Session 3

A Computer Tool to Improve Questionnaire Design

FCSM Seminar on the Funding Opportunity in Survey Research
Introduction to Session 3, "A Computer Tool To Improve Questionnaire Design"

Chair, Robert Parker, U.S. General Accounting Office

The subject of our 3rd session today is "A Computer Tool To Improve Questionnaire Design" and features a paper by a number of faculty members of the University of Memphis, headed by Professor Arthur Graesser, who will make this morning's presentation. Before introducing our speaker, I'd like to say that I am very pleased to be chairing this session, because I strongly support research designed to help reduce nonsampling errors and to increase response rates. The work described in this paper looks like a promising step in that direction, and I look forward to hearing comments from our discussants.

Now let me introduce our speaker. Professor Graesser is presently a full professor in the Department of Psychology and an adjunct professor in Mathematical Sciences at the University of Memphis. He is currently a co-director on the Institute for Intelligent Systems and director of the Center for Applied Psychological Research. Dr. Graesser received his Phd in psychology from the University of California at San Diego and has as his primary research interests cognitive science and discourse processing. He is currently editor of the journal Discourse Processing. In addition to publishing over 200 articles, he has written 2 books and edited several others.

Our first discussant will be Terry DeMaio, a principal researcher in the Census Bureau's Center for Survey Methods Research. She has been at the Census Bureau for 25 years, working on research issues related to nonresponse and questionnaire design. She currently heads a group that conducts research on the Bureau's demographic surveys. Terry received her graduate training in sociology at University of Indiana.

Our second discussant will be Fran Featherston. Fran is a senior survey researcher at the General Accounting Office and has extensive research in the design and analysis of a wide variety of surveys. Fran received a Phd in political science from the University of Michigan.

I want to thank Professor Graesser and our two discussants for their presentations. I also would like to add additional comments on the QUAID computer tool. First, it would seem to me that this tool would be useful not only to survey designers but also to managers in statistical agencies. Using QUAID, or some derivative program, for all surveys could provide managers with the knowledge that the questions in their surveys have been designed in a way to reduce comprehension problems by respondents. Second, it would seem that a next major step in the development of QUAID would be the ability to apply it simultaneously to groups of similar questions on a single survey.

A Computer Tool to Improve Questionnaire Design

Arthur C. Graesser, Ashish B. Karnavat, Frances K. Daniel, Elisa Cooper,
Shannon N. Whitten, and Max Louwerse
University of Memphis

Paper presented at the Funding Opportunity in Survey Research Seminar, June 11, 2001, Bureau of Labor Statistics, Washington, DC.

Send correspondence to: Arthur C. Graesser, Department of Psychology, 202 Psychology Building The University of Memphis, Memphis, TN 38152-3230, (901) 678-2742, 901-678-2579 (fax), a-graesser@memphis.edu.

This research was funded by grants awarded to the first author by the Statistical Research Division of the United States Census Bureau (1998-1999, 43-YA-BC-802930) and the National Science Foundation (SBR 9977969).

Abstract

We have developed a computer tool (called QUAID) that assists survey methodologists who want to improve the wording, syntax, and semantics of questions on surveys and questionnaires. QUAID stands for "Question Understanding Aid." The input to QUAID consists of a question on a questionnaire, whereas the output is a list of potential problems with the question, including: (1) unfamiliar technical term, (2) vague or imprecise relative term, (3) vague or ambiguous noun-phrase, (4) complex syntax, and (5) working memory overload. QUAID is now available on the web (www.psyc.memphis.edu/quaid.html). This web facility encourages researchers to send us problematic questions so that we can iteratively assess and improve the performance of QUAID. We have performed analyses that assess how well QUAID diagnoses these five problems with questions, sampled from a corpus of 11 surveys provided by the US Census Bureau. We have also collected eye- tracking data while college students answer 69 questions.

