

QUESTION UNDERSTANDING AID (QUAID) A WEB FACILITY THAT TESTS QUESTION COMPREHENSIBILITY

ARTHUR C. GRAESSER
ZHIQIANG CAI
MAX M. LOUWERSE
FRANCES DANIEL

Abstract When respondents do not understand the meaning of a survey question, they will not supply valid and reliable answers. Survey methodologists should therefore benefit from computer tools and other analytical schemes that help them identify problems with questions with respect to comprehension difficulty. We developed a Web facility called Question Understanding Aid (QUAID; www.psyc.memphis.edu/quaid.html) that assists survey methodologists in identifying problems with the wording, syntax, and semantics of questions on questionnaires. The survey methodologist enters the question into the Web facility, along with any context information and answer alternatives that accompany the question. QUAID quickly returns a list of potential problems with question comprehension, including unfamiliar technical terms, vague or imprecise relative terms, vague or ambiguous noun phrases, complex syntax, and working memory overload. This article describes QUAID and some empirical studies that have assessed the validity and utility of QUAID's critiques of questions. The output of QUAID was compared with the judgments of experts in language, discourse, and cognition during the development of the tool. In one evaluation, expert survey methodologists critiqued and revised problematic questions, whereas in a second evaluation survey methodologists evaluated the

ARTHUR C. GRAESSER is a professor in the University of Memphis Department of Psychology. ZHIQIANG CAI is a research scientist in the University of Memphis Department of Psychology. MAX M. LOUWERSE is a professor in the University of Memphis Department of Psychology. FRANCES DANIEL is a doctoral student in the University of Illinois, Chicago, Department of Psychology. This research was funded by grants awarded to Graesser by the Statistical Research Division of the U.S. Census Bureau (1998–99, 43-YA-BC-802930), the National Science Foundation (NSF; SES 9977969), and the Office of Naval Research (ONR; N00014-00-1-0917). Any opinions, findings, and conclusions or recommendations expressed in this material are ours and do not necessarily reflect the views of NSF, ONR, or the U.S. Census Bureau. Address correspondence to Arthur C. Graesser; e-mail: a-graesser@memphis.edu.

doi:10.1093/poq/nfj012

© The Author 2006. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

quality of original problematic questions, questions revised with the assistance of QUAID, and questions revised without QUAID. In a third evaluation, eye-tracking data were collected while respondents read questions on a computer screen and answered questions aloud. Respondents had a tendency to give up processing difficult questions too early (called an early exit), which potentially threatens the validity of the respondents' answers. Survey methodologists are encouraged to use QUAID and further evaluate its validity and utility.

Questions on a survey should elicit valid and reliable answers from respondents in a short amount of time. The goals of validity, reliability, and efficiency cannot be met if respondents have trouble comprehending the questions. Survey methodologists must somehow identify questions that are difficult for respondents to comprehend. They also need some theoretical and empirical guidance for identifying what is wrong with problematic questions so they can be appropriately repaired.

In order to achieve these goals, survey methodologists have developed several methods for pretesting survey questions. Methods of pretesting include conventional pretests, formal appraisal systems, panels of experts, cognitive interviews with respondents, respondent debriefings, behavior coding, and analysis of response latencies (Presser et al. 2004). The most common method is conventional pretesting in which a small number of interviewers collect data from a small number of respondents and then gather in a debriefing session to discuss their perceptions of question problems. In a formal appraisal system, experts identify particular problems with questions that are anticipated according to an analytical coding scheme (Lessler and Forsyth 1996). A successful scheme helps survey methodologists identify problems that otherwise would be missed when they rely exclusively on their own intuition. In cognitive interviews, the respondents either provide "think aloud" protocols or verbally respond to interviewers' question probes during the process of answering the survey questions (Willis, DeMaio, and Harris-Kojetin 1999). Some of the problems with questions can be articulated by respondents, whereas others are too subtle to capture the respondents' attention or are too difficult for respondents to express verbally. Therefore, there are advantages to including behavior coding methods whereby experts systematically analyze the behavior of a small sample of respondents during the process of completing a survey. Behavioral signals that respondents are having difficulties with a question include long pauses after the question is asked, puzzled facial expressions, and requests for clarification (Fowler and Cannell 1996; Schober and Conrad 1997). Although there is a rich variety of pretesting methods, these methods have rarely been systematically compared in evaluations of reliability, validity, and cost-effectiveness with respect to time and money (Presser and Blair 1994).

A new approach for pretesting is to build a computer program that identifies problems with questions in a theoretical, principled, and systematic fashion

(Graesser et al. 2000). This approach was pursued in the present project. We developed a computer program called Question Understanding Aid (QUAID) that critiques questions on five classes of comprehension problems. QUAID attempts to automate both the *detection* and the *diagnosis* of problems, which may be missed by the alternative pretesting methodologies. The computer program implements new analytical methods of diagnosing comprehension problems with computational modules that have recently been developed in computer science, computational linguistics, discourse processing, and cognitive science.

QUAID was originally inspired by models in cognitive psychology and survey methodology that dissect different stages of question answering (Jobe and Mingay 1991; Lessler and Sirken 1985; Schwarz and Sudman 1996; Sirken et al. 1999; Sudman, Bradburn, and Schwarz 1995; Tourangeau 1984). Most of these models include the stages of question interpretation, memory retrieval, comparative judgment, and response selection. The fidelity and variability of question *interpretation* among respondents are known to be one of the serious sources of error that threaten the reliability and validity of answers to questions (Fowler and Cannell 1996; Groves 1989; Lessler and Kalsbeek 1993; Schober and Conrad 1997). Therefore, revising questions to minimize interpretation problems is one important strategy for reducing measurement error. QUAID was designed to diagnose interpretation problems but was not designed to facilitate other stages of questions answering (i.e., memory retrieval, judgment, and response selection).

