Behavior Research Methods, Instruments, & Computers
2000, 32 (2), 254-262

QUAID: A questionnaire evaluation aid for survey methodologists

ARTHUR C. GRAESSER, KATJA WIEMER-HASTINGS,
ROGER KREUZ, and PETER WIEMER-HASTINGS
University of Memphis, Memphis, Tennessee

and

KENT MARQUIS
United States Census Bureau, Washington, D.C.

QUAID (question-understanding aid) is a software tool that assists survey methodologists, social scientists, and designers of questionnaires in improving the wording, syntax, and semantics of questions. The tool identifies potential problems that respondents might have in comprehending the meaning of questions on questionnaires. These problems can be scrutinized by researchers when they revise questions to improve question comprehension and, thereby, enhance the reliability and validity of answers. QUAID was designed to identify nine classes of problems, but only five of these problems are addressed in this article: unfamiliar technical term, vague or imprecise relative term, vague or ambiguous noun phrase, complex syntax, and working memory overload. We compared the output of QUAID with ratings of language experts who evaluated a corpus of questions on the five classes of problems. The corpus consisted of 505 questions on 11 surveys developed by the U.S. Census Bureau. Analyses of hit rates, false alarm rates, d' scores, recall scores, and precision scores revealed that QUAID was able to identify these five problems with questions, although improvements in QUAID's performance are anticipated in future research and development.

A good survey or questionnaire contains questions that elicit valid and reliable answers from respondents in a short amount of time. One of the challenges to survey researchers and social scientists is to design questions that achieve these general objectives. Researchers in the field of survey methodology have proposed models that dissect the many stages of question answering (Cannell, Miller, & Oksenberg, 1981; Schwarz & Sudman, 1996; Sudman, Bradburn, & Schwarz, 1995; Tourangeau, 1984), such as question interpretation, memory retrieval, judgment, and response selection. The fidelity and variability of question *interpretation* among respondents is known to be one of the serious sources of error that threaten the reliability and validity of answers to questions (Fowler & Cannell, 1996; Groves, 1989; Lessler & Kalsbeek, 1993; Schober & Conrad, 1997). This is indeed one of the basic truths that has been established in the field known as CASM, the cognitive aspects of survey methodology (Jobe & Mingay, 1991; Lessler & Sirken, 1985; Sirken & Fuchsberg, 1984; Sirken et al., 1999). In essence, if the respondent misinterprets the question, the respondent will virtually never provide a valid answer to the question. Therefore, revising questions to minimize interpretation problems is one important strategy for reducing measurement error.

The computer tool investigated in this research focuses on the interpretation of questions, as opposed to other components of the question-answering process. QUAID (which stands for question-understanding aid) has particular modules that perform a critique of each question for potential comprehension difficulties at various levels of language, discourse, and world knowledge. For example, the critique identifies words that are unfamiliar to most respondents, vague predicates (verbs, adjectives, or adverbs), ambiguous noun phrases, questions with complex syntax, and questions that overload working memory (WM). The *identification* of these problems by the computer will be useful to the extent that they are problems that end up being missed by survey methodologists because of fatigue or training deficits. The computer aid would be even more useful if it offered suggestions about the *revision* of problematic questions, but question revision is beyond the scope of QUAID.

It is overly optimistic to expect a computer to perfectly comprehend questions at all levels of language, discourse, and world knowledge. During the last 10 years, the Department of Defense has evaluated the best computer models of information extraction in the fields of artificial intelligence, computational linguistics, and cognitive science

This research was partially funded by grants from the U.S. Census Bureau (43-YA-BC-802930) and the National Science Foundation (SBR 9720314 and SBR 9977969). Previous versions of the QUAID tool had different names: QQEA (QUEST Questionnaire Evaluation Aid) and Cochlea. We thank Scott Allen for his feedback on an earlier draft of this article. Correspondence concerning this article should be sent to A. C. Graesser, Department of Psychology, Campus Box 526400, University of Memphis, Memphis, TN 38152-6400 (e-mail: a-graesser@memphis.edu).

(DARPA, 1995; Jacobs, 1992; Lehnert, 1997). There has been noticeable progress in automating some components of language that lie within the span of a sentence, but there has been limited progress in handling deep comprehension and lengthy stretches of discourse. The good news, nevertheless, is that the computer aid does not need to be perfect in order to be useful. Rather than solving all of the problems that confront the designers of questionnaires, it can offer advice about those components for which it can deliver reliable feedback. Some of these components are so complex, technical, or subtle that they are invisible to the unassisted human eye, even the eye of an expert on questionnaire design or the eye of on accomplished computational linguist. For example, it would be impossible for these experts to catch all of the problems in sentence syntax and WM load. A computer aid would be useful even if it produced occasional errors in diagnosis. Such faulty diagnoses would be eliminated when the human experts scrutinize the computer output. We envision a computer aid that is used collaboratively with a human expert on questionnaire design, so the human can always supersede and make the final decision about each suggestion offered by the computer. The computer aid would be analogous to the *spellcheck* facility in most word processing packages; the computer suggests incorrect spellings, but it is the human writer who ultimately decides the proper spelling of each word. In essence, the computer does not replace the survey methodologist but is a tool that facilitates the work of the expert.

