPONDENT LIVES] drivers license that is still good?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 49. Do you happen to own an automobile at the present time?
1 YES
2 NO [SKIP QUESTIONS 50-52]
8 DON'T KNOW [SKIP QUESTIONS 50-52]
9 REFUSED [SKIP QUESTIONS 50-52]

Question 50. (IF YES TO QUESTION 49) Is it registered in your name alone, or in your (wife's) (husband's) name also?
1 OWN NAME
2 WIFE/HUSBAND
8 DON'T KNOW
9 REFUSED

Question 51. (IF YES TO QUESTION 49) Does the car have [INSERT STATE WHERE RESPONDENT LIVES] plates or plates from some other state?
1 STATE WHERE RESPONDENT CURRENTLY LIVES
2 SOME OTHER STATE
8 DON'T KNOW
9 REFUSED

Question 52. (IF YES TO QUESTION 49) What year and make of car is it?

YEAR _________ MAKE __________ 8 DON'T KNOW 9 REFUSED

Question 53. Have you received a ticket for parking in the past 12 months?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 54. Have you received a ticket for going through a red light in the past 12 months?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 55. During the last 12 months, have you been charged by a policeman for speeding?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 56. During the last 12 months, have you been charged by a policeman for driving under the influence of liquor?
1 YES
2 NO
8 DON'T KNOW
9 REFUSED

Question 57. In what year were you born?

YEAR: __________ -8 DON'T KNOW -7 REFUSED
[HARD RANGE 1891-1988]

Question 58. May I ask your age?

AGE: __________ -8 DON'T KNOW -7 REFUSED
[HARD RANGE 18-99]

Sources for questions from record check studies

Belli, Robert F., Michael W. Traugott, Margaret Young, and Katherine A. McGonagle. 1999. "Reducing Vote Over-Reporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring." *Public Opinion Quarterly,* 63:90-108.

Blumberg SJ, Cynamon ML. 1999. Misreporting Medicaid enrollment: Results of three studies linking telephone surveys to state administrative records. In: Cynamon ML, Kulka RA, editors. Proceedings of the Seventh Conference on Health Survey Research Methods, Sep 24-27; Williamsburg, VA. Hyattsville (MD): Department of Health and Human Services; 2001; Publication No. (PHS) 01-1013; p. 189-95.

Duncan, Greg J and Daniel H. Hill. 1985. "An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data." *Journal of Labor Economics,* 3: 508-532.

Dykema, Jennifer, James Lepkowski, and Steven Blixt. 1997. "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study." Pp. 287-310 in *Survey Measurement and Process Quality,* edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, N. Schwarz, and D. Trewin. New York: John Wiley & Sons Inc.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity." *Public Opinion Quarterly,* 72: 847–865

Locander, William, Seymour Sudman, and Norman Bradburn. 1976. "An Investigation of Interview Method, Threat and Response Distortion." *Journal of the American Statistical Association*, 71:269-275.

Loftus, Elizabeth F., Kyle D. Smith, Mark R. Klinger, and Judith Fiedler. 1992. "Memory and Mismemory for Health Events." Pp. 102-137 in *Questions about Questions: Inquiries Into the Cognitive Bases of Surveys*, edited by Judith M. Tanur. Thousand Oaks, CA: Sage.

Parry, Hugh J. and Helen M. Crossley. 1950. "Validity of Responses to Survey Questions." *Public Opinion Quarterly,* 14: 61-80.

REFERENCES

AAPOR. 2006. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 4th edition. Lenexa, Kansas: American Association for Public Opinion Research.

Aday, Lu Ann. 1989. *Designing and Conducting Health Surveys*. San Francisco: Jossey-Bass.

Alwin, Duane F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, NJ: John Wiley and Sons, Inc.

Andrews, Frank M. 1994. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." *Public Opinion Quarterly* 48(2):409-442.

Anderson, D.R. (2008). *Model Based Inference in the Life Sciences: A Primer on Evidence.* New York: Springer.

Bassili, John N. and B. Stacey Scott. 1996. "Response Latency as a Signal to Question Problems in Survey Research." *Public Opinion Quarterly* 60(3):390-99.