Introduction

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. So how do survey methodologists identify questions that are difficult for respondents to comprehend? One method is to have experts identify particular problems with questions (Lessler & Forsyth, 1996). A second approach is to collect verbal protocols from respondents as they answer questions (Willis, DeMaio, & Harris-Kojetin, 1999); some of the problems with questions can be articulated by respondents. A third approach is to observe behaviors, such as pauses or requests for clarification, that suggest that the respondents are struggling with a particular question (Fowler & Cannell, 1996; Schober & Conrad, 1997).

A fourth approach is to build a computer model that identifies problems with questions in a theoretically principled or systematic fashion (Graesser, K. Wiemer-Hastings, Kreuz, P. Wiemer-Hastings, & Marquis, 2000). Building such a computer requires the coordination of several fields, including computer science, computational linguistics, discourse processing, cognitive science, and survey methodology. This fourth approach was pursued in the present project. We have developed a computer program (called QUAID) that critiques questions on different comprehension problems.

Researchers in CASM (Cognitive Aspects of Survey Methodology) have adopted models that dissect different stages question-answering (Jobe & Mingay, 1991; Lessler & Sirken, 1985; Sudman, Bradburn, & Schwarz, 1995; Schwartz & Sudman, 1996; Tourangeau, 1984; Sirken, Hermann, Schechter, Schwarz, Tanur, & Tourangeau, 1999). The stages included in most of these models are question interpretation, memory retrieval, judgment, and response selection. The inaccuracy 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 & Kalsbeck, 1993; Schober & Conrad, 1997). Therefore, revising questions to minimize interpretation problems is one important strategy for reducing measurement error. QUAID was designed to diagnose interpretation problems, as opposed to other stages of questions answering (memory retrieval, judgment, and response selection).

QUAID stands for Question Understanding Aid. It has particular modules that critique each question on potential comprehension difficulties at various levels of language, discourse and world knowledge. The critique identifies words that are unfamiliar to most respondents, vague predicates (verbs, adjectives, adverbs), ambiguous noun-phrases, questions with complex syntax, and questions that overload working memory (Graesser, K. Wiemer-Hastings, Kreuz, P. Wiemer-Hastings, & Marquis, 2000). The identification of such problems should be useful to the survey methodologist if the computer tool can accurately identify the questions with potential problems and can point out what the problems are. Some of these problems might otherwise be missed because of fatigue or training deficits in the survey researcher who writes, revises, and pretests the

questions. Most survey researchers do no have extensive training in linguistics, discourse processing or cognitive science, so QUAID should be a valuable augmentation to the standard tools of the survey methodologist.

This paper is a progress report on our development and evaluation of QUAID. Section 1 presents a succinct overview of QUAID. The is a web facility that survey methodologists can use to obtain critiques of questions with QUAID. Our hope is that survey methodologists use this web facility and send us problematic questions; these will be used in future tests and refinements of QUAID. The second section reports a recent evaluation of the performance of QUAID. That is, how well can it accurately discriminate questions with particular problems, when compared to expert evaluations as a gold standard. The third section describes an eye tracking study that recorded the eye fixations and eye movements while respondents answer survey questions. We are currently assessing the extent to which eye tracking patterns reveal problems with questions.

QUAID (Question Understanding Aid)

This section briefly describes the QUAID computer tool. QUAID can handle 5 problems with questions, as described shortly. The questionnaire designer first types a question into QUAID. Then QUAID critiques the question on the 5 different components. There are three levels of each critique that vary in specificity, from succinctly identifying a problem to a lengthy description of the nature of the particular problem.

Graesser's previous research has identified 12 problems with questions that periodically occur in surveys (Graesser, Bommareddy, Swamer, & Golding, 1996; Graesser, Kennedy, P. Wiemer-Hastings, & Ottati, 1999). Many of these problems have been incorporated in various analytical coding schemes of survey methodologists. The current version of QUAID reliably handles the five problems below.

(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 (e.g., *try, large*, *frequently*).

(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*).

(4) Complex syntax. The grammatical composition is embedded, dense, structurally ambiguous, or not syntactically well-formed.

(5) Working memory overload. Words, phrases, or clauses impose a high load on immediate memory.