TWELVE COMMON PROBLEMS WITH QUESTIONS

Graesser, Bommareddy, Swamer, and Golding (1996) identified 12 potential problems with questions that periodically occur on questionnaires and that would be anticipated by a cognitive computational model of human question answering (called QUEST, as will be discussed shortly). These 12 problems are presented in Table 1. Graesser, Bommareddy, et al. reported that approximately one out of five questions on everyday forms and questionnaires suffers from at least one of the problems in Table 1. They conducted a study in which expert judges (who were trained on the QUEST model and the 12 problems in Table 1) were asked to identify problematic questions and the specific problems with each problematic question. There were five questionnaires in one of the studies that was conducted: the 1040 income tax form (75 questions), the 1990 census form (102 questions), an application for graduate admission to the University of Memphis (44 questions), a dentist intake form (74 questions), and an application for a job at Kinko's (42 questions). The likelihood of a question's having a particular problem listed in Table 1 varied from .006 to .057.

A cognitive computational model of human question answering, called QUEST, provided the theoretical foundation for investigating problems with questions. It is beyond the scope of this article to describe the details of this QUEST model (Graesser, Baggett, & Williams, 1996; Graesser & Franklin, 1990; Graesser, Gordon, & Brainerd, 1992; Graesser & Hemphill, 1991; Graesser, Lang, & Roberts, 1991), but a brief sketch of the mechanism is needed to convey the value in grounding the computer tool in a cognitive computational model.

Table 1
Problems With Questions
(Graesser, Bommareddy, Swamer, & Golding, 1996)

1. *Unfamiliar technical term*. There is a word or expression that very few respondents would know the meaning of.
2. *Vague or imprecise predicate or relative term*. The values of a predicate (i.e., main verb, adjective, or adverb) are not specified on an underlying 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, structurally ambiguous, or not well formed syntactically.
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 may 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 question 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.

QUEST specifies the computational procedures and strategies that humans execute when they answer 19 different categories of questions. Some of these categories are open-class questions that permit a small number of legal response alternatives, such as *verification* questions (Is X true? "Are you a citizen of the United States?") and *disjunctive* questions (Is X, Y, or Z the case? "Are you male or female?"). Some question categories invite short answers, such as *concept completion* questions (Who? What? When? Where? "Who is your physician?") and *quantification* questions (How many? What is the value of a variable? "How many children do you have?"). Many of the question categories invite lengthy descriptions in the answers, such as *causal antecedent* questions (What caused event X to occur? "Why did you lose your job?"), *goal orientation* questions (What goals motivated action X? "Why did you move to Tennessee?"), and *comparison* questions (How is X similar to/different from Y? "What is the difference between a dividend and interest?"). A hybrid question is an amalgamation of two question categories. For example, the following question would be a hybrid between the goal orientation and disjunctive categories: "Why did you move to Tennessee? ____for a job; ____for family; ___other."

The QUEST model has four major components, which together generate the answers to questions. These are (1) *question interpretation*: QUEST parses the question syntactically, identifies referents of nouns, segregates presuppositions, interprets predicates (i.e., verbs and adjectives), and isolates the focus of the question, and the question category is also identified in this component; (2) *access to relevant information sources*: QUEST activates the relevant generic knowledge structures (e.g., scripts, stereotypes, and other packages of world knowledge) and specific knowledge structures (i.e., episodic memories); (3) *pragmatics*: QUEST identifies the common ground (shared knowledge) and the goals of the questioner and respondent; and (4) *convergence to relevant answers*: QUEST searches through the vast landscape of relevant knowledge structures and produces the very small subset of nodes that constitute the good answers to the question. Some of these components are similar to, but not strictly identical with, the models of the question response process in the survey methodology literature (Cannell et al., 1981; Sudman et al., 1995; Tourangeau, 1984).

Most of the potential problems with questions listed in Table 1 are familiar to experts in survey methodology who have devised checklists and other methods for diagnosing specific flaws with problematic questions (Bickart & Felcher, 1996; Fowler, 1993; Jobe & Mingay, 1991; Lessler & Forsyth, 1996). It should be noted that our list of 12 problems with questions is probably not exhaustive, but it did handle 96% of the problems that we identified when examining dozens of forms and questionnaires (Graesser, Bommareddy, et al., 1996). The list of problems will presumably grow somewhat as the science of questionnaire design evolves further. Although the 12 categories are conceptually distinct and, therefore, mutually exclusive, they are sometimes interdependent and correlated. For example, a question might suffer from having an unclear purpose (Category 9) if there is an unfamiliar technical term (Category 1) or if the respondent is unlikely to know an answer (Category 12). Any given question can suffer from multiple problems.

In order to illustrate some of the problems in Table 1, consider the following problematic question. This question is on a questionnaire that hundreds of women have completed in a women's health clinic in Memphis.