The fact that QUAID both detects and diagnoses problems with questions would presumably be useful to researchers and practitioners in the survey world. When QUAID detects a problem with a question, that is a signal that the survey researcher should consider modifying the wording of the question to minimize comprehension problems. QUAID diagnoses particular problems with a question so there is some specific guidance on what needs to be changed to repair the question. It is the survey methodologist who has the burden of generating a new revised question, however. QUAID detects and diagnoses problems, but it does not automatically *revise* and *repair* the questions.

It currently is an open question as to whether QUAID provides critiques that are valid. This article reports some initial evaluations of QUAID, but more research is needed to fully evaluate whether its critiques will successfully guide survey methodologists to revise problematic questions in a fashion that improves the reliability and validity of the answers to the questions. We encourage survey researchers to use QUAID as another window for pretesting questions and to evaluate whether the revised questions are an improvement in formal evaluations.

QUAID ultimately incorporated five modules that critique questions on potential comprehension difficulties at various levels of language, discourse, and world knowledge. The critique identifies (1) words that are unfamiliar to many adults, (2) unclear relative terms (i.e., verbs, adjectives, adverbs),

(3) vague or ambiguous noun phrases, (4) questions with complex syntax, and (5) questions that overload working memory. These five classes of problems not only are computationally tractable (i.e., within the capabilities of modern computer technologies) but also were very frequent in a corpus analysis of questions in information surveys (Graesser et al. 1996; Graesser et al. 1999).¹

The remainder of this article is divided into three sections. The first section describes QUAID's Web facility and the five problems with the questions it identifies. The second section describes a series of studies that perform some empirical assessments of the validity and utility of QUAID. QUAID is valid and useful to the extent that it helps survey methodologists do a better job diagnosing problems with questions and revising problematic questions. Another test of validity is whether it accounts for the eye-tracking performance of respondents while they answer questions. The reported empirical investigations provide *prima facie* evidence of the validity and utility of QUAID's output. The third section points out some limitations with the current version of QUAID, directions for further development, and practical uses of the tool.

QUAID

QUAID runs on the Web (www.psyc.memphis.edu/quaid.html) and can be used by the public at no cost. Online registration is needed in order to use the tool. The Web version of QUAID will handle only one question at a time for the general public. We have in-house utilities for processing a set of survey questions in batch mode; these are available to researchers at a modest cost.

Figure 1 shows the main QUAID Web page, along with an example question and output for illustration. The right banner presents a list of hypertext links to help facilities. At the top is an instruction link with information on how to use the tool. Then there are links to help facilities for each of the five problems with questions. There are multiple levels of help for each problem, starting with a brief description of the problem (i.e., as in the appendix) and ending with numerous example questions with problems within a particular category. The question is submitted to QUAID through the windows at the left of figure 1.

There are three slots of information that the user can enter: the Question, the Context, and the Response Items. The Question is the main question that is being asked, whereas the Response Items (if any) are the answer options that the respondent may select. The Context slot includes sentences presented to

1. A corpus analysis was conducted on questions in information surveys (such as those collected by the U.S. Census Bureau), the 1040 income tax form, an application to graduate school at a university, a dentist intake form, and an application form for a job at Kinko's in order to determine potential problems with questions. Researchers trained in cognitive science, psycholinguistics, and discourse processes identified 12 problems with questions that occur in surveys, which are listed and briefly described in the appendix.

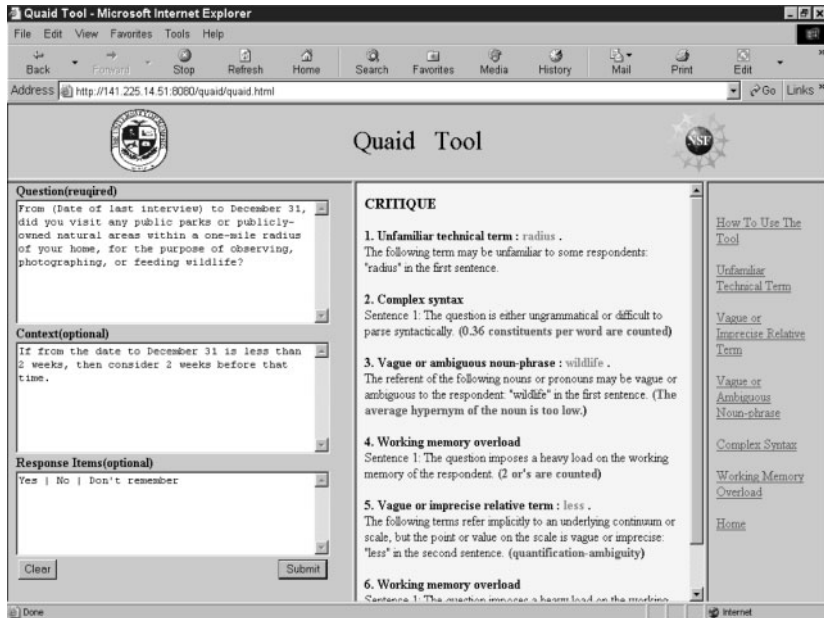


Figure 1. Interface of Web page for QUAID.

the respondent that clarify the meaning of the question and instructions on how the respondent is supposed to formulate an answer. The content of the three slots is illustrated in the example below and in figure 1:

QUESTION: From the date of the last interview to December 31, did you visit any public parks or publicly owned natural areas within a one-mile radius of your home, for the purpose of observing, photographing, of feeding wildlife?

CONTEXT: If from the date to December 31 is less than 2 weeks, then consider 2 weeks before that time.

RESPONSE ITEMS: Yes | No | Don't Remember

The user enters the question in the windows either by typing or by pasting information from other applications, such as text editors. After submitting the question and its related information, QUAID's critique appears within a few seconds in the center of the Web page. There is a list of all problems QUAID has identified and a short description of each problem. There also is a help facility that defines each problem more completely and gives examples of particular problems. The critique and help facility together allow the survey methodologist to dissect and repair the problem with a particular question.