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. 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 3 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 one 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's critique of each question is a list of problems it identified. For example, if a question had a one problem with each of the 5 categories, 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.
>
> WORKING MEMORY OVERLOAD: The question imposes a heavy load on the working memory of the respondent.

In addition to this short feedback, there are two additional levels of extended help that define each problem more completely and that give examples of particular problems. This extended help allows the survey methodologist to dissect and repair the problem with a particular question.

It is beyond the scope of this paper to provide the technical details of how QUAID identifies problems (see Graesser, et al., 1996, 1999; 2000). QUAID adopts both theoretical and empirical criteria when deciding whether a question has a problem. Regarding theory, the process of developing QUAID involved exploring a large space of features, modules, and mechanisms in computational linguistics 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 subordinate clauses, the number of relative clauses, and so forth (see Jurafsky & Martin, 2000 for recent developments in computational linguistics and natural language processing in artificial intelligence). We used correlation analyses to explore which of the alternative measures of syntactic complexity best predicted the ratings of syntactic complexity that were provided by language experts. As another example, unfamiliar technical terms were identified by accessing computer lexicons that specify the frequency of words in the English language.

QUAID currently runs on a Pentium computer with a Linux operating system. The software includes a number of processing modules written in different computer languages (Java, LISP, C). QUAID is currently available on the web (www.psyc.memphis.edu/quaid.html), available to the public for free. However, individuals will not be able to use QUAID unless they provide us their names, address, email, telephone number, and other pertinent information. QUAID users must also agree to our analyzing their questions for research purposes, in exchange for their free use of the facility. The originator of the questions will be kept anonymous, in compliance with the ethical use of human subjects in research. We will use these questions for the evaluation and refinement of QUAID. QUAID currently handles only one question at a time, whereas a future version of QUAID will accommodate a set of survey questions.

Performance of QUAID when Compared to Human Experts as a Gold Standard

This section discusses how well QUAID fares in detecting problems with questions when using human experts as the standard for a correct identification of a problem. So truth is defined as the judgment of human experts. It should be noted that a problem spotted by human experts may be a continuous variable, rather than a discrete variable (i.e., problem versus no problem). Thus, a question Q is said to have problem P on a continuum that varies from 0 to 1.0; this we define as problem score. Intermediate values of the problem score reflect differences among experts and different strengths of problemhood within the judges. We considered different thresholds of the problem score when declaring whether there is a problem with a question. That is, a question Q was said to have problem P if the problem score of experts met or exceeded some threshold T.

Graesser et al. (2000) conducted a study that assessed how well experts can identify the five problems with questions. Experts evaluated a corpus of 550 questions on the five problems (2750 judgments altogether). The three experts were extensively trained on the problems with questions and had a graduate degree in a field that investigated the

mechanisms of language, discourse, and/or cognition. The experts judged whether or not each question had each of the 5 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. The problem score was computed as: (sum of expert ratings – 3) / 9. A question was defined as having a problem P if the problem score $\geq$ threshold T.

Eleven surveys were selected for testing QUAID. These included: *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); *Survey of Program Dynamics: Adult Questionnaire*. These surveys were furnished by the United States Census Bureau.

Signal detection analyses were performed on the data after we classified questions as being problematic versus non-problematic 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

A high *d'* score means that the QUAID tool does an excellent job discriminating between questions that are problematic versus non-problematic, at least according to the standard of the human experts. The d' score is a pure measure of the ability of QUAID to discriminate problems with questions, after controlling for guessing biases. Another useful measure is called a **problem likelihood**, which is the proportion of questions that are classified as problematic according to the experts (given some threshold T on the problem scores).

There have been previous evaluations of QUAID on the corpus of 550 questions provided by the US Census Bureau (Graesser et al., 2000; Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, & Kreuz, 2000). These previous evaluations support the claim that QUAID has discriminative validity in identifying all five problems with questions, as defined by the experts. Table 1 summarizes the results of the evaluation reported in these

studies. Table 1 presents the different performance measures for the 5 categories of problems with questions. These include the hit rates, false alarm rates, *d'* scores, and problem likelihood scores. We selected suitable threshold values of problem scores that optimized hit rates, *d'* scores and problem likelihood scores.