> Did your mother, father, full-blooded sisters, full-blooded brothers, daughters, or sons ever have a heart attack or myocardial infarction? ( ) NO ( ) YES

It could be argued that this question suffers from most of the problems that are listed in Table 1. This question imposes *WM overload* in at least two ways. The first noun phrase is long and cumbersome; the respondent is forced to keep track of a long list of six or more family members. The respondent is asked whether each of these family members has had a heart attack or myocardial infarction, so there is a 6×2 matrix of implicit, embedded questions for those respondents who believe that a heart attack might be different from a myocardial infarction. A long list or matrix of questions is too much to keep track of in a WM that has limited capacity (Baddeley, 1986; Just & Carpenter, 1992). The question potentially has an *ambiguous noun phrase* for respondents with adoptive parents. This is especially the case for those who do not induce the purpose of the questionnaire—namely, to assess whether there are particular medical problems in the respondent's biological history. The expression "myocardial infarction" is undoubtedly an *unfamiliar technical term* for the majority of the respondents. For most respondents who are childless and from small families, there would be *incorrect presuppositions*; they would not have any full-blooded sisters, full-blooded brothers, daughters, and/or sons. It might be difficult or impossible to know whether some family members have had a heart attack or an infarction, so the question potentially suffers from Problems 11 and 12 in Table 1. This is especially true for respondents who were not raised by their biological parents.

The value of the QUAID tool is that it would help the survey methodologist to identify the 12 problems with questions and to revise the questions to correct the problems. Graesser, Kennedy, Wiemer-Hastings, and Ottati (1999) conducted a study that supports the claim that such a tool is likely to uncover problems that are frequently missed by (1) respondents who give feedback in a pretest phase and (2) judges who are trained to identify problems with questions. Survey researchers have frequently advocated the collection of think-aloud protocols from a sample of respondents during pretesting (Bickart & Felcher, 1996; Jobe & Mingay, 1991; Lessler & Sirken, 1985; Willis, Royston, & Bercini, 1991). Graesser et al. (1999) reported, however, that most of the problems in Table 1 are completely missed by respondents who give a critique of a survey during pretesting. The only problems that adult respondents can reliably identify are Problems 1 (unfamiliar technical term) and 3 (vague or ambiguous noun phrase). Graesser et al. (1999) also raised concerns that expert survey methodologists might miss many of the problems if they are not adequately trained in linguistics, discourse, and cognition. Our strong claim is that a computer aid (such as QUAID) provides a deeper and more detailed analysis of questions than that supplied by an expert in questionnaire design. It is an open empirical question, however, as to how well experts agree in identifying the 12 problems and how well the output of QUAID would compare with the experts. In the present study, answers to such questions will be explored.

GOALS OF THE PRESENT STUDY

The purpose of this article is to describe the QUAID tool and to report some data by which its performance may be evaluated. The current version of QUAID performs a critique of questions on the basis of the first nine problems with questions that are listed in Table 1. However, at this stage of developing the tool, we are satisfied with the

performance of only the first five problem modules, so the focus of this article will be on such problems as unfamiliar technical term, vague or imprecise relative term, vague or ambiguous noun phrase, complex syntax, and WM overload. We compared the output of QUAID with ratings of language experts who evaluated a corpus of questions on the five classes of problems. The corpus consisted of 505 questions on 11 surveys developed by the U.S. Census Bureau. After describing QUAID, we will report data on how well the tool compares with the decisions of experts in language, discourse, and cognitive psychology.

QUAID (QUESTION-UNDERSTANDING AID)

This section describes the QUAID computer tool. QUAID is grounded in a model of human cognition (QUEST), in addition to incorporating contemporary developments in computational linguistics (such as lexicons and syntactic parsers). QUAID has nine interface options, corresponding to the nine problems with questions. The computer user can turn each of the nine options on or off, depending on whether the user desires feedback on a component. There is also a *help* facility for each component; the user can read the help messages in order to learn about the particular type of problem with questions. The questionnaire designer first types a question into QUAID. Then QUAID performs a critique of the question on the nine different components (or as many of the nine as the user desires). We will focus on the first five problems with questions listed in Table 1, because we have not yet completed an adequate empirical analysis of the performance of Problems 6–9. Problems 6–9 were very infrequent in the corpus of surveys we analyzed, so we could not adequately test QUAID's performance.

QUAID currently runs on a Pentium computer with a Linux operating system. The software was developed in the LISP programming language. Individuals who are interested in acquiring the software should contact the first author of this article.

When a question is submitted to QUAID, there are three slots of information that get entered: focal question, context, and answer options. The focal question is the main question that is being asked, whereas the answer options (if any) are the response options that the respondent selects. The context slot includes sentences 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 following question.

