Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 503–529

Evaluating Survey Questions: A Comparison of Methods

Ting Yan[1], *Frauke Kreuter*[2], *and Roger Tourangeau*[3]

This study compares five techniques to evaluate survey questions — expert reviews, cognitive interviews, quantitative measures of reliability and validity, and error rates from latent class models. It is the first such comparison that includes both quantitative and qualitative methods. We examined several sets of items, each consisting of three questions intended to measure the same underlying construct. We found low consistency across the methods in how they rank ordered the items within each set. Still, there was considerable agreement between the expert ratings and the latent class method and between the cognitive interviews and the validity estimates. Overall, the methods yield 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.

Key words: Cognitive interviews; expert reviews; latent class analysis; measurement error; question pretests; reliability; validity.

1. Introduction

Survey researchers have a variety of techniques at their disposal for evaluating survey questions (see Presser et al. 2004b). These range from cognitive interviews (e.g., Willis 2005), to the conventional pretests recommended in many questionnaire design texts (e.g., Converse and Presser 1986), to behavior coding of various types (Maynard et al. 2002; van der Zouwen and Smit 2004), to question wording experiments (e.g., Fowler 2004), to the application of statistical procedures, such as latent class analysis (e.g., Biemer 2004) or structural equation modeling (Saris and Gallhofer 2007), that provide quantitative estimates of the level of error in specific items. These different evaluation techniques do not bear a close family resemblance. Although they all share the general goal of helping question writers to evaluate survey questions, they differ in their underlying assumptions,

[1] NORC at the University of Chicago, 1155 E. 60th Street, Chicago IL 60637, U.S.A.
Email: tingyan@umich.edu
[2] Joint Program in Survey Methodology, University of Maryland, College Park, 1218Q LeFrak Hall, Maryland, U.S.A. Institute for Employment Research/LMU, Nuremberg/Munich, Germany. Email: fkreuter@umd.edu
[3] Joint Program in Survey Methodology, University of Maryland, 1218Q LeFrak Hall, Maryland, U.S.A. Survey Research Center, University of Michigan, Ann Arbor, U.S.A. Email: RogerTourangeau@Westat.com
Acknowledgments: An earlier version of this article was presented at the 2008 Meeting of the American Association for Public Opinion Research. The work reported here is supported by the National Science Foundation under Grants SES 0550002 and 0549916 to the authors. We would like to thank the Methodology, Measurement, and Statistics (MMS) Program and Dr. Cheryl Eavey for their support. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Cat Tourangeau for her able research assistance; Norman Bradburn, Ashley Landreth, Stanley Presser, and Nora Cate Schaeffer for their expert reviews of the questions; and Willem Saris for providing SQP predictions for item quality. In addition, we thank Paul Beatty and Willem Saris for valuable comments on earlier versions of the manuscript.

the data collection methods they use, the types of problems they identify, the practical requirements for carrying them out, the type of results they generate, and so on.

To illustrate the differences across techniques, consider cognitive interviewing and the use of latent class modeling methods for evaluating survey items. Cognitive interviewing is a practice derived from the protocol analyses used in the work of Simon and his collaborators. Loftus (1984) first pointed out the potential relevance of Simon's work to the testing of survey items more than 25 years ago. The key assumptions of protocol analysis and its latter-day survey descendant, cognitive interviewing, are that the cognitive processes involved in answering survey questions leave traces in working memory (often intermediate products of the process of formulating an answer) and that respondents can verbalize these traces with minimal distortion (see Ericsson and Simon 1980). Cognitive interviews added several techniques to the think-aloud methods introduced by Simon and his colleagues, especially the use of specially designed probes, or follow-up questions; responses to these probes are also thought to provide important clues about how respondents come up with their answers and about potential problems with those processes. Some researchers see later developments of cognitive interviews as a departure from the original paradigm proposed by Ericsson and Simon, and argue that the use of probes has largely supplanted the use of think-aloud methods in cognitive interviewing (see, for example, Schaeffer and Presser 2003, p. 82; see also Beatty and Willis 2007; Gerber 1999). Cognitive interviews are rarely subjected to formal analyses; instead, the questionnaire testing personnel, often staff with advanced degrees, draw conclusions about the questions from their impressions of the verbal reports produced by respondents during the cognitive interviews (see Willis 2005 for a thorough discussion of cognitive interviewing).

At the other end of the continuum stands the application of quantitative methods, such as latent class modeling, to assess problems in survey questions. In a series of papers, Biemer and his colleagues (Biemer 2004; Biemer and Wiesen 2002; Biemer and Witt 1996) have used latent class models to estimate error rates in survey items designed to assess such categorical constructs as whether a person is employed or not. Latent class analysis is sometimes described as the categorical analogue to factor analysis (e.g., McCutcheon 1987, p. 7). It is used to model the relationships among a set of observed categorical variables that are indicators of two or more latent categories (e.g., whether one is truly employed or unemployed). In contrast to cognitive interviews, latent class analysis is a statistical technique that yields quantitative estimates. It uses maximum likelihood methods to estimate parameters that represent the prevalence of the latent classes and the probabilities of the different observed responses to the items conditional on membership in one of the latent classes.

Given the large number of different evaluation methods and the large differences between them, it is an important theoretical question whether the different methods *should* yield converging conclusions and, if not, whether they should be used alone or in combination with each other. In practice, the choice between techniques is often dictated by considerations of cost and schedule, and it is an important practical question whether clear conclusions will result even if different methods are adopted.

The answers to the questions of whether converging conclusions should be expected and how to cope with diverging conclusions about specific items depend in part on how researchers conceive of the purpose of the different evaluation methods. Much work on

question evaluation and pretesting tends to treat question problems as a binary characteristic – the question either has a problem or it does not. Question evaluation methods are used to identify the problems with an item and group questions into two categories – those with problems that require the item to be revised, and those without such problems. Under this conceptualization, all of the question evaluation methods flag some items as problematic and others as non-problematic, even though the methods may differ in which items they place in each category. Of course, questions may have problems that differ in seriousness, but ultimately questions are grouped into those that require revision and those that do not. Different question evaluation methods are, then, compared on their success in identifying question problems and correctly placing items into one of the two categories. Presser and Blair's (1994) study is a classic example of such a conceptualization. Implicit in such work is the assumption that if *any* method reveals a problem with an item, that problem should be addressed. That assumption has been challenged recently by Conrad and Blair (2004; see also Conrad and Blair 2009), who argue that the "problems" found in cognitive interviews may well be false alarms.

Recently, the field of question evaluation and pretesting has seen a shift towards a more general conceptualization of question problems and goals of question evaluation methods (e.g., Miller 2009). Survey questions measure the construct they are supposed to measure more or less well (Saris and Gallhofer 2007). Thus, it is possible to conceive of question problems as a matter of degree, and the purpose of question evaluation methods is to determine the degree of fit between question and construct (Miller 2009).

There is limited empirical work comparing different question evaluation methods, especially work comparing qualitative methods (like cognitive interviews and expert reviews) with quantitative methods (like measurements of reliability and validity). The few prior studies that have been done seem to suggest that the consistency between the different methods is not very high, even at the level of classifying items as having problems or not (see Presser and Blair 1994; Rothgeb et al. 2001; and Willis et al. 1999, for examples). Table 1 provides a summary of the major studies comparing question evaluation techniques.