QUAID's critique uncovered a number of potential problems with the above example. In fact, this example question exhibits all five of the problems

with questions that QUAID attempts to spot. We see that *radius* is an unfamiliar technical term, *less* is a vague or imprecise relative term, and *wildlife* is a vague or ambiguous noun phrase. The syntax is complex because QUAID states that *0.36 constituents per word are counted*; this means that there is a large density of structural embeddings in the sentence composition of the Question slot. The precise meaning of this parameter can be reconstructed from the help facilities. The question also imposes a large load on working memory because *two or's are counted*. Sentences with two or more *or's* tax working memory because the respondent needs to keep track of a mental table of options that potentially has many cells. When crossing public parks versus natural areas with observing, versus photographing, versus feeding wildlife, there are six cells ($2 \times 3 = 6$) that need to be considered. The number of cells quickly becomes unmanageable when there are more dimensions and more alternative values per dimension.

It is beyond the scope of this article to provide the technical details of how QUAID identifies problems. These can be reconstructed, to some extent, from the subsequent description of the algorithms and the help facilities on the Web site. As an example, unfamiliar technical terms are identified by accessing electronic lexicons that specify the frequency of words in the English language. If the frequency for a word is too low (i.e., below some threshold), then the word is flagged as a potential technical term. However, it is the survey methodologist who ultimately decides whether it is likely to be unfamiliar to the target population of respondents.

In the case of syntax, there are metrics that compute the number of constituents at the top level of a syntactic parse tree (i.e., which is similar to the diagrams of sentence structures that most teachers cover in grade school classes), the number of subordinate clauses, the number of relative clauses, and the number of modifiers of nouns in noun phrases. Jurafsky and Martin (2000) is an excellent source that covers the recent developments in computational linguistics and natural language processing in artificial intelligence. During the process of designing and developing QUAID, we used correlational analyses to explore which of the alternative measures of syntactic complexity best predicts expert ratings of syntactic complexity, as reported in previous publications (Graesser et al. 2000; Graesser et al. 2001).²

Below is a succinct description of the five problems with questions and how we computed the problems algorithmically:

2. Graesser et al. (2000) conducted a study that collected ratings from three experts who were extensively trained on problems with questions and had a deep understanding of cognitive science, psycholinguistics, and discourse processes. They evaluated a corpus of 550 questions on the first nine of the 12 problems in the appendix. The following rating scale was used in making these judgments for each potential problem with a question: 1 = *definitely not a problem*, 2 = *probably not a problem*, 3 = *probably a problem*, and 4 = *definitely a problem*. These ratings were correlated with the components and parameters of QUAID during the design and development of the tool.

1. *Unfamiliar technical term.* There is a word or expression for which few respondents would know the meaning. A word is tagged as unfamiliar if it is not found in an available electronic lexicon or is below the threshold in the following lexicons with frequency or familiarity metrics: Francis and Kucera's (1982) written frequency, Thorndike and Lorge's (1944) written frequency, Brown's (1984) verbal frequency, and the familiarity rating in the Coltheart's (1981) MRC Psycholinguistics Database. Each threshold is determined by varying candidate thresholds and identifying the value that maximizes correlations with expert ratings of questions as to whether they contain unfamiliar technical terms.
2. *Vague or imprecise predicate or relative term.* The values of a predicate (e.g., quantitative adjective or adverb) are not specified on a functional scale or underlying quantitative continuum (e.g., *many*, *few*, *rarely*, *frequently*). For example, when the word *frequently* is used on a survey question, how often does an event need to occur before it should count as being "frequent"? Obviously, this depends on what the event is; ten would be frequent for the number of traffic accidents that a person experiences per year but would not be frequent for the number of visits to a fast food restaurant per year. QUAID simply identifies such relative terms that are potentially vague or imprecise, whereas the survey methodologist makes the final judgment as to whether such terms would be problematic for most respondents. QUAID has a data file with lists of words that flag potential problems with relative terms. Quantitative adjectives and adverbs are a closed class of words (i.e., a relatively small finite number that can be listed), whereas quantitative verbs (e.g., *increase*) are an open class and cannot be exhaustively listed. The current version of QUAID does a much better job on the quantitative adjectives and adverbs than on verbs.
3. *Vague or ambiguous noun phrase.* The referent of a noun phrase, noun, or pronoun is unclear or ambiguous (e.g., *items*, *amount*, *it*, *there*). A word is flagged as problematic if one or more of the following conditions is met: (a) the hypernym value is less than a threshold when its entry is looked up in an electronic lexicon called WordNet (Miller et al. 1990), (b) the polysemy value of a head noun is greater than a threshold when its entry is looked up in WordNet, (c) the concreteness value of a noun is less than a threshold when its entry is looked up in the MRC database, or (d) it is in a list of vague terms. The hypernym value is an index of a word's abstractness according to WordNet's theoretical taxonomic hierarchy (e.g., Honda → automobile → vehicle → device → object → entity). A low hypernym value for a word means that it is very superordinate in the hierarchy, relatively abstract, and vulnerable to vagueness or ambiguity. The polysemy value is the number of different senses of a word. For example, the word *bank* has at least two

senses, such as an edge of a river and a location that handles money. Words that have many senses are more vulnerable to ambiguity. According to QUAID, concrete words are less vulnerable to ambiguity and vagueness than abstract terms are. The concreteness norms of the MRC database were therefore expected to predict experts' judgments of questions with problematic noun phrases, and indeed they did. As with all metrics in QUAID, a threshold was determined by varying candidate thresholds and identifying the value that maximizes correlations with expert ratings of questions as to whether they contain a particular problem (in this case, vague or ambiguous noun phrases).