Table 1: Comparison of QUAID and human experts in detecting problems with questions

	Hit Rate	False alarm Rate	d' score	Problem likelihood
(1) Unfamiliar technical term	.86	.41	1.31	.09
(2) Vague or imprecise relative term	.94	.53	1.48	.10
(3) Vague or ambiguous noun-phrase	.95	.61	1.37	.04
(4) Complex syntax	.29	.03	1.33	.07
(5) Working memory overload	.29	.04	1.20	.08

Several conclusions can be drawn from the data in Table 1. First, the QUAID tool was able to discriminate problematic questions because the *d'* scores were significantly above zero. Second, 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 3 problem categories (.86 to .95), but so were the false alarm rates (.41 to .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. So the survey methodologist would have many questions flagged as problems, but would have to spend extra time rejecting many questions that are not problematic. An improved QUAID needs to have computational methods of not being fooled by false alarms. In contrast, problem 4 (complex syntax) and problem 5 (WM overload) had low hit rates and extremely low false alarm rates. In these cases, QUAID needs to have more sensitive algorithms and metrics for picking up problematic questions. For all 5 problems, the problematic likelihood scores were quite low (ranging from .04 to .10). Thus, only 1 out of 10 to 25 questions suffered from a particular problem.

During the course of our research project, we have been exploring improved computational procedures for identifying problems with questions. We recently have been particularly interested in improving the complex syntax evaluator because it had previously shown a poor ability to detect problematic questions. In order to provide a more sensitive assessment, we desired a sample of questions that were more evenly split between problematic and nonproblematic questions. Therefore, we selected a sample of 94 questions from the original 550 questions in the question corpus; this restricted corpus had a higher incidence of problematic questions. First, we selected the top 50 problematic questions, using problem score measures that integrated over the 5 problems. Second, we randomly selected 50 questions from the sample of 550; 6 of these were in the first set of problematic questions, so we ended up with 94 questions in total.

Table 2 presents the recent performance evaluation of QUAID. The old version of QUAID is compared with the revised version of QUAID. Table 2 also contrasts a lower threshold (T = .33) with a higher threshold (T = 44) of problem scores. As the threshold gets higher, the greater extent to which expert judges believe there is a problem with a question. As the threshold increases, there automatically is a lower problem likelihood score; when averaging over the 5 question problems, the problem likelihood scores were .38 and .18 for the low versus high thresholds, respectively. Similarly, the *d'* scores generally increase as a function of higher thresholds (as do hit rates and false alarm rates). So when the experts have a stronger belief there is a problem with a question, the accuracy of QUAID shows a similar improvement.

The most interesting data contrasts the performance of the old versus the revised version of QUAID. We spent considerable effort improving the syntax component and that clearly paid off. The hit rates and *d'* scores increased dramatically for syntactic complexity. In the future, we plan on giving greater attention to working memory overload module, now that that there has been reasonable progress on syntactic complexity. This is because one aspect of working memory load consists of syntactic complexity. In contrast to the dramatic increases in the performance of the syntax module, there were modest gains in unfamiliar technical terms and vague/ambiguous noun-phrases. The vague and imprecise relative term component is almost finished, so improvements are not anticipated on that module.

There is some question of what performance index to maximize in our QUAID tool. We plan on having two versions of QUAID, one that maximizes hit rates and one that maximizes *d'* discrimination. If we maximize on hit rate, then QUAID will identify most of the problems, but at the cost higher false alarms. So QUAID will alert the survey methodologist that there might be a problem, but the survey methodologist will have to make frequent decisions that these potential problems should be dismissed. If we maximize on *d'* scores, then QUAID will be identifying problems less often, but the decisions will be more accurate. The use of the different versions will depend on the goals of the survey methodologist (i.e., completeness versus timeliness).