Basili, John N. 2000. "Editor's Introduction: Reflections on Response Latency Measurement in Telephone Surveys." *Political Psychology* 21: 1-6.

Beatty, Paul and Gordon Willis. 2007. "Research Synthesis: The Practice of Cognitive Interviewing." *Public Opinion Quarterly* 71(2):287-311.

Belson, William A. 1981. *The Design and Understanding of Survey Questions*. London: Gower Publishing.

Bickart, Barbara and Marla E. Felcher. 1996. "Expanding and Enhancing the Use of Verbal Protocols in Survey Research." Pp. 115-142 in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman. San Francisco: Jossey-Bass.

Bischoping, Katherine. 1989. "An Evaluation of Interviewer Debriefing in Survey Pretests." In *New Techniques for Pretesting Survey questions*, edited by C. Cannell. L. Oksenberg. F. Fowler. G. Kalton, and K. Bischoping, ch. 2. Ann Arbor, MI: Survey Research Center.

Blair, Johnny and Ronald Czaja. 2005. *Designing Surveys: A Guide to Decisions and Procedures*. 2nd edition. Thousand Oaks, CA: Pine Forge Press.

Blair, Johnny and Fred Conrad. 2011. “Sample Size for Cognitive Interview Pretesting.” *Public Opinion Quarterly* 75: 636-658.

Blair, Johnny and Stanley Presser. 1993. "Survey Procedures for Conducting Cognitive Interviews to Pretest Questionnaires: A Review of Theory and Practice." In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Blair, Johnny, Fred Conrad, Allison Ackerman, and G. Claxton. 2006. “Using Behavior Coding to Validate Cognitive Interview Findings.” In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Blair, Johnny, Allison Ackerman, Linda Piccinino, and Rachel Levenstein. 2007. “The Effect of Sample Size on Cognitive Interview Findings.” In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Campbell, Donald T. and Fiske, Donald W. 1959. “Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.” *Psychological Bulletin* 56(2):81-105.

Campanelli, Pamela C., Elizabeth A. Martin, and Jennifer M. Rothgeb. 1991. “Use of Respondent and Interviewer Debriefing Studies as a Way to Study Response Error in Survey Data.” *The Statistician* 40: 253-264.

Coltheart, Max. 1981. “The MRC Psycholinguistic Database.” *Quarterly Journal of Experimental Psychology* 33A:497-505.

Converse, Jean M. and Stanley Presser. 1984. *Survey Questions: Handcrafting the Standardized Questionnaire*. Thousand Oaks, CA: Sage.

Conrad, Frederick G. and Johnny Blair. 2004. "Aspects of Data Quality in Cognitive Interviews: the Case of Verbal Reports." Pp. 131-47 in *Methods for Testing and Evaluating Survey Questionnaires,* edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin and E. Singer. Hoboken, New Jersey: John Wiley and Sons.

Conrad, Fred and Johnny Blair. 2009. “Sources of Error in Cognitive Interviews.” *Public Opinion Quarterly* 73: 32-55.

Conrad, Frederick G. and Michael F. Schober. 2000. “Clarifying Question Meaning in a Household Survey.” *Public Opinion Quarterly* 64: 1-28.

Couper, Mick P. and Frauke Kreuter. 2013. “Using Paradata to Explore Item Level Response Times in Surveys.” *Journal of the Royal Statistical Society: Series A* 176: 271-286.

Demaio, Terry and Landreth, Ashley. 1993. “Examining Expert Review as a Pretest Method.” In *Proceedings of the 4th Conference on Questionnaire Evaluation Standards*. Mannheim, Germany: ZUMA.

DeMaio, Theresa J. and Ashley Landreth. 2004. "Do Different Cognitive Interviewing Techniques Produce Different Results." Pp. 89-108 in *Methods for Testing and Evaluating Survey Questionnaires,* edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin and E. Singer. Hoboken, New Jersey: John Wiley and Sons.

Draisma, Stasja and Wil Dijkstra. 2004. "Response Latency and (Para) Linguistic Expressions as Indicators of Response Error." Pp. 131-47 in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin and E. Singer. Hoboken, New Jersey: John Wiley and Sons.