4. *Complex syntax.* The grammatical composition is embedded, dense, structurally ambiguous, or not well formed syntactically. Syntactic parsers are used in conjunction with Brill's (1995) part-of-speech classifier when syntactic analyses are performed. A sentence is flagged as having complex syntax if one of the following conditions is met: (a) the number of words before the main verb of the main clause is greater than a threshold, (b) the number of modifiers of a noun in a noun phrase is greater than a threshold, or (c) the mean number of higher-level constituents per word is greater than a threshold. A sentence with many higher-level constituents per word is structurally complex, has many levels of structure, or is not syntactically well formed.
5. *Working memory overload.* Words, phrases, or clauses impose a high load on immediate memory. A sentence is flagged if one or more of the following conditions is met: (a) the number of higher-level constituents per word is greater than a threshold, (b) the number of conjunctions is greater than a threshold, or (c) a threshold is exceeded on the number of words that signify logical operations (*if, or, and, no, not*). These characteristics of questions require the respondent to keep in mind a large number of words, constituents, linguistic levels, variables, values, or alternative possibilities.

Evaluations of QUAID's Validity and Utility

This section reports some empirical studies that assess the validity and utility of QUAID's critiques of questions. In one set of studies, expert survey methodologists critiqued and revised problematic questions. QUAID's validity would be confirmed if its output included the critiques provided by experts. However, the agreement might not end up being particularly high if QUAID can diagnose important problems that are missed by experts. A low agreement between experts and QUAID would suggest that QUAID simply offers another view on question problemhood, unless there were some grounding in a defensible gold standard (such as response errors).

In another study, survey methodologists evaluated the quality of original problematic questions, questions revised by a previous sample of survey methodologists with the assistance of QUAID, and questions revised without QUAID. The validity and utility of QUAID would be supported if the revised questions received higher evaluations when the survey methodologists had used QUAID, compared with their revisions without QUAID. Moreover the revised questions with access to QUAID should be judged as better than the original questions.

In the final study, we collected eye movements and fixations while respondents answered questions aloud. We expected different eye-tracking behavior for questions with problems than for those without problems. The critiques of QUAID would be validated if there are interpretable differences between problematic and unproblematic questions according to QUAID.

EXPERT SURVEY METHODOLOGISTS REVISE AND CRITIQUE PROBLEMATIC QUESTIONS

If QUAID is a useful tool, then it should identify important problems with questions that are sometimes missed by survey methodologists. Consequently, survey methodologists should end up critiquing and revising questions more effectively with the assistance of the tool than without it. In this study, 12 expert survey methodologists revised and critiqued a sample of problematic questions either with the assistance of QUAID or without the assistance of QUAID. The expert survey methodologists were volunteers from the U.S. Census Bureau, the Bureau of Labor Statistics, the General Accounting Office, and the National Center for Health Statistics. Sixty of the most problematic questions from the 550 U.S. Census questions investigated by Graesser et al. (2000) were randomly divided into three groups of items (sets A, B, and C). The experts first revised and critiqued the questions in one of the sets without the use of QUAID. They subsequently revised and critiqued the questions in another set with the assistance of the QUAID tool. The critique consisted of a list of descriptions about what is wrong with each question. There was a space for the expert to revise the question. Each expert was offered \$100 for participating in the experiment, but most of the experts declined the payment and participated for intellectual curiosity rather than financial incentives.

As an example, consider the following question from one of the surveys that suffers from a number of problems: “At any time during the last 12 months, were you or any member of your household enrolled in or receiving benefits from free or reduced-priced meals at school through the Federal School Lunch program or the Federal School Breakfast program?” According to QUAID, this question has four out of the five problems. *Enrolled* is counted as an unfamiliar technical term, whereas *household* is flagged as a vague or ambiguous noun phrase (i.e., a low hypernym count). These words would

probably not be viewed as problematic according to most survey methodologists, but they are nevertheless detected by QUAID. The other two problems are more obvious and presumably would lead to a revision by survey methodologists. There is complex syntax because 16 words precede the main verb of the main clause. This problem routinely occurs for sentence constructions that have subordinate clauses preceding the main clause and other forms of left-embedded syntax. There is working memory overload because the sentence has many conjunctions (particularly *or*), so the respondent needs to keep track of a mental matrix with many combinations. The survey methodologist would hopefully identify these problems and revise the question accordingly. For example, the revision below would handle the problems of syntactic complexity and working memory overload: "Some families participate in a Federal School Lunch program or a Federal School Breakfast program. These programs provide free or reduced-priced meals at schools. Did you or any member of your household receive these benefits during the last 12 months?"

Two research assistants analyzed the critiques of the survey methodologists by scoring whether the five problems with questions were identified. That is, for each question critiqued by a particular expert, the researcher decided whether a particular problem, *P*, was or was not articulated in those protocols in which there was no QUAID tool to use. The two researchers had moderate to good reliability in making these judgments; the kappa's are .60, .58, .69, .54, and .56 for problems 1 through 5, respectively. It is informative to note that the expert survey methodologists did not identify most of the problems that QUAID identified. The probabilities that the experts identified a problem, given that QUAID did, were .10, .11, .46, .47, and .37, respectively. The experts did a much better job identifying vague or ambiguous noun phrases, complex syntax, and working memory load than they did identifying unfamiliar terms and vague/imprecise relative terms. The fact that the expert methodologists did not identify most of the problems that QUAID identified supports the general claim that QUAID uncovers some new problems that are not in the radar of expert survey methodologists.

Graesser et al. (1999) report that college student respondents are also not very good at identifying problems with questions. A sample of 24 college students at the University of Memphis provided written critiques of 38 problematic questions that were extracted from a census form, a dentist intake form, an application to graduate school, and a Kinko's job application. The questions had at least one of the problems in the appendix according to experts. The college students were instructed to write down two problems with each question. The results reveal that untrained literate adults are very poor at identifying the 12 problems listed in the appendix. The only two problem categories that they could identify in any discriminating fashion are unclear technical terms (category 1) and vague or ambiguous noun phrases (category 3). The other categories were not in their radar. This result supports the utility of QUAID over and above the use of think-aloud protocols and other forms of pretesting from

samples of respondents. The thoughts and critiques of questions by a sample of respondents will miss problems with questions that are identified by QUAID. Of course, we are assuming here that the additional problems identified by QUAID are valid problems that present comprehension difficulties.