> *Focal question*: From the date of the last interview to December 31, did you take one or more trips or outings in the United States, of at least 1 mile, for the primary purpose of observing, photographing, or feeding wildlife?
>
> *Context*: Do not include trips to zoos, circuses, aquariums, museums, or trips for scouting, hunting, or fishing.
>
> *Answer options*: Yes_____ No_____

QUAID allows a file to be entered that contains a list of questions on the survey, as long as each question is segmented into these three slots. The user can then scroll, one at a time, through the list of questions, for QUAID to evaluate. The user clicks on an "Analyze Question" option when the user is ready for QUAID to perform a critique of the question.

QUAID's critique of each question is a list of problems it identified. For example, if a question had one problem with each of the five categories in Table 1, QUAID would print out the following five summary messages.

> *Unfamiliar technical term*: The following term may be unfamiliar to some respondents: <unfamiliar technical term>
>
> *Imprecise relative term*: The following term refers implicitly to an underlying continuum or scale, but the point or value on the scale is vague or imprecise: <problematic term>
>
> *Vague or ambiguous noun phrase*: The referent of the following noun may be vague or ambiguous to the respondent: <problematic term>
>
> *Complex syntax*: The question is either ungrammatical or difficult to parse syntactically.
>
> *WM overload*: The question imposes a heavy load on the WM of the respondent.

In addition to this short feedback, there is a *help* facility that defines each problem more completely and gives examples of particular problems. This help facility allows the survey methodologist to dissect and repair the problem with a particular question. It should be noted, however, that QUAID does not perform a complete analysis of particular problems with a particular question, such as *what* syntactic constituents are problematic or *where* the WM overload occurs. The help facility provides clues about likely problems that frequently occur, but it is up to the survey methodologist to reconstruct the pathology of a particular question. Nevertheless, knowing *that* a problem occurs is a prerequisite to identifying the exact source of the problem and how it can be fixed.

QUAID adopts both theoretical and empirical criteria when deciding whether questions have a problem. Regarding theory, the process of developing QUAID involved exploring a large space of features, feature combinations, algorithms, metrics, and parameters that are potentially diagnostic for identifying a particular class of problems with questions. For example, in the case of syntax, there were metrics that computed the number of constituents at the top level of a parse, the number of levels of constituents in the parse (i.e., depth), the number of subordinate clauses, the number of relative clauses, and so forth (see Allen, 1995, for an excellent discussion of syntactic parsers and metrics of difficulty). We used correlational analyses to explore which of the alternative measures of syntactic complexity best predicted ratings of syntactic complexity that were provided by experts in language, discourse, and cognition (as will be discussed later). It is beyond the scope of this article to document the total set of criteria that we tested for each problem. Instead, we will

specify which criteria were selected in the current version of QUAID. It suffices to say that QUAID will be undergoing cycles of revision to explore additional criteria for identifying problems.

Unfamiliar Technical Term

Each word in the focal question and answer has a word frequency in the English language. QUAID has tested out a number of databases and lexicons that have information about word frequency and familiarity in the English language, including Francis and Kučera (1982), the MRC psycholinguistic database (Coltheart, 1981), and the WordNet lexicon (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). The best criterion that we found for determining whether a word was unfamiliar was whether the word either (1) has an MRC familiarity value of less than 500 or (2) is not in the WordNet lexicon.

Vague or Imprecise Relative Term

There is an adverb, adjective, or main verb that refers implicitly to an underlying continuum or scale. However, the point on the continuum, or the value on the scale, may be vague, imprecise, or ambiguous to the respondent (Moxey & Sanford, in press; Sanford, Moxey, & Paterson, 1996). For example, *sometimes*, *often*, and *rarely* are "relative adverbs" that may present problems to the population of respondents. Will the respondents know how frequently the event needs to occur in order to count as *frequently*? Will respondents agree? Examples of relative adjectives are *moderate, severe,* and *difficult*. Examples of relative verbs are *try*, *work*, and *hurt*. A decision needs to be made whether the term is sufficiently vague, imprecise, or ambiguous that it will present a problem to the population of respondents. QUAID has an exhaustive list of the relative adjectives and adverbs in the English language that specify frequency, intensity, quantity, and temporality. A partial list of the relative verbs is available, but this was only able to handle the main verbs in the corpus of the surveys that were tested. QUAID regards a question as having a problem in this category if there is a relative term in either the focal question or answer options.

Vague or Ambiguous Noun Phrase

The referent of a noun phrase, noun, or pronoun is potentially vague or ambiguous. Ambiguous nouns sometimes have two or more senses, so the respondent may not know which sense is relevant to the question. For example, *project* may refer to a cluster of low-income houses or to a major work activity. Abstract words are frequently vague or ambiguous. An ambiguous noun may refer to two or more entities in the discourse context, so the respondent is uncertain which entity was intended in the question. For example, *sibling* may refer to the respondent's sibling or the sibling of the respondent's child. Pronouns (*it*, *that*, *he*) often have such ambiguities. A decision needs to be made as to whether the term is sufficiently vague or ambiguous that it will present a problem to the population of respondents. QUAID currently identifies a word as raising problems if one or more of the following criteria are met: (1) the concreteness value of the word in the MRC database is below the threshold of 179; (2) the average number of hypernyms for the nouns (i.e., more general nouns in a semantic hierarchy) is less than 3.24; (3) the head of a noun phrase (with no attachments) has a polysemy value of greater than 19 when consulting WordNet; or (4) the word is a member of a list of vague noun phrases (which includes pronouns). Once again, this is the combination of features and parameters that best predicted the judgments of human experts.