It is apparent from Table 1 that large disagreements across methods exist about which items have problems or which problems they have. There are several possible reasons for discrepant results across evaluation methods. The different methods may identify different types of problems. For instance, Presser and Blair (1994) found that interviewer debriefings are likely to pick up problems with administering the questions in the field, whereas cognitive interviews are likely to detect comprehension problems. The two methods may yield complementary sets of real problems. In addition, the methods may not be all that reliable. Partly, this unreliability may reflect differences in what the researchers count as problems and in how they conduct different types of evaluations. Several studies have examined whether multiple implementations of the "same" method yield similar conclusions about a set of items; the results suggest that unreliability within a method is often high (e.g., DeMaio and Landreth 2004; Presser and Blair 1994; Willis et al. 1999). Finally, another reason for disagreement across methods is that some of the evaluation methods may not yield valid results (cf. Presser et al. 2004a; on the potential for invalid conclusions, see Conrad and Blair 2004; 2009).

Table 1. Studies comparing question evaluation methods

Paper	Methods tested	Criteria	Conclusions
Fowler and Roman (1992)	1. Focus groups 2. Cognitive interviews 3. Conventional pretest 4. Interviewer ratings of items 5. Behavior coding	• Number of problems found • Type of problem found	1. Focus groups and cognitive interviews provide complementary information 2. Results from two sets of cognitive interviews (done by separate organizations) are similar 3. Interviewer debriefing identifies more problems than interviewer ratings and ratings identify more problems than behavior coding 4. All five methods provide useful information
Presser and Blair (1994)	1. Conventional pretests 2. Behavior coding 3. Cognitive interviews 4. Expert panels	• Number of problems found • Type of problem found • Consistency across trials with the same method	1. Conventional pretests and behavior coding found the most interviewer problems 2. Expert panels and cognitive interviews found the most analysis problems 3. Expert panels and behavior coding were more consistent across trials and found more types of problems 4. Behavior coding was most reliable but provided no information about the cause of a problem, did not find analysis problems, and did not distinguish between respondent-semantic and respondent-task problems 5. Expert panels were most cost-effective 6. Most common problems were respondent-semantic
Willis, Schechter, and Whitaker (1999)	1. Cognitive interviewing (done by interviewers at two organizations) 2. Expert review 3. Behavior coding	• Number of problems found • Consistency within and across methods regarding the presence of a problem (measured by the correlation across methods and organizations between the percent of the time items were classified as having a problem)	1. Expert review found the most problems 2. The correlation between behavior coding trials was highest (.79), followed closely by the correlation between the cognitive interviews done by two organizations (.68)

Table 1. Continued

Paper	Methods tested	Criteria	Conclusions
		• Type of problems	3. Across methods of pretesting and organizations, most problems were coded as comprehension/ communication; there was a high rate of agreement in the use of sub-codes within this category across techniques
Rothgeb, Willis and Forsyth (2001)	Three organizations each used three methods to test three questionnaires 1. Informal expert review 2. Formal cognitive appraisal 3. Cognitive interviewing	• Number of problems found • Agreement across methods based on summary score for each item (summary scores ranged from 0 to 9 based on whether the item was flagged as a problem item by each technique and each organization)	1. Formal cognitive appraisal (QAS) found most problems but encouraged a low threshold for problem identification 2. Informal expert review and cognitive interviewing found similar numbers of problems, but found different items problematic 3. Results across organizations were more similar than across techniques: Moderate agreement across organizations in summary scores (r's range from .34 to .38) 4. Communication and comprehension problems were identified most often by all three techniques
Forsyth, Rothgeb and Willis (2004)	1. Informal expert review	• Conducted randomized experiment in a RDD survey that compared the original items pretested in 2001 study with revised items designed to fix problems found in the pretest	1. Items classified as high in interviewer problems during pretesting also had many problems in the field (according to behavior coding and interviewer ratings)
(Note: This study is a follow-up to Rothgeb et al. 2001)	2. Formal cognitive appraisal (QAS) 3. Cognitive interviewing	• Classified items as low, moderate, or high in respondent and interviewer problems, based on behavior coding data and interviewer ratings	2. Items classified as high in respondent problems during pretesting also has many problems in the field. 3. Items classified as having recall and sensitivity problems during pretesting had higher nonresponse rates in the field.

Table 1. Continued

Paper	Methods tested	Criteria	Conclusions
			4. The revised items in the experimental questionnaire produced nonsignificant reductions in item nonresponse and problems found via behavior coding, but a significant reduction in respondent problems (as rated by the interviewers); however, interviewers rated revised items as having more interviewer problems.
DeMaio and Landreth (2004)	1. Three cognitive interview methods (three different "packages" of procedures carried out by three teams of researchers at three different organizations) 2. Expert review	• Number of problems identified • Type of problem identified • Technique that identified the problem • Frequency of agreement between organizations/methods	1. The different methods of cognitive interviewing identified different numbers and types of problems 2. Cognitive interviewing teams found fewer problem questions than expert reviews, but all three organizations found problems with most questions for which two or more experts agreed there was a specific problem 3. The problems identified by the cognitive interviewing teams were also generally found by the experts 4. Different teams used different types of probes 5. Cognitive interviews done on revised questionnaires found that only one team's questionnaire had fewer problems than the original
Jansen and Hak (2005)	1. Three-Step Test Interview (cognitive interviews with concurrent think-alouds followed by probes and respondent debriefing) 2. Expert review	• Number of problems found • Places in questionnaire where problems were found • Type of problem found	1. Three-step test interview identified more problems than expert reviews 2. Three-step test-interview identified unexpected problems stemming from non-standard drinking patterns and from local norms regarding drinking alcohol

A limitation on the comparison studies summarized in Table 1 is that it is rarely clearly evident whether the problems identified by a given technique actually reduce the validity or accuracy of the answers in surveys. As Groves and his colleagues note: "[The assumption is that] questions that are easily understood and that produce few other cognitive problems for the respondents introduce less measurement error than questions that are hard to understand or that are difficult to answer for some other reason" (Groves et al. 2009, p. 259). As a result, the problems detected by the qualitative methods should in theory be related to quantitative measures of response validity. Question problems identified by the qualitative methods could also be attributed to lower reliability of survey items if the problems are not systematic in their effects (for example, some respondents misinterpret the questions in one way while other respondents interpret the questions in another way).

Most of the studies in Table 1 compare several qualitative techniques to each other; this is unfortunate since the ultimate standards by which items should be judged are quantitative – whether the items yield accurate and reliable information. The study described here attempts to fill this gap in the literature. We compare results from both qualitative and quantitative assessments of a set of items, including estimates of item validity and reliability, and assess how well the conclusions from qualitative methods for question evaluation stack up against the conclusions from more direct quantitative estimates of validity and reliability.

As one reviewer noted, some question "problems" may not lead to response error but interrupt the flow of the interview. Both types of problems are typically addressed in the evaluation and pretesting process. We completely agree with this view, and for the remainder of this paper, we use the term "problem" to refer to suspected or purported problems identified by a given evaluation method without implying that these "problems" actually reduce the value of the data. Still, we believe that question evaluations are mainly done to ensure that the data that are ultimately collected are valid and accurate and that the main value of question evaluation methods is in improving data quality rather than improving the flow of the questions.

2. Comparing Five Evaluation Methods

The five methods we compare include two qualitative methods (expert reviews and cognitive interviews) and three quantitative methods (measures of validity and reliability and estimated error rates from latent class analysis). We chose expert reviews and cognitive interviews because they are popular methods for evaluating survey questions. We included latent class analysis because of its ability to estimate error rates without an external gold standard. And last but not least, we included validity and reliability because these are the ultimate standards a good item should meet.