EXPERT SURVEY METHODOLOGISTS SELECT BETWEEN ORIGINAL AND REVISED QUESTIONS

If QUAID is a useful tool, then the revisions that expert survey methodologists produce should be perceived as better when they have QUAID at their disposal. In this study, another sample of 12 expert survey methodologists completed a two-alternative forced-choice (2AFC) test in which pairs of questions were presented and they judged which of the two questions would be easier to comprehend for most respondents. We paired original questions with questions revised by survey methodologists with the assistance of QUAID (called *SM + QUAID revisions*). Similarly, we paired original questions with questions revised by survey methodologists who had no access to QUAID (called *SM-alone revisions*). The revised questions were randomly sampled from the previous experiment in which experts revised problematic questions. Random guessing in the 2AFC test would produce a score of .50. A preference score measured the extent to which their choice of the revised question (R) was above .50, computed as $[(R - .50)/(1 - .50)]$. A preference score of 0 constitutes chance responding, whereas a preference score of 1 signifies that the revised question is always preferred over the original question. The preference score for SM + QUAID revisions is .30, which is significantly above 0 (with 95 percent confidence intervals of $\pm .24$). The preference score for SM-alone revisions is .17, which also is significantly above 0 (with 95 percent confidence intervals of $\pm .10$). The difference in preference scores for SM + QUAID revisions versus SM-alone revisions is only significant at a liberal one-tailed test at a reduced alpha, $t(11) = 1.77$, $p = .10$. However, the effect size of the SM + QUAID condition compared to the control is a respectable 1.3 standard deviation units. Although the results could be more robust statistically, they are consistent with the claim that survey methodologists will improve the comprehensibility of problematic questions more with the QUAID tool than without the QUAID tool.

The above experiment was replicated except that college students completed the 2AFC test instead of the expert survey methodologists. The preference scores for college students are not significantly different for SM + QUAID question revisions and SM-alone question revisions, .13 versus .18, respectively. This difference is in the wrong direction and not statistically significant. This result is consistent with the claim that respondents have difficulty judging question quality and problems in comprehension (see also Graesser et al. 1996; Graesser et al. 1999). Expertise in survey methodology or in cognitive science, discourse, and language is needed to make reliable

judgments on the difficulty of interpreting questions (Presser et al. 2004). 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 (Presser and Blair 1994).

EYE TRACKING WHILE ANSWERING QUESTIONS

The previous studies support the general claim that QUAID detects many of the subtle aspects of language, discourse, and world knowledge that are not detected by experts and respondents off-line. In this final study, we conducted an eye-tracking experiment to determine whether readers can detect these subtleties online. The collection of eye-tracking data provides a different method of diagnosing problematic questions with respect to question interpretation. Eye-tracking patterns presumably serve as a sensitive index of online comprehension processes. If a question is difficult to comprehend, then there should be multiple fixations on words and regressive eye movements. Words that are difficult to interpret should have longer total fixation times than words that are easy to interpret. We collected eye-tracking data in order to assess whether the problems identified by QUAID are manifested in eye movements and gaze durations.

We adopted some working assumptions about eye movements in the present study. One assumption is that there is a sufficiently close connection between the display item viewed and the content being thought about, as well as between the time fixating on display items and the amount of cognitive processing (Just and Carpenter 1980; Rayner 1998). A second assumption is that eye movements provide an important window for dissecting the cognitive processes in tasks that require deeper levels of processing, which would include the answering of survey questions. Eye-tracking analyses have indeed provided an illuminating method of investigating problem solving (Grant and Spivey 2003; Knoblich, Ohlsson, and Raney 2001), reasoning (Graesser et al. 2005; Just and Carpenter 1992), and the comprehension of stories, informational texts, and advertisements (Just and Carpenter 1980; O'Brien et al. 1997; Rayner et al. 2001).

In the eye-tracking experiment we conducted, college students read and answered questions selected from the corpus of 550 survey questions.³ There were three sets of questions. One set was run only on college students who were hunters because there were questions from surveys that presupposed the

3. The eye-tracking equipment was an Applied Science Laboratory Model 501 eye tracker with a head-mounted device. A head-mounted apparatus was used so that participants could move their heads and speak during data collection. Participants were calibrated both before the experimental session began and throughout the session to ensure reliable data. During calibration, participants viewed nine points on a $1,024 \times 768$ computer monitor, and the eye tracker recorded corresponding x-y coordinates. The temporal resolution of the eye tracker was 60 Hz. The spatial resolution was a .50 degree angle horizontally and a .40 degree angle vertically.

respondent was a hunter. There were three surveys in set A. The questions students read and answered included a sample of problematic and unproblematic questions in the order that they appeared in the following surveys: Adolescent Self-Administered Questionnaire (ten questions), Hunting and Fishing Questionnaire (17 questions), and American Community Survey (five questions). A question was defined as problematic if it contained at least one problem according to QUAID. The sample administered the second set (B) included college students who were parents because some of the questions presupposed the respondent was a parent. Set B included the National Health Survey–Child Prevention Module (seven questions), the 1998 National Health Interview Survey Basic Module–Adult Core (15 questions), and the American Community Survey (five questions). A third set (C) could accommodate any college student as a participant and included the Nonconsumptive User’s Questionnaire (eight questions), the Adult Core IV (nine questions), the American Community Survey (five questions), and the U.S. Census 2000 Dress Rehearsal (five questions). The numbers of participants who completed sets A, B, and C are 7, 9, and 9, respectively. After the eye-tracking data were collected, the participants completed a Wechsler Abbreviated Scale of Intelligence (Psychological Corporation 1999) and an information sheet about demographic information and university training.

During each trial, the participant advanced to the next question by hitting a space bar. Then the question appeared on the screen. The participant read the question and answered the question aloud. We recorded the eye-tracking data while the students read the question, audio recorded their answers, and videotaped the computer screen.