Complex Syntax

The grammatical composition of the focal question is embedded, dense, ambiguous, or ungrammatical. There are thousands of ways that a question can have a problem with its grammatical composition. For example, a verb may be missing. There may be too many clauses or adjectives to hold in memory by the time a main verb or a noun appears. The verb may not agree with the subject noun in number (singular vs. plural) or some semantic feature. QUAID uses a part-of-speech tagger that was developed by Brill (1995) and a SCOL syntactic parser that was developed by Abney (1997). SCOL generates the most likely syntactic tree structure that would be assigned to the focal question or context sentence. A sentence is considered to have a complex syntax if one or both of the following conditions are met. First, there are more than 12 constituents at the top level of the parse of the sentence. This criterion presupposes that a sentence with a problem would have at least 12 words, so it would not be applied to sentences with 11 or fewer words. Second, there are 10 or more NX constituents in a sentence; an NX is a noun phrase with no attachments (e.g., prepositional phrases). Both of these criteria are not applicable to shorter sentences. However, not surprisingly, it is the lengthier sentences that tend to have difficulties with syntactic complexity. Versions of QUAID in the future are expected to improve on the syntactic component, but the existing version is satisfactory for this initial assessment of complex syntax.

Working Memory Overload

It is widely acknowledged that comprehension is constrained by a WM that is limited in capacity (Goldman, Varma, & Cote, 1996; Just & Carpenter, 1992; Kintsch, 1998). Capacity limits both the number of processing operations that can be executed during a time span and the number of units that can be preserved in a passive storage buffer. The implications of these WM limitations on questionnaire design are perfectly obvious. Questions should be written in a fashion that minimizes the load on WM. Unfortunately, many questions pack a large number of clauses, qualifiers, and prepositional phrases into a single question. Sentences with *right-branching* syntax are easy to process, because they first present the main clause (e.g., an assertion or a question) and subsequently add on clauses and phrases that qualify the first clause.

In contrast, sentences with a *left-embedded* syntax are difficult, because the main clause is never finished until the end of the sentence and WM must maintain the unfinished information. Thus, some of the problems with syntactic complexity also predict problems with WM overload. Another feature of a question that imposes a heavy WM load is a large number of Boolean alternatives to consider (i.e., and, or). Consider the following problematic question from the 1990 U.S. census:

> Do you have a physical, mental, or other health condition that has lasted for 6 or more months and which limits the kind of work you do at the job?

In order to answer this question, the respondent must consider each cell in a mental matrix of alternatives.

QUAID adopted two criteria for identifying questions that impose a high WM load, because these two criteria significantly predicted the ratings of human experts. First, there are more than 12 constituents at the top level of the parse of the sentence. This criterion also served as a criterion in the syntactic complexity component. Second, there were more than two conjunctions (i.e., and, or, if) in the sentence, which is an indicator of complex Boolean expressions.

Once again, QUAID also has components corresponding to Problems 6–9 in Table 1. However, these will not be presented and discussed in the present article.

EVALUATION OF QUESTIONS BY HUMAN EXPERTS

Experts evaluated a corpus of questions on the first nine problems listed in Table 1. The three experts were all extensively trained on the nine problems with questions. All three experts had a master's degree or a doctoral degree in a field that investigated the mechanisms of language, discourse, and/or cognition. Each expert judged whether a question had any of the nine problems. The following rating scale was used in making these judgments: 1, *definitely not a problem*; 2, *probably not a problem*; 3, *probably a problem*; and 4, *definitely a problem*.

Corpus of Surveys Developed at the U.S. Bureau of Census

Eleven surveys were selected for testing QUAID. These included the following: *Hunting and Fishing Questionnaire*, third detailed interview, 1991 (Form FH-3C); *Nonconsumptive User's Questionnaire*, third detailed interview, 1991 (form FH-4C); *1993 Survey of Working Experience of Young Women* (form LGT-4161); 1996 *American Community Survey* (form ACS-1); *United States Census 2000 Dress Rehearsal* (form DX-2); *Adolescent Self-Administered Questionnaire: Survey of Program Dynamics* (form SPD-18008); 1998 *National Health Interview Survey Basic Module: Adult Core* (version 98.1); 1998 *National Health Interview Survey Basic Module: Household Composition* (version 98.1); 1998 *National Health Interview Survey: Child Prevention Module* (version 98.1); *Crime Incident Report: National Crime Victimization Survey* (form NCVS-2); and *Survey of Program Dynamics: Adult Questionnaire*. All of these surveys were furnished by the U.S. Census Bureau.