We begin by describing each of these methods and reviewing the prior studies that have examined them; then in the Section 3, we describe how we compared them.

2.1. Expert Reviews

One relatively quick and inexpensive method for evaluating draft survey questions is to have experts in questionnaire design review them for problems. Not surprisingly, expert

reviews have become a common practice in questionnaire development (Forsyth and Lessler 1991). As Willis et al. (1999) point out, expert reviews can be conducted individually or in group sessions. In addition, the experts can rely exclusively on their own judgments, making informal assessments that typically yield open-ended comments about the survey items to be evaluated, or they can be guided by formal appraisal systems that provide a detailed set of potential problem codes.

Four studies have examined the effectiveness of expert reviews, and they differ somewhat in their findings (see Table 1 for details). Two of the studies found that expert reviews identified more problems than other methods, such as cognitive interviews (Presser and Blair 1994; Willis et al. 1999), but Rothgeb and her colleagues (2001) reported that expert reviews identified roughly the same number of problems with questions as cognitive interviews, and that the two methods identified different questions as problematic. Finally, Jansen and Hak (2005) report that their three-step cognitive testing procedure found more problems than an expert review. The three-step variant on cognitive interviewing developed by Jansen and Hak (2005) begins with a concurrent think-aloud, follows that with probing the attempt to clarify observed during the think-aloud portion of the interview, and concludes with a debriefing interview to explore the respondent's problems in answering the questions. In these studies, there is no independent evidence that the "problems" identified by the experts or those found in cognitive interviews are, in fact, problems for the respondents in the survey. Expert reviews are especially likely to identify problems related to data analysis and question comprehension (Presser and Blair 1994; Rothgeb et al. 2001). In addition to turning up lots of potential problems, expert reviews are less expensive than cognitive interviews or behavior coding (Presser and Blair 1994).

2.2. *Cognitive Interviewing*

As we noted earlier, cognitive interviewing relies on verbalizations by respondents to identify problems with the questions. Even though cognitive interviewing has become popular among survey practitioners, there is little consensus about the exact procedures that cognitive interviewing encompasses or even about the definition of cognitive interviewing (Beatty and Willis 2007). Beatty and Willis (2007) offer a useful definition; cognitive interviewing is "the administration of 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 its author intends" (p. 288). They also noted that cognitive interviews have been carried out in various ways. Some cognitive interviewers use think-alouds (either concurrent or retrospective), but others rely mainly on probes (either scripted or generated on the fly by the interviewers) intended to shed light on potential problems in the response process.

The evidence regarding the effectiveness of cognitive interviewing is inconsistent (see Table 1). Some studies have found that cognitive interviews detect fewer problems than expert reviews (Jansen and Hak 2005; Presser and Blair 1994; Willis et al. 1999), but Rothgeb and colleagues (2001) found that the two methods identified about the same number of problems. Cognitive interviews may find more problems than behavior coding

(Presser and Blair 1994), or the opposite may be true (Willis et al. 1999). In addition, Presser and Blair (1994) found that cognitive interviews identified more problems than conventional pretesting. Rothgeb and colleagues (2001) showed that cognitive interviews detected fewer problems than the formal appraisal method.

Willis and Schechter (1997) carried out several experiments testing whether predictions based on cognitive interviewing results were borne out in the field, and concluded that the predictions were largely confirmed. Other studies show that cognitive interviewing produces reasonable consistency across organizations at least in the number of problems identified (Rothgeb et al. 2001; Willis et al. 1999), and Fowler and Roman (1992) claim there is reasonable agreement across two sets of cognitive interviews done by different organizations but do not attempt to assess the level of agreement quantitatively. The results of Presser and Blair (1994) are less reassuring; they argue that cognitive interviews were less consistent across trials than expert reviews or behavior coding in the number of problems identified and in the distribution of problems by type.

2.3. Reliability and Validity

Expert reviews and cognitive interviews generally produce only qualitative information, typically in the form of judgments (either by the experts or the cognitive interviewers) about whether an item has a problem and, if so, what kind of problem. Still, most survey researchers would agree that the ultimate test a survey question must meet is whether it produces consistent and accurate answers — that is, whether the question yields reliable and valid data. These quantitative standards are rarely employed to pretest or evaluate survey questions because they require the collection of special data. For example, the reliability of an item can be assessed by asking the same question a second time in a reinterview, but this entails carrying out reinterviews. Or validity might be assessed by comparing survey responses to some gold standard, such as administrative records, but that requires obtaining the records data and matching them to the survey responses.

The most common strategy for estimating the reliability of survey items is to look at correlations between responses to the same questions asked at two different time points, a few weeks apart (e.g., O'Muircheartaigh 1991). This method of assessing reliability assumes that the errors at the two time points are uncorrelated. As Saris and Gallhofer (2007, pp. 190–192) note, the correlation between the same item (say, y_1) administered on two occasions (y_{11} and y_{12}) is not a pure measure of reliability, but is the product of the reliabilities of the item at time 1 (r_{11}) and time 2 (r_{12}) and the correlation between the true scores over time (s):

$$\begin{aligned}\rho(y_{11}, y_{12}) &= r_{11} s r_{12} \\ &= r_1^2 s\end{aligned} \tag{1}$$

The equation simplifies if we assume that the reliability of the item remains the same across the two occasions; the result is shown in the second line of Equation 1 above. Since the stability over time (s) is a characteristic of the true score rather than of the items, it follows that ranking a set of items (y_1, y_2, y_3) that measure the same construct by their correlations with themselves over two occasions is identical to ranking them by their reliability. The major drawback of estimating reliability through over time correlations is

the possibility of correlated errors in the test and retest due to learning or memory effects. Because we administered the items in different surveys conducted several weeks apart, we believe that any learning or memory effects are likely to have had only minimal impact on our ranking of the items by their test-retest reliability.

A simple approach for assessing the validity of survey items is to measure the correlations between each of the items to be evaluated and other questions to which they ought, in theory, to be related. Again, this is not a pure measure of validity (see Saris and Gallhofer 2007, p. 193). The correlation between an item of interest (y_1) and some other variable (x) is the product of the reliability (r_1) of y_1, its validity (v_1), and the true correlation (ρ) between the underlying constructs measured by x and y_1:

$$\rho(y_1, x) = r_1 v_1 \rho \tag{2}$$

However, as Equation 2 shows, because ρ is a property of the underlying constructs, ranking a set of items tapping the same construct by their correlation with some other variable is equivalent to ranking them by their overall accuracy – that is, by the product of the reliability (which reflects only random measurement error) and the validity (which reflects only systematic error).

Alternative measures of validity and reliability can be obtained using the SQP program of Saris and Gallhofer (Saris and Gallhofer 2007). Based on a meta-analysis of 87 multitrait-multimethod (or MTMM) experiments, the SQP program produces estimates of reliability, validity, and quality (a product of reliability and validity). Reliability is defined as one minus the random error variance over the total variance, and quality is defined as the proportion of the observed variance explained by the latent construct (Saris and Gallhofer 2007).