There is one fundamental problem with using the expert ratings as the gold standard of spotting problems with question interpretation. The experts have only moderate agreement on the identification of these problems (see Graesser et al., 2000) and they miss many of the subtle analyses of language, discourse, and world knowledge. Therefore, we need a more objective measure of identifying questions with particular problems. Our hope is that eye tracking data will provide a more objective measure. Therefore, we conducted a study on eye tracking during question answering. This is reported in the next section.

Table 2: Recent comparison of QUAID and human experts in detecting problems with questions

	Hit rate	False alarm rate	d' score	Problem likelihood
(1) Unfamiliar technical term				
Threshold = .33				
Old QUAID	.82	.44	1.06	.30
Revised QUAID	.96	.71	1.20	.30
Threshold = .44				
Old QUAID	.93	.49	1.50	.15
Revised QUAID	1.00	.75	1.64	.15
(2) Vague or imprecise relative term				
Threshold = .33				
Old QUAID	.77	.50	.74	.38
Revised QUAID	.77	.50	.74	.38
Threshold = .44				
Old QUAID	.90	.52	1.29	.22
Revised QUAID	.90	.52	1.29	.22
(3) Vague or ambiguous noun-phrase				
Threshold = .33				
Old QUAID	.88	.64	.82	.46
Revised QUAID	.90	.56	1.13	.46
Threshold = .44				
Old QUAID	1.00	.73	1.70	.08
Revised QUAID	1.00	.71	1.76	.08
(4) Complex syntax				
Threshold = .33				
Old QUAID	.28	.16	.42	.41
Revised QUAID	.62	.38	.62	.41
Threshold = .44				
Old QUAID	.39	.15	.76	.24
Revised QUAID	.91	.34	1.75	.24
(5) Working memory overload				
Threshold = .33				
Old QUAID	.40	.12	.90	.37
Revised QUAID	.40	.12	.90	.37
Threshold = .44				
Old QUAID	.63	.12	1.50	.20
Revised QUAID	.63	.12	1.50	.20

Eye Tracking While Answering Questions

The collection of eye tracking data provides a different method of diagnosing problematic questions with respect to question interpretation. Eye tracking patterns serve as a sensitive index of on-line comprehension processes. If a question is difficult to comprehend, then there should be a high density of multiple fixations on words and regressive eye movements. Words that are difficult to interpret should have long fixation times. We collected eye tracking data in order assess whether the problems identified by QUAID are manifested in eye movements and gaze durations.

We conducted a study on 9 college students who read and answered 69 questions selected from the corpus of 550 survey questions. The 69 questions included 45 problematic questions and 24 random questions. We had to exclude questions that were too long to fit on a computer screen. The eye tracking equipment was an Applied Science Laboratory Model 501 eye tracker with a head mounted device. Thus, the respondents could move their heads while reading and answering the questions.

During each trial, the participant advanced to the next question by hitting a bar in presence of a READY signal. Then the question appeared on the screen. The participant read the question and answered the question aloud. We recorded the eye tracking data while they read the question, audio recorded their answers, and videotaped the computer screen. The eye tracking portion of the study lasted 30 minutes, 10 for calibration of the eyes and 20 minutes for collecting data on the 69 questions. There were 6 different random orders of the questions. After collecting the eye tracking data, the participants completed a Wechsler Abbreviated Scale of Intelligence (Psychological Corporation, 1999) and an information sheet about demographic information and university training.

One index of comprehension difficulty is multiple fixations on a word. If comprehension runs smoothly, the reader would move ahead in a linear fashion, with only one eye fixation per word. However, there will be multiple fixations and regressive eye movements to the extent that there are problems interpreting words, noun-phrases, clauses, and sentences. The index of comprehension difficulty was therefore scored as number of eye fixations per word, given that there was at least one fixation on the word.