The corpus of questions in the sample included the first 50 items in each survey or all of the items if there were fewer that 50 questions. Some items had multiple questions; in these cases, we selected the first question within the item. When we prepared the files with the questions in this question corpus, we removed instructions to the interviewer and extraneous symbols and codes that frequently occur in the Census files. We also segregated the question into three portions: (1) the focal question, (2) context sentences, and (3) answer options. Some of the questions were deleted because they were opinion questions, rather than factual questions about the respondent. The final corpus had 505 usable questions. Also, we originally split the corpus of questions into a training corpus and a test corpus; the training corpus consisted of the odd-numbered questions whereas the test corpus consisted of the even-numbered questions. The purpose of doing this was to tune QUAID to maximize performance on the training corpus but to use the test corpus to evaluate the generality of the performance of QUAID. When we tuned the training corpus, we made sure that all of the words from the questions were in the relevant lexicons and that all of the vague relative terms and noun phrases were identified. We also tuned threshold parameters (as in the case of word frequency and syntactic complexity) so that there was a maximum correlation with the judgments of the experts. However, performances on the training and test samples were indistinguishable, so we decided to collapse these samples in the present article.

Scoring the Experts' Ratings of Problems With Questions

Table 2 presents a summary of the problem evaluation ratings by the experts. Three measures are reported in the table, as defined below.

> Problem incidence = Proportion of questions in which at least one expert had a rating of 3 or 4,
>
> Problem score = (sum of expert ratings −3) / 9,
>
> Interjudge reliability = Proportion of agreements among pairs of experts (1–2 vs. 3–4 split).

A number of conclusions can be drawn from the data in Table 2. First, the five problems were not rare occurrences

Table 2
Problems Identified by Human Experts

Problem	Problem Incidence	Problem Score	Interjudge Reliability
Unfamiliar technical term	.238	.131	.83
Vague or imprecise relative term	.403	.184	.73
Vague or ambiguous noun phrase	.486	.184	.69
Complex syntax	.328	.151	.77
Working memory overload	.274	.147	.81

in the corpus of questions, even though the questions had been pretested and scrutinized by personnel at the U.S. Census Bureau. Second, the interjudge reliability among the judges was significantly above chance but hardly impressive. The proportion of common decisions on the 1–2 versus 3–4 rating split varied between .69 and .83, which is rather modest. Other measures of reliability (i.e., correlations among ratings, and Kappa scores) were significant in the majority of the cells, but rather low.

There are plausible explanations for the variability among experts. First, it was discovered during debriefing that the three judges weighted the various criteria differently when they made the judgments. Second, the judges may have experienced some problems of fatigue while making the 4,545 ratings (9 problems × 505 questions). Third, the detection of some problems is very subtle, so subtle that they end up being missed by language experts. This outcome indeed justifies the need for the QUAID tool; the tool will reveal problems that even language experts end up missing sometimes. Graesser, Bommareddy, et al. (1996; Graesser et al., 1999) argued that a computer tool would prove useful to the extent that it spots problems that are missed by survey methodologists and language experts. Thus, the survey methodologist plus the QUAID tool together should do better than the survey methodologist alone. This conclusion has an interesting implication. It is not clear what should serve as the gold standard for declaring that there is a problem with a question. We adopted the human experts as a standard, but the possibility remains that the QUAID tool is better than the human in detecting some problems.

COMPARISON OF QUAID AND HUMAN EXPERTS

This section evaluates how well QUAID fares in detecting problems with questions when human experts are used as the gold standard for a correct identification of a problem. So, truth is defined as the judgment of human experts. It should be noted, however, that the problem incidence (and the problem score) of human experts is a continuous variable, not a discrete variable. Therefore, we need to consider different thresholds of problem incidence when declaring whether there is a problem with a question. The most lenient criterion *threshold* is .11; if any of the three human experts assigned a rating of 2 (with the ratings of the other two experts being 1), the problem incidence score would be .11. This criterion is undoubtedly too lenient, but we will nevertheless use this as one extreme for assessing the performance of QUAID. The other criterion thresholds had more intermediate problem incidence scores: .33, .44, and .56. Given a criterion threshold of *T*, a question was declared to be a problem for human experts if the problem incidence score was *T* or greater; the question was not declared a problem if the problem incidence score was less than *T*. As the criterion threshold increases, the human experts would consider fewer questions to be problematic. This is reflected in the problem likelihood score, the proportion of 505 questions in the corpus that are classified as problematic for criterion *T*. As the threshold criterion increases (i.e., becomes more stringent), the problem likelihood necessarily decreases.

Signal detection analyses can be performed on the data, once we have classified questions as being problematic versus nonproblematic for any given criterion threshold *T*. Using the terminology of signal detection theory, a target item is a question that human experts regard as a problem (given threshold *T*), whereas a nontarget item is a question that human experts regard as nonproblematic. The following metrics can then be computed.

Hit rate = p(computer sees problem | human sees problem),

False alarm rate (FA) = p(computer sees problem | human sees no problem),

d' score = computer's discriminative ability to identify problem, in theoretical standard deviation units.