Dykema, Jennifer, James Lepkowski, and Steven Blixt. 1997. “The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study.” Pp. 287-310 in *Survey Measurement and Process Quality,* edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, N. Schwarz, and D. Trewin. New York: John Wiley & Sons Inc.

Ehlen, Patrick, Michael F. Schober, and Frederick G. Conrad. 2007. “Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces.” *Discourse Processes* 44:245-265.

Eisenhower, Donna. 1994. “Design Oriented Focus Groups and Cognitive Interviews: A Comparison.” In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Ericsson and Simon. 1980. *Protocol Analysis*. Cambridge, MA: The MIT Press.

Esposito, James L. 2004. “Iterative Multiple-Method Questionnaire Evaluation Research: A Case Study.” *Journal of Official Statistics* 20: 143-183.

Esposito, James L., Pamela C. Campenell, Jennifer Rothgeb, and Anne E. Polivka. 1991. “Determining Which Questions are Best: Methodologies for Evaluating Survey Questions.” In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Esposito, James L. and Jennifer M. Rothgeb. 1997. "Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment." Pp. 541-71 in *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw and C. Dippo. New York: John Wiley and Sons.

Forsyth, Barbara H., Judith T. Lessler, and Michael L. Hubbard. 1992. "Cognitive Evaluation of the Questionnaire." Pp. 13-52 in *Survey Measurement of Drug Use*, edited by C. F. Turner, J. T. Lessler and J. C. Gfroer. Rockville, MD: U.S. Department of Health and Human Services.

Forsyth, Barbara, Jennifer M. Rothgeb, and Gordon B. Wills. 2004. "Does Pretesting Make a Difference? An Experimental Test." Pp. 525-546 in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin and E. Singer. Hoboken, New Jersey: John Wiley and Sons.

Fowler, Floyd J. 1995. *Improving Survey Questions*. Thousand Oaks, CA: Sage.

Fowler, Floyd J. and Charles F. Cannell. 1996. "Using Behavior Coding to Identify Cognitive Problems With Survey Questions." Pp. 15-36 in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman. San Francisco: Jossey-Bass.

Fowler, Floyd J. and Roman. 1992. A Study of Approaches to Survey Question Evaluation. Unpublished Manuscript. University of Massachusetts.

Graesser, Arthur C., Katja Wiemer-Hastings, Peter Wiemer-Hastings, and Roger Kreuz. 2000. "The Gold Standard of Question Quality on Surveys: Experts, Computer Tools, Versus Stastical Indices." In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Graesser, Arthur C., Zhqiang Cai, Max M. Louwerse, and Frances Daniel. 2006. "Question Understanding Aid (QUAID): A Web Facility That Tests Question Comprehensibility." *Public Opinion Quarterly* 70(1):3-22.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. Hoboken, NJ: John Wiley and Sons.

Hess, Jennifer, Eleanor Singer, and John Bushery. 1999. "Predicting Test-Retest Reliability From Behavior Coding." *International Journal of Public Opinion Research* 11:346-360.

Holbrook, Allyson, Young Ik Cho, and Timothy Johnson. 2006. “The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties.” *Public Opinion Quarterly* 70:565-595.

Hughes, Kristen Ann. 2004. “Comparing Pretesting Methods: Cognitive Interviews, Respondent Debriefing, and Behavior Coding.” In *Survey Methodology Research Report Series*. Washington D.C.: U.S. Census Bureau.

Hunt, Shelby D., Richard D. Sparkman, Jr., and James B. Wilcox. 1982. “The Pretest in Survey Research: Issues and Preliminary Findings.” *Journal of Marketing Research* 19: 269-73.

Jobe and Hermann 1996. “Implications of Models of Survey Cognition for Memory Theory.” In *Basic and Applied Memory Research*, edited by D. Herrmann, C. McEvoy, C. Herzog, P. Hertel and M. Johnson, 193-205. Hillsdale, NJ: Erlbaum.

Knauper, Barbel, Robert F. Belli, Daniel H. Hill, A. Regula Herzong. 1997. “Question Difficulty and Respondents Cognitive Ability: The Effect on Data Quality.’’ *Journal of Official Statistics* 13:181–99.

Krosnick, Jon A., 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5:213-236.