Table 3 shows this fixation frequency index for the content words of one of the questions. Content words include nouns, pronouns, adjectives, and main verbs. The function words and other minor words were not counted because they are known to have short fixation times. The fixation frequencies clearly increase as the readers progress further in the sentences, when the working memory load is higher and the syntactic complexity is more taxing. The mean fixation frequencies were 1.14, 1.44, 2.08, and 2.57 for the content words on lines 1, 2, 3, and 4, respectively. Table 4 shows that gaze durations on individual words show the same pattern. The mean daze durations (measured in milliseconds) are 225, 290, 397, and 633 milliseconds for lines 1, 2, 3, and 4, respectively.

We are currently analyzing fixation frequencies and gaze durations of the words in the 69 questions. The mean fixation frequency per content word (and mean gaze duration) should be significantly higher for the problematic questions than the nonproblematic questions. Moreover, gaze durations should be comparatively high for unfamiliar technical terms, unclear relative terms, and vague or ambiguous noun-phrases. Regressive eye movements should occur at points in the sentence when the syntactic complexity and/or working memory load are high. These predictions are currently being tested in our laboratory.

Table 3: Fixation frequencies for content words in an example question.

1.27 1.00 1.27 1.00
Do the people who do not live and eat

2.52 1.00 1.33 1.00 1.33
at your house have direct access from the

1.70 2.00 2.55
outside or through a common hallway to a

2.77 2.70 2.25
separate living quarter?

1.00 1.33 5.73 12.10 3.00
Yes; No; Refused; Don't know

Table 4: Gaze durations for content words in an example question.

310 190 220 180
Do the people who do not live and eat

500 240 290 200 220
at your house have direct access from the

210 400 580
outside or through a common hallway to a

760 490 650
separate living quarter?

120 290 880 2600 530
Yes; No; Refused; Don't know

References

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, CA: Jossey-Bass.

Graesser, A.C., Bommareddy, S., Swamer, S., & Golding, J.M. (1996). In N. Schwartz and 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., 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., Wiemer-Hastings, K., Kreuz, R., Wiemer-Hastings, P., & Marquis, K. (2000). QUAID: A questionnaire evaluation aid for survey methodologists. Behavior Research Methods, Instruments, and Computers, 32, 254-262.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (2000). The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices. Proceedings of the Section on Survey Research Methods of the American Statistical Association. (pp. 459-464).

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

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

Jurafsky, D., & Martin, J.H. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Prentice. .

Lessler, J.T., & Forsyth, B.H. (1996). A coding system for appraising questionnaires. In N. Schwartz and S. Sudman (Eds), Answering questions: Methodology for determining cognitive and communicative processes in survey research (pp. 259-292). 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 and Society, 63, 565-581.

Psychological Corporation (1999). Wechsler Abbreviated Scale of Intelligence. San Antonio, TX: Harcourt Brace.

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

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

Sirken, M.G., Hermann, D.J., Schechter, S., Schwarz, N., Tanur, J.M., & Tourangeau, R. (1999)(Eds.), 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, and R. Tourangeau (Eds.), Cognitive aspects of survey methodology: Building a bridge between disciplines. Washington, DC: National Academy of Sciences.

Willis, G.B., DeMaio, T.J., & Harris-Kojetin, B. (1999). Is the bandwagon headed to the methodological promisef land? Evaluating the validity of cognitive interviewing techniques. In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J.M. Tanur, & R. Tourangeau (Eds.), Cognition and survey methods research (pp. 133-153). New York: Wiley.

Discussion: "A Computer Tool to Improve Questionnaire Design"

Theresa J. DeMaio
U. S. Census Bureau

The QUAID model is certainly a computational challenge, and interesting from the point of view of cognitive linguistics. And from the perspective of a survey methodologist, it has the potential to be a useful diagnostic tool. I'd like to focus my comments today on three aspects of the paper: the choice of human experts for comparison of problem detection with the QUAID computer model, the results of the comparison with human experts, and the search for a new gold standard.