2.4. Latent Class Analysis (LCA)

As we already noted, latent class analysis is a statistical procedure that has been used to identify survey questions with high levels of measurement error. Proponents of the use of LCA in questionnaire development argue that it does not require error-free gold standards. Instead, it takes advantage of multiple indicators of the same construct and models the relationship between an unobserved latent variable (a.k.a., the construct) and the multiple observed indicators. The indicators are not assumed to be error-free. However, the errors associated with the indicators have to be independent conditional on the latent variable. This assumption – the local independence assumption – is almost always made in applications of LCA models. When this is satisfied, LCA produces unbiased estimates of the unconditional probabilities of membership in each of the latent classes (e.g., $P(c = 1)$

Table 2. Key parameters in latent class models

	Latent class	
Observed value	c = 1	c = 2
$u_1 = 1$	$P(u_1 = 1 \mid c = 1)$	$P(u_1 = 1 \mid c = 2)$
$u_1 = 2$	$P(u_1 = 2 \mid c = 1)$	$P(u_1 = 2 \mid c = 2)$
Unconditional probabilities	$P(\mathrm{c} = 1)$	$P(\mathrm{c} = 2)$

in Table 2 below). These unconditional probabilities represent the prevalence of each class in the population. LCA also produces estimates of the probability of each observed response conditional on membership in each latent class. For example, in a two-class model like the one in Table 2, the probability that a binary item u_1 is equal to 1 conditional on being in the first latent class ($c = 1$) is $p_{1|1} = P(u_1 = 1|c = 1)$, and the probability that this particular item is equal to 2 conditional on being in the first latent class is $p_{2|1} = P(u_1 = 2|c = 1)$.

Two of the conditional probabilities in Table 2 represent error rates. These are the probabilities of a false positive ($P(u_1 = 1|c = 2)$) and false negative response ($P(u_1 = 2|c = 1)$) to the question, given membership in latent class c. A high false positive or false negative probability signals a problem with a particular item. The primary purpose of applying LCA to the evaluation of survey questions is to identify questions that elicit error-prone responses – that is, questions with high rates of false positives or false negatives. When the local independence assumption is not satisfied (e.g., when the responses to three items measuring the same underlying construct are correlated even within the latent classes), then the LCA estimates of the unconditional and conditional probabilities may be erroneous.

Biemer and his colleagues have carried out several studies that use LCA to identify flawed survey questions and to explore the causes of the problems with these items (Biemer 2004; Biemer and Wiesen 2002). For example, Biemer and Wiesen (2002) examined three indicators used to classify respondents regarding their marijuana use and used LCA estimates to pinpoint why the multi-item composite indicator disagreed with the other two indicators. The LCA results indicated that the problem was the large false positive rate in the multi-item indicator (Biemer and Wiesen 2002).

A recent paper by Kreuter, Yan, and Tourangeau (2008) attempted to assess the accuracy of the conclusions from such applications of LCA. Kreuter and her colleagues conducted a survey of alumni from the University of Maryland that included several questions about their academic records at the university. They compared the survey answers to university records. They also fit LCA models to the survey responses and found that the LCA approach generally produced qualitative results that agreed with those from the comparison with the records data; the item that the LCA model singled out as having the largest estimated misclassification rate was also the one with the largest disagreement with the university records according to a traditional "gold standard" analysis. However, the quantitative estimates of the error rates from the LCA models often differed substantially from the error rates found in comparisons to the records data.

3. Research Design and Methods

In this study, we carried out two large-scale web surveys that allow us to measure the reliability of the answers for some of our items across two interviews (see Equation 1) and the construct validity of the items by examining the relation of each item to other questions in the same survey (as in Equation 2). We examined a total of fifteen items, five triplets consisting of items intended to measure the same construct. All fifteen items were assessed by four experts, tested in cognitive interviews, and investigated by latent class modeling. Six of the items were administered as part of a two-wave web survey that allowed us to

measure both the reliability and construct validity of the items; the nine remaining items were administered in a one-time web survey, and we used the data from this survey to estimate the construct validity for these items.

3.1. Questions

The five triplets concerned a range of constructs — evaluations of one's neighbors, reading habits, concerns about one's diet, doctor visits in the past year, and feelings about skim milk.

One member of each of the triplets administered as part of a two-wave web survey was deliberately "damaged," that is, it was written so as to have more serious problems than the other two items in the triplet. For example, the neighborhood triplet asks respondents to evaluate their neighbors:

1a. How much do you agree or disagree with this statement? People around here are willing to help their neighbors. (Strongly agree, Agree, Disagree, Strongly Disagree)
1b. In general, how do you feel about people in your neighborhood?
  0.1. They are very willing to help their neighbors.
  0.2. They are somewhat willing to help their neighbors.
  0.3. They are not too willing to help their neighbors.
  0.4. They are not at all willing to help their neighbors.
1c. How much do you agree or disagree with this statement? People around here are willing to help other people. (Strongly agree, Agree, Disagree, Strongly Disagree)

The third item was written to be vaguer and therefore worse than the other two items. All fifteen items making up the five triplets are included in Appendix 1.

3.2. Expert Reviews

We asked four experts in questionnaire design to assess all fifteen items. Two of the experts were authors of standard texts on questionnaire design; the third has written several papers on survey questions and taught classes on questionnaire design; and the fourth was an experienced staff member of the unit charged with testing questions at one of the major statistical agencies in the United States.

We told the experts that we were doing a methodological study that involved different methods of evaluating survey questions but did not give more specific information about the aims of the study. We asked them to say whether each item had serious problems (and, if it did, to describe the problems) and also to rate each item on a five-point scale. The scale values ranged from "This is a very good item" (=1) to "This is a very bad item" (=5). We used the average of the four ratings of each item to rank order the items.

3.3. Cognitive Interviews

All fifteen of the items were tested in interviews carried out by five experienced cognitive interviewers at the Survey Research Center (SRC) at the University of Michigan. Three versions of the questionnaire were tested, each containing one item from each of the five triplets plus some additional filler items. Respondents were randomly assigned to get one version of the questionnaire. A total of 15 cognitive interviews were done on each version.

The respondents were adults (18 years old or older) recruited from the Ann Arbor area and paid $40 for participating. (Respondents were also reimbursed for their parking expenses.) The respondents included 22 females and 23 males. Sixteen were 18 to 34 years old; 15 were 35 to 49 years old; and 14 were 50 years or older. Thirty of the respondents were white; ten were African-American; and five characterized themselves as "Other." Fourteen had a high school diploma or GED; 25 had at least some college; and six had more than a four-year college degree.

The interviews took place at SRC's offices and were recorded. An observer also watched each interview though a one-way mirror. The cognitive interviewers asked the respondents to think aloud as they formulated their answers, administered pre-scripted "generic" probes (such as "How did you arrive at your answer?" or "How easy or difficult was it for you to come up with your answers?"; see Levenstein et al. 2007 for a discussion of such probes), and followed up with additional probing ("What are you thinking?" or "Can you say a little more?") to clarify what the respondents said or how they had arrived at an answer. (Our cognitive interviews thus included both concurrent probes and immediate retrospective probes.) After the respondent completed each item, both the interviewer and the observer checked a box indicating whether he or she thought the respondent had experienced a problem in answering the question. The interviewer and observer also indicated the nature of the problems they observed (that is, whether the problem involved difficulties with comprehension, retrieval, judgment or estimation, reporting, or some combination of these). We counted a respondent as having had a problem with an item if both the interviewer and the observer indicated the presence of a problem.