Our analysis of the eye-tracking data turned out to be an intense data-mining exercise because of the complex patterns of eye movements that were manifested. We began with the simple but incorrect assumption that readers would linearly read through the question one word at a time from left to right and top to bottom. This is the pattern that most readers tend to manifest, with notable exceptions, in the case of other printed text (Just and Carpenter 1980; Rayner 1998). This was not a correct assumption in our analysis of questions, however. Respondents would frequently look ahead at the response options and would go back and forth among the focal question, context, and response options. These complex strategic activities made it unproductive to look at the standard measures of processing difficulty, such as first fixation durations on a word, number of fixations on a word, total reading time per word, percentage of regressions, and skipping likelihood.

We did find some support for QUAID’s detection of unfamiliar technical terms (problem 1). The content words (i.e., nouns, main verbs, adjectives) that are unfamiliar technical terms according to QUAID had longer first fixation durations, longer total reading times, more fixations, and lower skip rates than did content words that are not technical terms. These word frequency effects are well known in the literature on reading from text (Just and Carpenter 1980;

Rayner 1998). This unsurprising finding is perhaps valuable for survey methodologists because it confirms the intuition that technical terms are often manifested in eye-movement behavior. Respondents who fail to spend more time on these technical terms may have difficulties comprehending the question. Questions with several technical terms will run the risk of being misinterpreted or adding a significant amount of processing time that some respondents may not invest.

We originally had hoped that eye tracking would help us detect words that are vague or imprecise relative terms (problem 2) or are vague or ambiguous noun phrases (problem 3). There are a few patterns in the eye-tracking data that have distinctive signatures for these two problems, but the effects are extremely subtle, often not statistically significant, and in need of replication.

Another plausible but incorrect assumption about eye tracking is that difficult questions would require more processing time. In fact, the data reveal that there is a near-zero correlation between the mean reading time per word and the degree to which a question is problematic (e.g., number of problems or presence/absence of a problem). For example, the correlations between the average total reading time per content word in a question and the presence/absence of a problem are .00, -.05, -.06, -.11, and -.05 for problems 1, 2, 3, 4, and 5, respectively. The corresponding correlations with number of fixations per content word are .00, -.08, -.07, -.10, and -.05. If anything, the participants tended to speed up for the difficult questions, given that the correlations are all in the negative direction. One possible explanation of the low correlations is that our eye-tracking equipment did not have the resolution and calibration accuracy to measure word processing precisely. This explanation can be ruled out by the fact that there is a moderate positive correlation between the word length (number of letters) of content words and total reading time per content word ($r = .27$), number of fixations ($r = .39$), and first fixation times ($r = .42$). A second possible explanation of the counterintuitive result (i.e., near-zero negative correlations between processing time and question problems) is that respondents more efficiently recruit their cognitive resources when the difficulty of the question increases. There is a third possibility that is particularly intriguing, however. Readers may tend to give up processing a question very deeply when the question is difficult. That is, they essentially adopt an “early exit strategy” for problematic questions.

We did observe evidence for the early exit strategy in our analyses of complex syntax (problem 4) and working memory overload (problem 5). Consider first the complex syntax. The questions were segregated into items that do and those that do not have complex syntax according to QUAID; there were 55 questions with complex syntax and 21 without complex syntax. For each item, we identified the point in the item that corresponds to the last word in the focal question and scored that word as point 0. Words before point 0 were scored as negative, whereas words after point 0 (i.e., the answer options) were scored as positive. We then analyzed (a) the furthest point in the item on which there

was at least one eye fixation and (b) the point in the sentence when the answer was initially articulated by the participant (i.e., the voice onset). If a participant completed the interpretation of the sentence before the question was answered, then the values of both (a) and (b) would be positive. That is, they would read the question and subsequently answer the question; these are called “full question interpretations.” In contrast, errors in interpretation and the validity of the answers would presumably occur when the participants did not finish reading the question before they answered it. Sometimes they would answer the question before the final word in the question was reached and would never read further; these are called “early exits.” At other times they would answer the question before the final word in the question was reached but would read on to the end of the question during or after the articulation of the answer; these are called “early articulations.” There is another category called “simultaneous” when the onset of the answer is precisely during the final word of the focal question is read. Valid answers would be expected in the case of full question interpretations and simultaneous observations but not in the early exits and early articulations.

Table 1 shows the proportion of observations that were classified as full question interpretations, simultaneous, early articulations, and early exits. These proportions were computed for each of the 76 items. The unit of analysis is the question in statistical analyses. We performed an analysis of variance with a factorial design that crossed syntactic complexity (two levels) and processing status (three levels: full interpretation, early articulation, early exit). The simultaneous category was excluded because these are proportions (which add to 1.0 for the four categories) and one category needs to be excluded because of degrees of freedom. There is a statistically significant interaction between syntactic complexity and processing status, $F(2, 148) = 3.41, p < .05$. When the syntax of the question is complex, there were more early articulations and early exits and fewer full question interpretations. Moreover, in a follow-up analysis we segregated participants with high versus

Table 1. Reading of the Question Item and Onset of Answer Articulation: Syntactic Complexity and Working Memory Load

| | Syntax | | Working Memory | |
|--------------------|-------------------|----------------|----------------------------|-------------------------|
| | No Complex Syntax | Complex Syntax | No Working Memory Overload | Working Memory Overload |
| Full Question | | | | |
| Interpretation | .63 | .48 | .59 | .46 |
| Simultaneous | .06 | .07 | .05 | .07 |
| Early Articulation | .23 | .32 | .27 | .32 |
| Early Exit | .08 | .13 | .09 | .15 |

low verbal ability (according to the psychometric test that was administered) and found that this trend was most pronounced for participants with comparatively lower verbal ability. These results are consistent with the conclusion that participants tend to give up when the question is difficult.

The same set of analyses is presented for working memory load, as shown in table 1. There were 41 items with high working memory load and 35 with low, according to QUAID. The statistical analyses support the same conclusions as the above interpretations for complex syntax.