Problem Detection by the Computer Model and Human Experts

In developing the computer model, Graesser and his colleagues have defined the truth as the correct identification of problems by human experts. In the results they report here, and in results they have reported in previous papers (Graesser, Kennedy, Wiemer-Hastings and Ottati, 1999; Graesser, Wiemer-Hastings, Wiemer-Hastings, and Kreuz, 2000), they compare the performance of the computer model in identifying questionnaire problems of specific types against the performance of human judgement. Their judges had graduate degrees in a field that investigated the mechanisms of language, discourse, and/or cognition. I think this would be a relevant criterion for the judges if the tool was for a purpose related to these disciplines. But since this is a tool to be used by survey methodologists, it would be much more appropriate if the results of the computer model were compared to evaluations of the same questions by questionnaire design experts in the field of survey methodology, who have familiarity with and expertise in identifying problems with survey questions experienced by respondents.

Their use of language experts makes an implicit assumption that the use of language in survey questions is the same as all other questions, and I think we know that this is not necessarily the case. "How many people live in your house or apartment?" may be interpreted differently when a survey interviewer talks to a recent illegal immigrant than when the immigrant is speaking to a friend or relative. Perhaps the results of the comparison of the computer with language experts would be the same as a comparison with questionnaire design experts – I wouldn't want to make predictions about the extent of any differences – but I would definitely feel more comfortable about the utility of QUAID as a diagnostic tool for surveys if I could see some data about how it compares to survey methodologists' evaluations of survey questions.

Results of the Comparison with Human Experts

I view QUAID as a preliminary questionnaire design tool, one that would be useful in identifying major problems in draft questionnaires during the initial questionnaire development process. As such, I don't see it as a competitor to either verbal protocols from respondents during think-aloud interviews or coding of the interaction between respondents and interviewers during field interviews. To my mind, its use would precede either of these two methods. It is more similar to an expert review and cognitive appraisal methods. So a questionnaire designer might want to make a choice between QUAID, expert reviews, or questionnaire appraisals (Lessler & Forsyth, 1996; Willis & Lessler, 1999) in the early stages of questionnaire development.

In this context, I was interested in the last column of Table 1, in which the authors note the problem likelihood, that is, the likelihood that each of the five problems of interest was identified in a question. These scores ranged from .04 (which means that a problem of this kind was detected in 1 out of 25 questions) to .10 (which means that a problem was detected in 1 out of 10 questions). Summed together, the problem likelihood that any of these five problems would be identified is .38, or 4 out of every ten questions. This is an upper bound, since more than one of these problems could apply to a single question. These scores seem very low to me. The questions came from 11 survey questionnaires conducted by the Census Bureau, and I'd like to think that Census Bureau surveys are this good, but I don't really believe it.

Research has been conducted on the expert review methodology and the questionnaire appraisal system, which I said I view as QUAID's main competitors, and these methods identify a much higher percentage of questionnaire problems. In 1991, Presser and Blair (1994) conducted experimental research in which expert reviews were conducted, along with other pretest methods. Two independent expert reviews were conducted on a 140-item questionnaire. One of the expert reviews identified 182 problems, and the other identified 140 problems.

More recently, Jennifer Rothgeb and her colleagues (Rothgeb, Willis, & Forsyth, 2001) presented a paper at AAPOR last month in which they compared expert reviews with questionnaire appraisals. For an 83-item questionnaire that was rated on a problem scale of 0 to 3, the expert review yielded a mean problem score of 1.55 (that is, items were found to be problematic half the time) and the questionnaire appraisal yielded a mean problem score of 2.93. In other words, almost all the time, items were found to be problematic.

None of these comparisons are exactly equivalent, but there is enough similarity in the objectives and methods that I would expect a higher problem yield from QUAID. The greatest portion of the problems identified in both these research efforts dealt with question meaning, and four out of the five problems included in QUAID deal with question meaning as well. One difference between the QUAID results and the other research is that survey experts conducted the expert reviews and the questionnaire appraisals, while this was not the case for the gold standard for the QUAID evaluation. Perhaps there is some unique expertise that questionnaire design experts bring to bear when evaluating survey questions that is different than the experience of linguists.