3.4. Web Surveys: Reliability, Validity, and LCA Error Rates

3.4.1. Web Survey Data Collection

The two first triplets (see the neighborhood triplet — items 1a-1c — and the triplet of book items — 2a-2c — in Appendix I) were administered as part of two web surveys that were conducted about five weeks apart. The six questions were spread throughout the questionnaires in the two surveys. Respondents who completed the first web survey were invited to take part in the second one. They were not told that the second survey had any relationship to the first. The second survey was the subject of an experiment described in detail by Tourangeau et al. (2009). Briefly, the invitation to the second survey and the splash page (i.e., the first web screen shown to respondents once they logged on) for that survey systematically varied the description of the topic and sponsor of the survey. (Neither of the experimental variables affected the items we examine here.)

A total of 3,000 respondents completed the first survey. Half of the respondents came from Survey Sampling Inc.'s (SSI) Survey Spot frame, and the other half were members of the e-Rewards web panel. Both are opt-in panels whose members had signed up online to receive survey invitations via e-mail. The response rate (AAPOR 1; see American Association for Public Opinion Research 2008) for the first wave of the survey was 4.1% among the SSI members and 14.8% among the e-Rewards members. A total of 2,020 respondents completed the second wave of the survey. The response rate (AAPOR 1) for the second wave was 61.1% for the SSI members and 73.7% for the e-Rewards panel.

The first wave of the survey was conducted from January 25, 2007, to February 1, 2007; the second wave, from March 2, 2007, to March 19, 2007.

The response rates for this survey, particularly for the first wave, were quite low, and neither panel from which the respondents were drawn is a probability sample of the general population. As a result, Tourangeau and his colleagues (Tourangeau et al. 2009) attempted to measure the effects of any selection and nonresponse biases on the representativeness of the responding panel members. They compared the respondents from each wave of the survey to figures from the American Community Survey (ACS) on sex, age, race, Hispanic background, and educational attainment. In both waves, the web respondents did not depart markedly from the ACS figures on age, race, or Hispanic background. The web samples did underrepresent persons who were 18 to 29 years old (members of this group made up 14 percent of the wave 1 sample and 11 percent of the wave 2 sample, versus 21 percent of the population according to the ACS) and overrepresented those who were 60 years and older (28 percent and 32 percent in waves 1 and 2 of our survey, versus 22 percent in the ACS). The web samples also overrepresented college graduates (50 percent and 52 percent in the two waves, versus 25 percent in the ACS) and underrepresented those with less than a high school education (1 percent in both waves, versus 14 percent in the ACS). Of course, there could still be biases in the results we present, unless the data are missing at random (MAR), conditional on these variables.

The items making up the final three triplets – the diet triplet (items 5a-5c), the doctor visits triplet (items 6a-6c), and the skim milk triplet (items 7a-7c; see Appendix I) – were administered as part of a one-time web survey completed by 2,410 respondents. Half of these respondents came from the SSI Survey Spot panel, and the other half were from the Authentic Response web panel. The response rate (AAPOR 1) was 1.9% among the SSI members and 16.5% among the members of the Authentic Response panel. The survey was carried out from September 2 to September 23, 2008.

Again, because the web sample was a non-probability sample and the response rate was low, we compared the demographic makeup of the respondents in our second study sample to that of the American Community Survey. The results were similar to those for our earlier web survey. The web respondents in the second study also tended to be more highly educated and older than the U.S. adult population as a whole; in addition, they were more likely to be white (89 percent versus 77 percent in ACS) and less likely to be Hispanic (4 percent of our web respondents versus 13 percent in the ACS) than the U.S. general population. Again, this does not demonstrate an absence of bias in the results we present.

The nine target questions were spread throughout the questionnaire in the second web survey, with one item from each triplet coming at the beginning of the survey, one coming in the middle, and one coming at the end.

3.4.2. Reliability and Validity

Because we intended to apply latent class models to each target item, we first recoded the responses to all fifteen target items to yield dichotomies. For example, with item 1b (the second item in the neighborhood triplet, see Appendix I), we combined the first two response options and the last two. The results presented below in Tables 3 through 5 do not differ markedly if we do not dichotomize the items offering more than two response options, but treat them as scales instead.

We computed reliabilities for the neighborhood and book triplets (the first six items in Appendix I). Our reliability estimate was the correlation between responses to the same question in the two waves (after recoding the answers to yield dichotomies). This is the same approach summarized earlier in Equation 1. Similarly, the validity coefficients were the correlations between the dichotomized responses to the items in each triplet with some other item in the questionnaire. For example, for the three neighborhood items (items 1a, 1b, and 1c above), we correlated dichotomized responses (in the initial interview) with answers to the first item in the wave 1 questionnaire, which asked for an overall assessment of the respondent's neighborhood (see Appendix I for detailed wordings of all the questions examined in this article). This is the same approach described earlier (see Equation 2).

3.4.3. LCA Error Rates

We fit latent class models (like the one summarized in Table 2 above) to the data from the three items in each triplet, using the Mplus software (Muthén and Muthén 1998–2007). We dichotomized each item prior to fitting the latent class models. For each triplet, we fit a model with two latent classes and estimated the false positive and false negative rates for each of the three items presented in Appendix II. In ranking the items in each triplet, we used the sum of the two error rates for each item and labeled it as 'misclassification rate' in Tables 3 and 4.

4. Results

Tables 3 and 4 present the main results from the study. Table 3 displays the summary statistics for the six items included in the two-wave web survey. It shows the mean ratings of the experts for each item (with higher ratings indicating a worse item), the proportion of cognitive interviews in which the item was found to have a problem, the misclassification rates from the latent class modeling, and the validity and reliability coefficients for each of the items. Table 4 displays similar summary statistics for the nine items included in the second web study. Because the second web study was a single-wave survey, we could not compute reliability estimates for the nine items in that survey.

Both tables also provide ranking of the items within each triplet and standard errors for the main statistics. For the statistics derived from the web survey data (that is, the reliability and validity coefficients and the error rates from the latent class models), we used the "random groups" approach to calculate the standard errors for the statistics themselves as well as for the differences between pairs of statistics (see Wolter 1985, ch. 2, for a detailed description of the random groups technique). We randomly subdivided the sample into 100 replicates and used the variation in the statistic of interest across replicates to estimate the standard error:

$$SE(\hat{\theta}) = \left[\frac{1}{k}\sum\frac{(\hat{\theta}_i - \bar{\theta})^2}{(k-1)}\right]^{1/2} \tag{3}$$

where $\hat{\theta}_i$ is a statistic (such as a reliability coefficient) computed from replicate i and $\bar{\theta}$ is the mean of that statistic across all 100 replicates. We also used the random groups

Table 3. Indicators of item quality (and ranks), by item — Study 1

	Expert reviews		Cognitive interviews		LCA model error rates			Validity			Reliability		
	Mean rating (higher is worse)	SE	% with problems	SE	Full sample estimate	Mean across replicates	SE	Full sample estimate	Mean across replicates	SE	Full sample estimate	Mean across replicates	SE
Neighborhood items													
Item 1a	4.25 (1)	.48	26.7 (1)	11.8	.092 (1)	.088	.015	.318 (2)	.313	.028	.449 (2)	.449	.030
Item 1b	4.50 (1)	.29	21.4 (1)	11.4	.189 (3)	.158	.021	.341 (1)	.345	.030	.566 (1)	.599	.034
Item 1c	4.25 (1)	.25	40.0 (2)	13.1	.183 (2)	.145	.018	.322 (2)	.317	.026	.549 (1)	.550	.031
Book items													
Item 2a	3.75 (1)	.63	46.7 (1)	13.3	.203 (2)	.196	.011	.227 (1)	.219	.026	.680 (1)	.672	.019
Item 2b	3.50 (1)	.65	50.0 (1)	13.9	.013 (1)	.016	.003	.226 (1)	.215	.023	.717 (1)	.706	.017
Item 2c	2.75 (1)	.85	46.7 (1)	13.3	.067 (1)	.060	.007	.231 (1)	.219	.024	.725 (1)	.724	.018