Krosnick, Jon A., and Stanley Presser. 2010. ‘‘Question and Questionnaire Design.’’ In *Handbook of Survey Research*, 2nd ed., edited by P.V. Marsden and J.D. Wright, 263–313. Bingley, UK: Emerald Group Publishing Limited.

Lessler, Judith T. and Barbara H. Forsyth. 1996. “A Coding System for Appraising Questionnaires.” Pp. 259-92 in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman. San Francisco: Jossey-Bass.

Lessler, Judith, Roger Tourangeau, and William Salter. 1989. “Questionnaire Design in the Research Laboratory.” In *Vital and Health Statistics Series 6*. Hyattsville, MD: National Center for Health Statistics.

Levenstein, Rachel, Fred Conrad, Johnny Blair, Roger Tourangeau, and Aaron Maitland. 2007. “The effect of probe type on cognitive interview results: A signal detection analysis.” In *Proceedings of the Section on Survey Methods* (pp. 3850-3855). Alexandria, VA: American Statistical Association.

Loftus, Elizabeth. 1984. "Protocol Analysis of Responses to Survey Recall Questions." Pp. 61-64 in Cognitive Aspects of Survey Methodology, edited by T. B. Jabine, M. L. Straf, J. M. Tanur and R. Tourangeau. Washington, D.C.: National Academy Press.

Long, Jeffrey D. (2012). *Longitudinal Data Analysis for the Behavioral Sciences Using R*. Los Angeles: Sage.

Lord, Fredric M. and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Madans, Jennifer and Paul Beatty. 2012. “Evaluating Survey Questions: A Comparison of Methods - Discussion.” *Journal of Official Statistics* 28: 531-35.

Miller, Kristen. 2002. “A Comparison of Focus Group and One-on-One Cognitive Interviewing for Question Evaluation.” Paper Presented at the International Conference on Questionnaire Evaluation Design and Testing. Charleston, SC.

Mulligan, Kenneth, J. Tobin Grant, Stephen T. Mocabee, and Joseph Quin Monson. 2003. “Response Latency Methodology for Survey Research.” *Political Analysis* 11: 289-301.

Oksenberg, Lois, Charles Cannell, and Graham Kalton. 1991. “New Strategies for Pretesting Survey Questions.” *Journal of Official Statistics* 7: 349-365.

Olson, Kristen. 2010. "A Note on Questionnaire Evaluation by Expert Raters." *Field Methods*. 22: 295-318.

O’Muircheartaigh, Colm. 1991. Simple Response Variance: Estimation and Determinants. Pp. 551-74 in *Measurement Errors in Surveys*, edited by P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman. New York: John Wiley and Sons.

Ongena, Yfke P. and Wil Dijkstra. 2006. “Methods of Behavior Coding of Survey Interviews. *Journal of Official Statistics* 22: 419-451.

Payne, Stanley. 1951. *The Art of Asking Questions*. Princeton, NJ: Princeton University Press.

Pickery, Jan and Geert Loosveldt. 2001. “An Exploration of Question Characteristics that Mediate Interviewer Effects on Item Nonresponse.” *Journal of Official Statistics* 17: 337-350.

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

Presser, Stanley, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. 2004. *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, New Jersey: John Wiley and Sons.

Rothgeb, Jennifer M., Gordon B. Willis, and Barbara H. Forsyth. 2001. "Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results?" In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Saris, Willem and Irmtraud Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research.* Hoboken, NJ: John Wiley and Sons Inc.

Saris, Willem. 2012. "Discussion: Evaluation Procedures for Survey Questions." *Journal of Official Statistics* 28: 537-551.

Saris, Willem, William van der Veld, and Irmtraud Gallhofer. 2004. "Development and Improvement of Questionnaires Using Predictions of Reliability and Validity." Pp 275-297. in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin and E. Singer. Hoboken, New Jersey: John Wiley and Sons.

Schaeffer, Nora C. 1991. "Conversation With a Purpose – or Conversation? Interaction in the Standardized Interview." Pp. 367-391 in *Measurement Errors in Surveys,* edited by P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman. Hoboken, N.J.: John Wiley and Sons.