Since questionnaire appraisals and expert reviews identify more problems than cognitive interviews or behavior coding (Presser and Blair, 1994; Rothgeb et al, 2001), it seems that the knowledge of the survey experts leads them to identify potential problems that are not evidenced by respondents themselves. This is the equivalent of the False Alarm rate calculated by Graesser and his coauthors. I am not bothered by that as much as I am by the relatively low problem likelihood. I would urge them to focus their attempts to improve QUAID in that area. My perspective on this comes from my view that this is a tool for the initial stages of questionnaire development. Suspected problems that don't turn out to be serious can be addressed, but serious problems that never get detected can jeopardize a data collection effort.

What Should the Gold Standard Be?

Graesser and his colleagues have a question about what the gold standard should be for comparing the QUAID results against. They think that judgements of experts in language, cognition and world knowledge are problematic because they are not stable across multiple experts. They also consider getting feedback from respondents, but find this to be lacking because their judgements can be "insensitive to problems that allegedly exist." So they are moving on to eye tracking as an objective measure of on-line comprehension processes.

Eye tracking is an interesting notion. Cleo Redline, one of my colleagues at the Census Bureau, is investigating its use as a vehicle to evaluate visually administered instruments, and she also recently presented a paper at AAPOR (Redline and Lankford, 2001). Her research to date has focused on skip instructions on paper questionnaires, and she is planning to expand to studies of response to automated questionnaires and websites. Her concentration is on navigational issues, and keeping track of respondents' eye movements as they find their way through a questionnaire or a website makes intuitive sense.

I wonder, however, whether this technique can really be an objective measure of comprehension, as Graesser asserts. It seems to me that a big assumption must be made to state that multiple fixations on a word is an indicator of comprehension difficulty. That might be the case, of course, but it also could be that the respondent is absorbing the content of the major concepts of the question without difficulty. Furthermore, if the objective of this gold standard is to spot problems of the five types that QUAID can reliably detect, it is not clear how the eye tracking methodology can achieve this. I think some demonstration of the validity of this criterion measure is necessary before it is used in this way.

My view is that, since the stated objective of QUAID is to be "a computer tool that assists survey methodologists who want to improve the wording, syntax, and semantics of questions on surveys and questionnaires," the perspective of the survey methodologist is a logical place to start in assessing how well the computer tool works in terms of meeting its objective. QUAID would be useful if it could provide an easy, automated means for providing the same types of information about questionnaire problems that can already be obtained with a lot more effort through other means. There are probably other ways to look at the issue, and I would be open to other standards if they could improve on the information that is already available, but from my perspective that is the minimum standard that would make QUAID a viable method for testing survey questions.

In conclusion, I would encourage the authors to continue their development of the QUAID program and make it into a useful tool for questionnaire designers.

References

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

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., and Kreuz, R. (2000) "The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices," Proceedings of the American Statistical Association (Survey Research Methods Section), Alexandria, VA: American Statistical Association, forthcoming.

Lessler, J.T., and Forsyth, B. H. (1996) A coding system for appraising questionnaires. In N. Schwarz and S. Sudman (Eds.), Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research (pp. 389-402). San Francisco: Jossey-Bass.

Presser, S. and Blair, J. (1994) Survey pretesting: Do different methods produce different results? In P.V. Marsden (ed.), Sociological Methodology, Vol. 24, Washington, DC: American Sociological Association, pp. 73-104.

Redline, C.D. and Lankford, C.P. (2001) "Eye movement analysis: A new tool for evaluating the design of visually administered instruments (paper and web), paper prepared for presentation at the Annual Meetings of the American Association for Public Opinion Research, Montreal, Canada, May 2001.

Rothgeb, J.R., Willis, G.B., and Forsyth, B. (2001) "Questionnaire pretesting methods: Do different techniques and different organizations produce similar results?", paper prepared for presentation at the Annual Meetings of the American Association for Public Opinion Research, Montreal, Canada, May 2001.

Willis, G.B. and Lessler, J.T. (1999) The question appraisal system: A guide for systematically evaluating survey question wording. Final Report submitted to the Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion. Research Triangle Institute.