Table 4. Indicators of item quality (and ranks), by item — Study 2

	Expert reviews		Cognitive interviews		LCA model error rates			Validity		
	Mean rating (higher is worse)	SE	% with problems	SE	Full sample estimate	Mean across replicates	SE	Full sample estimate	Mean across replicates	SE
Diet items										
Item 5a	4.00 (1)	.48	60.0 (3)	13.1	.298 (1)	.304	.025	− .282 (2)	− .274	.023
Item 5b	4.50 (1)	.29	0.0 (1)	0.0	.468 (3)	.400	.028	− .404 (1)	− .405	.021
Item 5c	4.25 (1)	.25	13.3 (2)	9.1	.386 (2)	.349	.017	− .354 (1)	− .358	.020
Doctor visit items										
Item 6a	2.75 (1)	.48	46.7 (2)	13.3	.046 (1)	.038	.010	− .408 (1)	− .407	.011
Item 6b	3.00 (1)	.41	13.3 (1)	9.1	.042 (1)	.037	.005	− .419 (1)	− .412	.013
Item 6c	5.00 (2)	.00	46.7 (2)	13.3	.039 (1)	.035	.010	− .399 (2)	− .395	.012
Skim milk items										
Item 7a	4.25 (2)	.25	20.0 (1)	10.7	.262 (3)	.246	.015	− .207 (1)	− .215	.018
Item 7b	2.25 (1)	.63	57.1 (2)	13.7	.038 (1)	.043	.007	− .194 (1)	− .208	.018
Item 7c	3.50 (2)	.65	60.0 (2)	13.1	.061 (2)	.060	.008	− .172 (2)	− .184	.018

technique to estimate differences between pairs of statistics (e.g., between the reliabilities of items 1a and 1b). Because each evaluation method yields results on different metrics, we rank order the questions based on their performance on each method. These ranks ignore "small" differences, which we defined somewhat arbitrarily as differences of one standard error or less. These ranks are displayed in Tables 3 and 4 in parentheses.

For the neighborhood items (items 1a, 1b, 1c in Table 3), the validities of the items are quite similar, but item 1a seems to have the lowest reliability. The experts seem to agree with these quantitative results; they rated the items as not very different from each other and saw all three items as problematic. The latent class model picks out item 1a as having the *lowest* misclassification rate of the three items; that item was also the least reliable item. Cognitive interviewing was the only method that picked out the damaged item (item 1c) as worse than the other two items.

All three items in the book triplet (items 2a, 2b, 2c in Table 3) had similar estimated validities and also similar estimated reliabilities. Cognitive interviews and expert reviews do not find much difference between the three items in this triplet. The LCA model identifies item 2a as having the highest misclassification rate among the three items. None of the five methods picked out the damaged item (item 2c) as worse than the other two items.

For the diet items (items 5b, 5b, and 5c in Table 4), both the validity analysis and the cognitive interviews indicate that items 5a is the weakest item among the three, whereas the LCA picks it out as the best member of the set.

For the doctor visit items (items 6a, 6b, and 6c in Table 4), the experts agree with the validity analysis in finding 6c the weakest item in this triplet. The LCA method, however, did not seem to find much difference between them. Cognitive interviews produced the opposite conclusions, identifying item 6b as the best item.

Expert reviews and the LCA method both ranked item 7b as the best in this triplet on skim milk (items 7a, 7b, and 7c in Table 4). By contrast, cognitive interviews and the validity measure favored item 7a over the other two.

So far, we have considered only how the different methods rank order the items within each triplet; this corresponds with how a questionnaire designer might make a decision about the items in a given triplet. Table 5 presents a quantitative assessment of the agreement across methods; the table shows the matrix of correlations among the mean expert ratings, the proportion of cognitive interviews in which both the interviewer and observer thought the item exhibited problems, the misclassification rates from the LCA models, and the estimates of quality obtained from SQP predictions provided by Dr. Willem Saris. (We drop the reliability estimates from this analysis since they are available only for six of the items.) It is reasonable to compare the validity estimates used in Tables 3 and 4 within triplets, but across triplets the comparisons are confounded with strength of the underlying relationship between the construct tapped by our three items and the construct we are trying to predict (see Equation 2, presented earlier, where this relationship is represented by ρ). We therefore include the correlations of the other methods with a statistic we call the validity ratio in Table 5. The validity ratio is just the ratio between the validity estimate for a given item within a triplet and the lowest validity estimate for the items in that triplet. This ratio renders the correlations across triplets more comparable by removing the effect of ρ. Italicized entries in the table take the opposite of

Table 5. Correlations (and number of items) among quantitative indicators of item quality

				Validity analysis		
	Expert review	Cognitive interviews	Latent class model	Validity estimate	Validity ratio	Quality
Expert rating	–	− *.408* (15)	.526 (15)*	*.326* (15)	*.230* (15)	*.608 (15)**
Cognitive interview		–	− *.570* (15)*	− .560 (15)*	− .715 (15)*	− .070 (15)
Latent class model			–	*.201* (15)	*.757* (15)*	*.369 (15)*
Validity analysis						.063 (15)

Note: * indicates the $P < .05$ (two-tailed). The indicator from the expert review was the mean rating of the item across the four experts; for the cognitive interviews, it was the proportion of interviews in which both coders judged the item to have a problem; for the latent class analysis, it was the misclassification rate; for the validity analysis, the validity estimate refers to the correlation of the item with a conceptually related item as used in Tables 3 and 4; validity ratio is the ratio between the validity estimate for a given item within a triplet and the lowest validity estimate for the items in that triplet; and, for the quality measure, it was the prediction from the SQP program (provided by Dr. Willem Saris). Italics indicate that the entry takes the opposite of the direction expected.

the direction expected. Just to be clear, we expected the validity estimates, the item reliabilities, and the quality measure from the SQL model to be positively correlated with each other. These measures are all quantitative measures of item quality, with higher numbers indicating a "better" item. Similarly, we expected the expert ratings, the proportion of cognitive interviews finding a problem with an item, and the misclassification rates to be positively correlated with each other, since they all measure the degree to which an item has problems. Finally, the measures in the first group should correlate negatively with those in the second group.

As Table 5 makes clear, the correlations are not very high and several of them go in the wrong direction. The indicators seem to fall into two groups. The expert ratings show good agreement with the LCA misclassification rates. The correlation between the mean of the expert ratings and the misclassification rates from the LCA models was significant ($r = .526$, $p < .05$, based on $n = 15$ items). The cognitive interviews and the validity analyses also produce converging conclusions. The correlation between the proportion of interviews in which a problem was found with an item and the validity coefficient for the item was significant and, as expected, negative (a higher rate of problems found in the cognitive interviews was associated with lower validity estimates; $r = -.560$, $p < .05$); this correlation increases to $-.715$ when we use our validity ratio statistic in place of the original validity estimates.