The results of the eye-tracking analyses suggest that we need a more complex cognitive model of how participants answer survey questions. The model will need to predict the point in the sentence when the processing load becomes so difficult that the participant either (a) gives up, (b) processes the question more quickly, or (c) more diligently spends time and effort to construct an adequate interpretation. The model will no doubt need to be sensitive to individual differences among respondents.

There is at least one practical implication of these analyses on early exits. When surveys are conducted, there may be circumstances when respondents should be encouraged to read the entire question before answering it. This normally happens when the survey is given orally, but there is no assurance of this when the survey is completed on the Web or by paper and pencil. Web facilities have the potential to provide better control. The question context and focal question can remain on the computer display before the question is finished being comprehended and the respondent presses a button. Then the answer options would appear, and the answers could be collected. This more controlled procedure with Web technologies would presumably yield more valid responses, but that is an open question for future research.

Conclusions and Implications

QUAID is computer tool that potentially can help survey methodologists design, modify, and evaluate questions on comprehension difficulty. It provides another pretesting methodology at the researcher's disposal in addition to conventional pretests, formal appraisal systems, panels of experts, cognitive interviews with respondents, respondent debriefings, behavior coding, and analysis of response latencies (Presser et al. 2004). QUAID *identifies* problematic questions and specific properties of the questions that are problematic but does not *revise* the questions. Survey methodologists must (a) decide which problems identified by QUAID are likely to be problematic for a given population of respondents, (b) revise the questions that have particular problems, and (c) analyze the questions on other stages of processing, such as memory retrieval, judgment, and response selection.

QUAID's analysis of interpretation problems offers information that goes beyond most of the alternative methodologies. Many potential problems

uncovered by QUAID are not noted by survey methodologists, respondents, and experts in language, discourse, and cognition. For example, 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. The use of QUAID as a training tool for survey methodologists may be a sensible direction.

The accumulation of evidence suggests that QUAID's output has a modicum of validity. Our evaluations reveal that survey methodologists revise questions with QUAID better than they do without QUAID, at least according to judgments of question quality by expert survey methodologists. Our analyses of eye tracking reveal that questions identified as problematic by QUAID were processed differently than the nonproblematic questions. These findings lend support to the validity and utility of QUAID, but additional rigorous assessments are needed in future research. Experiments need to be conducted that assess whether response error rates on surveys are significantly lower in versions of surveys in which the questions are modified by survey methodologists with the assistance of QUAID. Revisions of questions with the assistance QUAID need to be systematically compared with alternative pretesting methodologies in formal tests of response error. We encourage survey methodologists to use QUAID, which is free to the public on the Web, and to perform systematic assessments that uncover its strengths, its weaknesses, and avenues for improvement.

The patterns of eye tracking reveal that question comprehension is guided by somewhat complex top-down processes. It is not the case that readers routinely read questions in a linear fashion one word at a time, left to right, one line at a time. Instead, they hop around between the question and response alternatives. More time is spent in the upper left quadrant than the lower right quadrant (Graesser et al. 2002). When the question has complex syntax or a high working memory load, readers in our studies tended to take an early exit from reading the question and generate an answer. In essence, they answered the question before they finished reading it. Such early exits could potentially threaten the validity of answers when surveys are completed by paper and pencil. A Web facility that forces the respondent to read the question before presenting the answer alternatives would presumably mitigate this problem. However, we acknowledge that there are conditions when it is preferable to present answer alternatives before presenting the question. One direction for future research is to model eye-tracking mechanisms during question comprehension and to investigate the processing of different question wordings and item layouts.

Future versions of QUAID could be enhanced in a number of ways. The accuracy of the five problems could be improved with more research and development in computational linguistics and corpus linguistics. An analysis of thousands of surveys would allow us to fine-tune virtually all of QUAID's components. QUAID could be enhanced by incorporating special-purpose

lexicons and corpora that are tailored to the population of respondents. For example, technical terms associated with financial jargon would be quite different on a survey targeted for accountants as opposed to the general population. QUAID could be enhanced by incorporating additional classes of problems that are summarized in the appendix. In particular, the identification of problematic presuppositions (category 6) and the detection of unclear question purpose (category 9) are good candidates for future efforts. This will require additional computational components that assess world knowledge. The current version of QUAID is only the beginning of our attempt to offer survey methodologists more sophisticated, intelligent technologies for doing their work.

Appendix: Problems with Questions

1. *Unfamiliar technical term.* There is a word or expression of which very few respondents would know the meaning.
2. *Vague or imprecise relative term.* The values of a predicate (i.e., main verb, adjective, or adverb) are not specified on a scale that allows comparisons along a continuum.
3. *Vague or ambiguous noun phrase.* The referent of a noun phrase, noun, or pronoun is unclear or ambiguous.
4. *Complex syntax.* The grammatical composition is embedded, dense, or structurally ambiguous.
5. *Working memory overload.* Words, phrases, or clauses impose a high load on immediate memory.
6. *Misleading or incorrect presupposition.* The truth-value of a presupposed proposition is false or inapplicable.
7. *Unclear question category.* It is difficult to determine what class of question is being asked.
8. *Amalgamation of more than one question category.* The question may be assigned to two or more different classes of questions.
9. *Unclear question purpose.* The respondent would not know why the question is being asked.
10. *Mismatch between question category and answer option.* The question invites one set of answer options that is different from the response options in the questionnaire.
11. *Difficult to access specific or generic knowledge.* A typical respondent would have difficulty recalling the information requested in the question.
12. *Respondent unlikely to know answer (no information source).* A typical respondent would not know the information requested in the question.

References

- Brill, Eric. 1995. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging." *Computational Linguistics* 21:543–66.
- Brown, G. D. A. 1984. "A Frequency Count of 190,000 Words in the London-Lund Corpus of English Conversation." *Behavioral Research Methods Instrumentation and Computers* 16:502–32.