Signal detection analyses are quite familiar to most experimental psychologists. A high d' score value would mean that the QUAID tool would do an excellent job of discriminating between questions that are problematic versus nonproblematic, at least when the human experts are the gold standard. A different way of analyzing the same data adopts the metrics used in the field of computational linguistics (DARPA, 1995; Lehnert, 1997). Computational linguists collect recall and precision scores. These measures are defined below, with H signifying the frequency of hits, FA signifying the frequency of false alarms, and M signifying the frequency of misses.

Recall score = H/(H + M) = hit rate,

Precision score = H/(H+FA).

The measures of both signal detection theory and computational linguistics will be reported in this section.

Table 3 reports the different performance measures for the five categories of problems with questions. That is, hit rates, false alarm rates, d' scores, recall scores, and precision scores are presented as a function of four different values of *T* (.11, .33, .44, and .56) for each of the five question categories. Problem likelihoods are also included. It follows mathematically that there is an inverse relationship between the threshold criterion *T* and problem likelihood. The intermediate values of *T* (.33 or .44) are the most feasible thresholds to consider when evaluating the data. A very low value of *T* (namely, .11) is too lenient a criterion, so a large number of questions would be classified as problematic by human experts; a high value of *T* (namely, .56) is so stringent that very few questions would be classified as problematic. However, we include data for the extreme values of *T* in order to unveil the performance of QUAID across a large continuum of thresholds.

A number of conclusions are supported by the data in Table 3. The most important conclusion is that QUAID is able to discriminatively identify problems with the

Table 3
Comparison of QUAID and Human Experts in Detecting Problems With Questions

Problem	Criterion of Human Experts	Hit Rate (Recall Score)	False Alarm Rate	d' Score	Precision Score	Problem Likelihood
Unfamiliar technical term	.11	.71	.34	.91	.45	.28
	.33	.79	.37	1.14	.32	.18
	.44	.86	.41	1.31	.17	.09
	.56	.79	.43	.99	.08	.05
Vague or imprecise relative term	.11	.77	.42	.94	.59	.44
	.33	.84	.47	1.08	.41	.28
	.44	.94	.53	1.48	.17	.10
	.56	.91	.57	1.16	.03	.02
Vague or ambiguous nounphrase	.11	.75	.49	.70	.62	.51
	.33	.80	.54	.74	.39	.30
	.44	.95	.61	1.37	.06	.04
	.56	1.00	.62	2.01	.01	.01
Complex syntax	.11	.12	.01	1.13	.88	.38
	.33	.12	.03	.70	.60	.25
	.44	.29	.03	1.33	.40	.07
	.56	.58	.04	1.95	.28	.02
Working memory overload	.11	.14	.02	.97	.75	.33
	.33	.21	.02	1.23	.69	.20
	.44	.29	.04	1.20	.34	.08
	.56	.47	.05	1.57	.22	.03

five classes of questions. When considering the criterion threshold value of .44, the d' scores for unfamiliar technical term, vague/imprecise relative term, vague/ambiguous noun phrases, complex syntax, and WM overload were 1.31, 1.48, 1.37, 1.33, and 1.20, respectively. All of these d' scores were statistically significant when we analyzed the frequency tables and computed chi-squares. That is, a chi-square test of association was computed on each 2×2 frequency table that includes the frequency of hits, misses, false alarms, and correct rejections. The d' scores were slightly lower for the more lenient .33 value of T (at least for Problems 1, 2, 3, and 4) and for the extremely lenient .11 value.

A second conclusion is that the hit rates and false alarm rates had remarkably different patterns among the five classes of questions. The hit rates were quite high for the first three problem categories (.84–.95 for $T = .44$), but so were the false alarm rates (.41–.61). QUAID does a good job in detecting these classes of problems, but at the expense of generating false alarms that may not be problematic under more careful analysis. Follow-up analyses could be conducted by having experts evaluate how many of the false alarms are truly unproblematic. If many of the false alarms are not really problems, the survey methodologist would have many questions flagged as problems but would have to spend extra time rejecting many questions that are not problematic. Future versions of QUAID need to find principled ways of reducing the false alarm rate without seriously lowering the hit rate. In contrast, Problem 4 (complex syntax) and Problem 5 (WM overload) had low hit rates and extremely low false alarm rates. In these cases, future versions of QUAID need to have more sensitive algorithms and metrics for picking out problematic questions. The recall scores and precision scores, measures that are standard in computational linguistics, are compatible with these conclusions. That is, there is a tradeoff between recall scores and precision scores. For the first three problem categories, the recall scores are more impressive than the precision scores; for Problems 4 and 5, the recall scores are less impressive than the precision scores. These analyses provide some informative guidance in modifying QUAID in the future.