There are two other significant correlations in the table and both are in the wrong direction. The correlation between the LCA misclassification rates and the proportions of cognitive interviews in which a problem was observed with an item was significant but negative ($r = -.570$, $p < .05$) – the higher the proportion of cognitive interviews revealing problems with the item, the lower the misclassification rate according to the LCA models. The LCA error rates also are significantly correlated (in the wrong direction) with our validity ratio statistic. The correlation between expert ratings and the quality measure was significant but positive ($r = .608$, $p < .05$) – the higher the experts' ratings (and the worse the items), the higher the predicted quality according to SQP.

5. Conclusions and Discussion

This article examined a variety of question evaluation methods. As the studies reviewed in Table 1 might suggest, the methods generated different results, giving inconsistent, even contradictory, conclusions about the items in a triplet. As shown in Table 5, even though we find considerable agreement with the expert ratings and the LCA results and the cognitive interview results and the validity analysis, most of the correlations among the indicators generated by each method take the opposite of the direction expected (see Table 5).

Why are the results not more consistent across different methods? One possibility is that the methods do not all give valid indications of problems with the items. All of the methods make assumptions, and these assumptions may often be violated in practice. In an earlier paper examining the use of LCA models to evaluate survey items, Kreuter et al. (2008) found that the LCA models often gave good qualitative results (e.g., correctly identifying the worst item among a set of items designed to measure the same construct) but were substantially off in their quantitative estimates of the error rates. LCA models

make strong assumptions and their results seem to be sensitive to the violation of those assumptions (e.g., Spencer 2008). The data here suggest they are not a substitute for direct estimates regarding item validity. Of course, the validity estimates we present are hardly perfect or assumption-free either; as we noted, they reflect both the properties of the items and the strength of the underlying relationship between the relevant constructs.

The more qualitative methods may be especially prone to yielding unreliable or invalid conclusions. As Presser and Blair (1994) first demonstrated, multiple rounds of expert reviews and cognitive interviews often yield diverging conclusions. More recently, Conrad and Blair (2004) have found that cognitive interviews may be prone to false positives in question evaluation, evidenced by the high percentage of items found to have problems (see also Levenstein et al. 2007 and Conrad and Blair 2009). Our results indicate some convergence between the cognitive interview results and the validity estimates. This was true even though the consistency across cognitive interviewers was quite low. Three of the cognitive interviewers did seven or more cognitive interviews and, for these three, we calculated the proportion of interviews in which a problem was found with each item. The correlations in these proportions across the fifteen items ranged from only .143 to .326. (The convergence across experts was a little higher; the median correlation in the expert ratings was .360). The relatively low agreement across cognitive interviewers and across experts may put a low ceiling on their convergence with quantitative measures of item performance such as the validity and reliability measures used here.

Another possible reason for the low consistency across methods is the low agreement among question evaluation methods about the *nature* of the problem. We calculated the proportion of the experts who saw each item as presenting a comprehension problem, a recall problem, or a problem with judgment or reporting, and we correlated these proportions with the proportion of cognitive interviews in which the interviewer indicated there was a problem of the same type. (Problems in judgment and reporting were relatively rare, which is why we combined those categories.) The correlations were $-.09$ and $-.33$ for comprehension and judgment/reporting problems; the correlation was .86 for recall problems. This picture does not change much if we look at the proportion of the time the observers of the cognitive interviews indicated that there was a problem of a given type; the correlations are very similar (.03 for comprehension problems, $-.47$ for judgment or reporting programs, and .80 for recall problems).

Thus, one potential source of the conflicting conclusions about an item is that the different question evaluation methods focus on different aspects of the questions and different types of problems. As one reviewer pointed out, the experts and the cognitive methods tend to concentrate more on how well the underlying constructs are measured and somewhat less on the response scales. By contrast, the latent class methods focus on the probabilities of errors and marginal distributions of responses whereas the quality measures from the SQP predictions emphasize purely on the effects of the form of the questions and the response scales.

Whatever the reason for the diverging results across question evaluation methods, 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 (cf. Presser et al. 2004a). Most

textbooks advocate applying more than one evaluation method in testing survey questions, and our results indicate that a multi-method approach to question evaluation may be the best course for the foreseeable future. The natural next steps for this research are to understand how to reduce the inconsistencies and to investigate how to best combine different evaluation methods while capitalizing on the strengths of each. We believe that there is no substitute for the traditional psychometric indicators and we recommend that more questionnaire evaluation studies include validity and reliability measures. This may be expensive, but there seems to be no low-cost qualitative substitute for these indicators of item quality. We believe that the methods used to evaluate survey questions should have a firmer scientific basis and, in our view, more studies with credible estimates of the validity and reliability of the items are needed if we are ever to understand how much confidence we can place on the different qualitative methods currently used to evaluate survey questions.

Appendix I: Items Used in the Study

Items included in two-wave web survey

Neighborhood Triplet

1a. How much do you agree or disagree with this statement? People around here are willing to help their neighbors. (Strongly agree, Agree, Disagree, Strongly Disagree)
1b. In general, how do you feel about people in your neighborhood?

1. They are very willing to help their neighbors.
2. They are somewhat willing to help their neighbors.
3. They are not too willing to help their neighbors.
4. They are not at all willing to help their neighbors.

1c. How much do you agree or disagree with this statement? People around here are willing to help other people. (Strongly agree, Agree, Disagree, Strongly Disagree)

Book Triplet

2a. Which, if any, of the following have you done in the past 12 months? . . . Read more than five books? (Yes, No)
2b. During the past year, how many books did you read?
2c. During the past year, about how many books, either hardcover or paperback, including graphic novels, did you read either all or part of the way through?

Question used in validity estimates for Neighborhood triplet (1a, 1b, and 1c)

3. The first few questions are about some general issues. First, how would you rate your neighborhood as a place to live? (Poor, Fair, Good, Very Good, Excellent)

Question used in validity estimates for Book triplet (2a, 2b, and 2c)

4. What is the highest level of education you've completed? (Grades 1 through 8, Less than High School Graduate, High School Graduate, Some college/Associates' degree, College graduate, Master's degree, Doctoral/Professional degree)

Items included in final web survey

Diet Triplet

5a. On a scale of 0 to 9, where 0 is not concerned at all and 9 is strongly concerned, how concerned are you about your diet? (Nine-point scale, with labeled endpoints)

5b. Would you say that you care strongly about your diet, you care somewhat about your diet, you care a little about your diet, or you don't care at all about your diet? (Strongly, Somewhat, A little, Not at all)

5c. Do you worry about what you eat or do you not worry about it? (Worry about what I eat; Do not worry about what I eat)

Doctor Visit Triplet

6a. The next item is about doctor visits — visits to a physician or someone under the supervision of a physician, such as a nurse practitioner or physician's assistant for medical care. During the last 12 months — that is, since [INSERT CURRENT MONTH] of 2007 — how many times have you visited a doctor? (Open-ended answer)

6b. Over the last 12 months, how many times have you seen a doctor or someone supervised by a doctor for medical care? (Open-ended answer)

6c. How many times have you seen a doctor over the past year? (0 times; − 2 times; 3–4 times; 5–6 times; 7 or more times)

Skim Milk Triplet

7a. Please indicate how you feel about the following foods . . . . Apples; Whole milk; Skim milk; Oranges (These items appeared in a grid, with a ten-point response scale; the end points of the scale were labeled "Like Very Much" and "Dislike Very Much")

7b. How much would you say you like or dislike skim milk? (Like very much; Like somewhat; Neither like nor dislike; Dislike somewhat; Dislike very much)

7c. How much would you say you agree or disagree with the statement "I like skim milk." (Agree strongly; Agree somewhat; Neither agree nor disagree; Disagree somewhat; Disagree strongly)