- Coltheart, Max. 1981. "The MRC Psycholinguistic Database." *Quarterly Journal of Experimental Psychology* 33A:497–505.
- Fowler, Floyd J., and Charles F. Cannell. 1996. "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, ed. Norbert Schwarz and Seymore Sudman, pp. 15–36. San Francisco: Jossey-Bass.
- Forbes, W. Nelson, and Henry Kucera. 1982. *Frequency Analysis of English Usage*. Boston: Houghton-Mifflin.
- Graesser, Arthur C., Sailaja Bommareddy, Shane Swamer, and Jonathan M. Golding. 1996. "Integrating Questionnaire Design with a Cognitive Computational Model of Human Question Answering." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, ed. Norbert Schwarz and Seymore Sudman, pp. 143–74. San Francisco: Jossey-Bass.
- Graesser, Arthur C., Ashish B. Karnavat, Frances K. Daniel, Elisa Cooper, Shannon N. Whitten, and Max M. Louwerse. 2001. "A Computer Tool to Improve Questionnaire Design." In *Statistical Policy Working Paper 33, Federal Committee on Statistical Methodology*, pp. 36–48. Washington, DC: Bureau of Labor Statistics.
- . 2002. "QUAID: A Computer Tool to Improve Questionnaire Design." Paper presented at the meetings of the American Association of Public Opinion Research, St. Petersburg, FL.
- Graesser, Arthur C., Tina Kennedy, Peter Wiemer-Hastings, and Victor Ottati. 1999. "The Use of Computational Cognitive Models to Improve Questions on Surveys and Questionnaires." In *Cognition and Survey Methods Research*, ed. Monroe G. Sirken, Douglas J. Hermann, Susan Schechter, Norbert Schwarz, Judith M. Tanur, and Roger Tourangeau, pp. 199–216. New York: Wiley.
- Graesser, Arthur C., Shulan Lu, Brent A. Olde, Elisa Cooper-Pye, and Shannon Whitten. 2005. "Question Asking and Eye Tracking during Cognitive Disequilibrium: Comprehending Illustrated Texts on Devices when the Devices Break Down." *Memory and Cognition* 33:1235–47.
- Graesser, Arthur C., Katja Wiemer-Hastings, Roger Kreuz, Peter Wiemer-Hastings, and Kent Marquis. 2000. "QUAID: A Questionnaire Evaluation Aid for Survey Methodologists." *Behavior Research Methods, Instruments, and Computers* 32:254–62.
- Grant, Elizabeth R., and Michael J. Spivey. 2003. "Eye Movements and Problem Solving: Guiding Attention Guides Thought." *Psychological Science* 14:462–66.
- Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.
- Jobe, Jared B., and David J. Mingay. 1991. "Cognition and Survey Measurement: History and Overview." *Applied Cognitive Psychology* 5:175–92.
- Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice.
- Just, Marcel A., and Patricia A. Carpenter. 1980. "A Theory of Reading: From Eye Fixations to Comprehension." *Psychological Review* 87:329–54.
- . 1992. "A Capacity Theory of Comprehension: Individual Differences in Working Memory." *Psychological Review* 99:122–49.
- Knoblich, Gunther, Stellan Ohlsson, and Gary E. Raney. 2001. "An Eye Movement Study of Insight Problem Solving." *Memory and Cognition* 29:1000–9.
- Lessler, Judith T., and Barbara H. Forsyth. 1996. "A Coding System for Appraising Questionnaires." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, ed. Norbert Schwarz and Seymore Sudman, pp. 259–92. San Francisco: Jossey-Bass.
- Lessler, Judith T., and William D. Kalsbeek. 1993. *Nonsampling Error in Surveys*. New York: Wiley.
- Lessler, Judith T., and Monroe G. Sirken. 1985. "Laboratory-Based Research on the Cognitive Aspects of Survey Methodology: The Goals and Methods of the National Center for Health Statistics Study." *Milbank Memorial Fund Quarterly/Health and Society* 63:565–81.
- Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. "Introduction to WordNet: An On-Line Lexical Database." *International Journal of Lexicography* 3:235–44.
- O'Brien, Ed J., Gary E. Raney, Jason E. Albrecht, and Keith Rayner. 1997. "Processes Involved in the Resolution of Anaphors." *Discourse Processes* 23:1–24.

- Presser, Stanley, and Johnny Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology* 24:73–104.
- Presser, Stanley, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Elinor Singer. 2004. "Methods for Testing and Evaluating Survey Testing." *Public Opinion Quarterly* 68:109–30.
- Psychological Corporation. 1999. *Wechsler Abbreviated Scale of Intelligence*. San Antonio: Harcourt Brace.
- Rayner, Keith. 1998. "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological Bulletin* 124:372–422.
- Rayner, Keith, Caren M. Rotello, Andrew J. Stewart, Jessica Keir, and Susan A. Duffy. 2001. "Integrating Text and Pictorial Information: Eye Movements when Looking at Print Advertisements." *Journal of Experimental Psychology: Applied* 7:219–26.
- Schober, Michael F., and Frederick G. Conrad. 1997. "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly* 60:576–602.
- Schwarz, Norbert, and Seymour Sudman, eds. 1996. *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco: Jossey-Bass.
- Sirken, Monroe G., Douglas J. Hermann, Susan Schechter, Norbert Schwarz, Judith M. Tanur, and Roger Tourangeau, eds. 1999. *Cognition and Survey Methods Research*. New York: Wiley.
- Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1995. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.
- Thorndike, Edward L., and Irving Lorge. 1944. *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.
- Tourangeau, Roger. 1984. "Cognitive Sciences and Survey Methods." In *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, ed. Thomas J. Jabine, Miron L. Straf, Judith M. Tanur, and Roger Tourangeau, pp. 73–100. Washington, DC: National Academy of Sciences.
- Willis, Gordon B., Theresa J. DeMaio, and Brian Harris-Kojetin. 1999. "Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques." In *Cognition and Survey Methods Research*, ed. Monroe G. Sirken, Douglas J. Hermann, Susan Schechter, Norbert Schwarz, Judith M. Tanur, and Roger Tourangeau, pp. 133–53. New York: Wiley.