CONCLUSIONS

The results of this study have supported the claim that the QUAID tool is able to identify questions that suffer from five different classes of problems. QUAID can significantly discriminate the problems that human experts also identify as problematic (vs. nonproblematic). Of course, there is room for QUAID to improve. The false alarm rate is high for unfamiliar technical terms, vague or imprecise relative terms, and vague or ambiguous noun phrases, whereas the hit rate needs to increase for complex syntax and WM overload. Nevertheless, the results are encouraging news for the development of a computer tool to assist designers of questionnaires and surveys.

One persistent question addresses the gold standard for identifying a question as problematic. We adopted human experts as the gold standard, but there are reasons for being skeptical about this approach. Human experts do not show a high amount of agreement in identifying particular problems with questions. Perhaps the modest interjudge reliability scores can be explained by the variability in their research background, by the subtlety of the theoretical components, or by fatigue. These reasons all converge on the value of a computer tool in assisting the survey methodologist. Indeed, the problems identified by the computer may end up being the better gold standard because of the rich set of analyses that can be performed. The matter of the appropriate gold standard awaits future research.

REFERENCES

Abney, S. (1997). *The SCOL manual (Version 0.1b)*. Unpublished manuscript, University of Tübingen (www.sfs.nphil.uni-tuebingen.de/\Tilde abney/).

Allen, J. (1995). *Natural language understanding*. Redwood City, CA: Benjamin/Cummings.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Bickart, B., & Felcher, E. M. (1996). Expanding and enhancing the use of verbal protocols in survey research. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 115-142). San Francisco: Jossey-Bass.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, **21**, 1-24.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389-437). San Francisco: Jossey-Bass.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.

DARPA (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. San Francisco: Morgan Kaufman.

Fowler, F. J. (1993). *Survey research methods*. Newbury Park, CA: Sage.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 15-36). San Francisco: Jossey-Bass.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage*. Boston: Houghton-Mifflin.

Goldman, S. R., Varma, S., & Cote, N. (1996). Extending capacity-constrained construction integration: Toward "smarter" and flexible models of text comprehension. In B. F. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 73-114). Hillsdale, NJ: Erlbaum.

Graesser, A. C., Baggett, W., & Williams, K. (1996). Question-driven explanatory reasoning. *Applied Cognitive Psychology*, **10**, S17-S32.

Graesser, A. C., Bommareddy, S., Swamer, S., & Golding, J. M. (1996). Integrating questionnaire design with a cognitive computational model of human question answering. In N. Schwartz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 143-174). San Francisco: Jossey-Bass.

Graesser, A. C., & Franklin, S. P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes*, **13**, 279-303.

Graesser, A. C., Gordon, S. E., & Brainerd, L. E. (1992). QUEST: A model of question answering. *Computers & Mathematics with Applications*, **23**, 733-745.

Graesser, A. C., & Hemphill, D. (1991). Question answering in the context of scientific mechanisms. *Journal of Memory & Language*, **30**, 186-209.

Graesser, A. C., Kennedy, T., Wiemer-Hastings, P., & Ottati, V. (1999). The use of computational cognitive models to improve questions on surveys and questionnaires. In M. G. Sirken, D. J. Hermann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey methods research* (pp. 199-216). New York: Wiley.

Graesser, A. C., Lang, K. L., & Roberts, R. M. (1991). Question answering in the context of stories. *Journal of Experimental Psychology: General*, **120**, 254-277.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Jacobs, P. S. (Ed.) (1992). *Text-based intelligent systems: Current research and practice in information extraction and retrieval*. Hillsdale, NJ: Erlbaum.

Jobe, J. B., & Mingay, D. J. (1991). Cognition and survey measurement: History and overview. *Applied Cognitive Psychology*, **5**, 175-192.

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, **99**, 122-149.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Lehnert, W. G. (1997). Information extraction: What have we learned? *Discourse Processes*, **23**, 441-470.

Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 259-291). San Francisco: Jossey-Bass.

Lessler, J. T., & Kalsbeek, W. (1993). *Nonsampling error in surveys*. New York: Wiley.

Lessler, J. T., & Sirken, M. G. (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 & Society*, **63**, 565-581.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). *Five papers on WordNet* (Rep. No. 43). Princeton, NJ: Princeton University, Cognitive Science Laboratory.

Moxey, L. M., & Sanford, A. J. (in press). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology*.

Sanford, A. J., Moxey, L. M., & Paterson, K. B. (1996). Attentional focusing with quantifiers in production and comprehension. *Memory & Cognition*, **24**, 144-155.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, **60**, 576-602.

Schwarz, N., & Sudman, S. (Eds.) (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco: Jossey-Bass.

Sirken, M. G., & Fuchsberg, R. (1984). Laboratory-based research on the cognitive aspects of survey methodology. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 183-197). Washington, DC: National Academy Press.

Sirken, M. G., Hermann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., & Tourangeau, R. (Eds.) (1999). *Cognition and survey methods research*. New York: Wiley.

Sudman, S., Bradburn, N. M., & Schwarz, M. (1995). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. J. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy of Sciences.

Willis, G., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology*, **5**, 251-267.

(Manuscript received October 29, 1999;
revision accepted for publication February 24, 2000.)