Question used in validity estimates for the Diet triplet (5a, 5b, 5c) and Skim Milk triplet (7a, 7b, 7c)

8. Indicate how much you favor or oppose each of the following statements . . . . "Maintaining healthy diet" (Strongly oppose, Somewhat oppose, Neither favor nor oppose, Somewhat favor, Strongly favor)

Question used in validity estimates for the Doctor Visit triplet (6a, 6b, and 6c)

9. How many different PRESCRIPTION DRUGS are you currently taking? (None, 1, 2, 3, 4, 5 or more)

Appendix II: False Positive and False Negative Rates, by Triplet and Item

	Neighborhood items (Triplet 1)		Book items (Triplet 2)		Diet items (Triplet 3)		Doctor visit items (Triplet 4)		Skim milk items (Triplet 5)	
	False positive	False negative	False positive	False negative	False positive	False negative	False positive	False negative	False positive	False negative
Item a	0.052	0.040	0.176	0.027	0.244	0.054	0.037	0.028	0.114	0.148
Item b	0.184	0.005	0.002	0.011	0.450	0.018	0.011	0.041	0.013	0.025
Item c	0.152	0.031	0.055	0.012	0.021	0.365	0.026	0.051	0.028	0.033

6. References

The American Association for Public Opinion Research (2008). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, (5th edition). Lenexa, Kansas: AAPOR.

Beatty, P.C. and Willis, G.B. (2007). Research Synthesis: The Practice of Cognitive Interviewing. Public Opinion Quarterly, 71, 287–311.

Biemer, P.P. (2004). Modeling Measurement Error to Identify Flawed Questions. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 225–246.

Biemer, P.P. and Wiesen, C. (2002). Measurement Error Evaluation of Self-Reported Drug Use: A Latent Class Analysis of the US National Household Survey on Drug Abuse. Journal of the Royal Statistical Society, Series A, 165, 97–119.

Biemer, P.P. and Witt, M. (1996). Estimation of Measurement Bias in Self-Reports of Drug Use with Applications to the National Household Survey on Drug Abuse. Journal of Official Statistics, 12, 275–300.

Conrad, F.G. and Blair, J. (2004). Aspects of Data Quality in Cognitive Interviews: The Case of Verbal Reports. In Questionnaire Development, Evaluation and Testing Methods, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 67–88.

Conrad, F.G. and Blair, J. (2009). Sources of Error in Cognitive Interviews. Public Opinion Quarterly, 73, 32–55.

Converse, J.M. and Presser, S. (1986). Survey Questions: Handcrafting the Standardized Questionnaire. Beverly Hills, CA: Sage.

DeMaio, T. and Landreth, A. (2004). Do Different Cognitive Interview Techniques Produce Different Results? In Methods for Testing and Evaluating Survey Questionnaires. S. Presser et al. (eds), pp. 891-08. Hoboken, NJ: John Wiley and Sons.

Ericsson, K.A. and Simon, H.A. (1980). Verbal Reports as Data. Psychological Review, 87, 215–257.

Forsyth, B.H. and Lessler, J.L. (1991). Cognitive Laboratory Methods: A Taxonomy. In Measurement Errors in Surveys, P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds). New York: John Wiley, 393–418.

Forsyth, B., Rothgeb, J., and Willis, G. (2004). Does Questionnaire Pretesting Make a Difference? An Empirical Test Using a Field Survey Experiment. In Questionnaire Development, Evaluation, and Testing, S. Presser, et al. (Eds.), pp. 525-546. Hoboken, NJ: John Wiley and Sons.

Fowler, F.J. (2004). The Case for More Split-Sample Experiments in Developing Survey Instruments. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 173–188.

Fowler, F.J. and Roman, A.M. (1992). A Study of Approaches to Survey Question Evaluation, Final Report for U.S. Bureau of the Census, Boston: Center for Survey Research.

Gerber, E. (1999). The View from Anthropology: Ethnography and the Cognitive Interview. In Cognition and Survey Research, M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (eds). New York: Wiley, 217–234.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2009). Survey Methodology. New York: Wiley.

Jansen, H. and Hak, T. (2005). The Productivity of the Three-Step Test-Interview (TSTI) Compared to an Expert Review of a Self-Administered Questionnaire on Alcohol Consumption. Journal of Official Statistics, 21, 103–120.

Kreuter, F., Yan, T., and Tourangeau, R. (2008). Good Item or Bad – Can Latent Class Analysis Tell? The Utility of Latent Class Analysis for the Evaluation of Survey Questions. Journal of the Royal Statistical Society, Series A, 171, 723–738.

Levenstein, R., Conrad, F., Blair, J., Tourangeau, R., and Maitland, A. (2007). The Effect of Probe Type on Cognitive Interview Results: A Signal Detection Analysis. In Proceedings of the Section on Survey Methods, 2007. Alexandria, VA: American Statistical Association, 3850–3855.

Loftus, E. (1984). Protocol Analysis of Responses to Survey Recall Questions. In Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines, T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau (eds). Washington, DC: National Academy Press.

Maynard, D.W., Houtkoop-Steenstra, H., Schaeffer, N.C., and van der Zouwen, J. (2002). Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview. New York: John Wiley and Sons.

McCutcheon, A.L. (1987). Latent Class Analysis. Newbury Park, CA: Sage.

Miller, K. (2009). Cognitive Interviewing. Paper presented at the Question Evaluation Methods Workshop at the National Center for Health Statistics.

Muthén, L.K. and Muthén, B.O. (1998–2007). Mplus User's Guide, (Fifth Edition). Los Angeles, CA: Muthén & Muthén.

O'Muircheartaigh, C. (1991). Simple Response Variance: Estimation and Determinants. In Measurement Error in Surveys, P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds). New York: John Wiley, 551–574.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J., and Singer, E. (2004a). Methods for Testing and Evaluating Survey Questions. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 1–22.

Presser S., Rothgeb J., Couper M.P., Lessler J.T., Martin E., Martin J., Singer E. (eds) (2004b). Methods for Testing and Evaluating Survey Questionnaires. New York: John Wiley.

Presser, S. and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? Sociological Methodology, 24, 73–104.

Rothgeb, J., Willis, G., and Forsyth, B. (2001). Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results. Proceedings of the Section on Survey Methods (2001). Alexandria, VA: American Statistical Association.

Saris, W.E. and Gallhofer, I.N. (2007). Design, Evaluation, and Analysis of Questionnaires for Survey Research. New York: John Wiley.

Schaeffer, N.C. and Presser, S. (2003). The Science of Asking Questions. Annual Review of Sociology, 29, 65–88.

Spencer, B.D. (2008). When Do Latent Class Models Overstate Accuracy for Binary Classifiers? Unpublished manuscript.

Tourangeau, R., Groves, R., Kennedy, C., and Yan, T. (2009). The Presentation of the Survey, Nonresponse, and Measurement Error. Journal of Official Statistics, 25, 299–321.

van der Zouwen, J. and Smit, J.H. (2004). Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and Question-Answer Sequences: A Diagnostic Approach. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 109–130.

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

Willis, G.B. and Schechter, S. (1997). Evaluation of Congitive Interviewing Techniques: Do the Results Generalize to the Field? Bulletin de Methodologie Sociologique, 55, 40–66.

Willis, G.B., Schechter, S., and Whitaker, K. (1999). A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What do They Tell US? In Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association. 28–37.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Received March 2010
Revised March 2012