Schaeffer, Nora C. and Jennifer Dykema. 2011. "Questions for Surveys Current Trends and Future Directions." *Public Opinion Quarterly* 75: 909-961.

Schaeffer, Nora C. and Douglas W. Maynard. 2002. "Occasions for Intervention: Interactional Resources for Comprehension in Standardized Interviews." Pp. 261-80 in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer and J. van der Zouwen. New York: John Wiley and Sons.

Schober, Michael F. and Frederick G. Conrad. 1997. "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly* 61: 576–602.

Schuman, Howard. 1966. "The Random Probe: A Technique for Evaluating the Validity of Closed Questions." *American Sociological Review* 21(2): 218-222

Spector, Paul E. 1992. *Summated Rating Scale Construction: An Introduction*. Thousand Oaks, CA: Sage.

Stapleton-Kudela, Marthat, Barbara H. Forsyth, Kerry Levin, Deirdre Lawrence, and Gordon Willis. 2006. "Cognitive Interviewing Versus Behavior Coding." In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Sudman, Semour and Norman Bradburn. 1974. *Response Effects in Surveys*. Chicago: Adline Publishing Company.

Sykes, Wendy and Jean Morton-Williams. 1987. “Evaluating Survey Questions.” *Journal of Official Statistics* 3: 191-207.

Tourangeau, Roger. 1984. "Cognitive Sciences and Survey Methods." Pp. 73-100 in Cognitive Aspects of Survey Methodology, edited by T. B. Jabine, M. L. Straf, J. M. Tanur and R. Tourangeau. Washington, D.C.: National Academy Press.

van der Zouwen, Johannes and Wil Dijkstra. 2002. Testing Questionnaires Using Interaction Coding. Pp. 427-47 in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer and J. van der Zouwen. New York: John Wiley and Sons.

van der Zouwen, Johannes, Willem E. Saris, Stasja Draisma, and William van der Veld. 2001. "Assessing the Quality of Questionnaires: A Comparison of Three Methods for 'Ex Ante' Evaluation of Survey Questions." Presented at the International Conference on Quality in Official Statistics, May 14-15, Stockholm, Sweden.

van der Zouwen, Johannes and Johannes H. Smit. 2004. "Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and Question-Answer Sequences: A Diagnostic Approach." Pp. 109-30 in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin and E. Singer. Hoboken, New Jersey: John Wiley and Sons.

Von Thurn, and Moore. 1994. “Results From a Cognitive Exploration of the 1993 American Housing Survey.” In Proceedings of the ASA Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Wellens. 1994. “The Cognitive Evaluation of the Nativity Questions for the Current Population Survey.” In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

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

Willis, Gordon B. 1991. "The Use of Behavior Coding to Evaluate a Draft Health-Survey Questionnaire." Paper presented at the annual meeting of the American Association for Public Opinion Research. Phoenix.

Willis, Gordon B. 2004. "Cognitive Interviewing Revisited: A Useful Technique, in Theory?" Pp. 23-43 in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin and E. Singer. Hoboken, New Jersey: John Wiley and Sons.

Willis, Gordon B. and Judith Lessler. 1999. *Questionnaire Appraisal System: QAS-99*. Rockville, MD: Research Triangle Institute.

Willis, Gordon B. and Susan Schechter. 1997. "Evaluation of Cognitive Interviewing Techniques: Do Results Generalize to the Field?" *Bulletin de Methodolgie Sociologique* 55:40-66.

Willis, Gordon B., Susan Schechter, and Karen Whitaker. 1999. "A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What Do They Tell Us?" In *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Wilson, Mark. 2005. *Constructing Measures: An Item Response Modeling Approach*. New York: Psychology Press.

Yan, Ting and Roger Tourangeau. 2008. “Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times,” *Applied Cognitive Psychology* 22: 51-68.

Yan, Ting, Frauke Kreuter, and Roger Tourangeau. 2012a. “Evaluating Survey Questions: A Comparison of Methods.” *Journal of Official Statistics* 28: 503-29.

Yan, Ting, Frauke Kreuter, and Roger Tourangeau. 2012b. “Evaluating Survey Questions: A Comparison of Methods - Rejoinder.” *Journal of Official Statistics* 28: 553-54.