THE GOLD STANDARD OF QUESTION QUALITY ON SURVEYS: EXPERTS, COMPUTER TOOLS, VERSUS STATISTICAL INDICES

Art Graesser, University of Memphis, Katja Wiemer-Hastings, Northern Illinois University,
Peter Wiemer-Hastings, University of Edinburgh, Roger Kreuz ,University of Memphis
Art Graesser, Department of Psychology, University of Memphis, Campus Box 526400, Memphis, TN 38152

Key Words: Survey questions, question quality

Abstract

QUAID (Question Understanding Aid) is a computer tool that assists survey methodologists who want to improve the wording, syntax, and semantics of questions on surveys. QUAID produces a list of potential problems with each question, including (1) unfamiliar technical term, (2) vague or imprecise relative term, (3) vague or ambiguous noun-phrase, (4) complex syntax, (5) working memory overload, and (6) misleading or incorrect presupposition. We assessed the incidence of these problems on a corpus of 11 surveys provided by the US Census Bureau. We are currently facing the challenge of assessing the validity of QUAID's critique of questions. The collection of multiple methods and measures is a lofty virtue, but this approach offers no principled foundation for rectifying discrepancies in performance measures. This presentation addresses the challenges of performance evaluation when there is no defensible gold standard for question quality.

Introduction

A good questionnaire contains questions that elicit valid and reliable answers from respondents in a short amount of time. This is the lofty goal that all survey methodologists want to achieve, but the routes to getting there can radically differ among researchers. One traditional approach has been to perform sophisticated statistical analyses that isolate different components of systematic variance and measurement error, as in the case of test-retest reliability assessments and item response theory (Groves, 1989). Another approach is to pretest questionnaires by having experts identify particular problems with questions (Lessler & Forsyth, 1996), by collecting verbal protocols from respondents as they answer questions (Willis, DeMaio, & Harris-Kojetin, 1999), or by observing behaviors that suggest that the respondents are struggling with particular questions (Fowler & Cannell, 1996). A rather different approach, one that is pursued in the present study, is to conduct interdisciplinary research that dissects the cognitive mechanisms of question answering, and then maps these mechanisms onto indices of performance (such as respondents' answers, expert evaluations of question flaws). Our approach integrates the fields of computer science, computational linguistics, cognitive science, and survey methodology. More specifically, we have developed a computer program that critiques questions on different types of cognitive problems.

Researchers in the field called CASM (Cognitive Aspects of Survey Methodology) have proposed models that dissect different stages question-answering (Jobe & Mingay, 1991; Lessler & Sirken, 1985; Schwartz & Sudman, 1996; Tourangeau, 1984; Sirken, Hermann, Schechter, Schwarz, Tanur, & Tourangeau, 1999), 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 & Kalsbeck, 1993; Schober & Conrad, 1997). 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 our recent 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 critique each question on potential comprehension difficulties at various levels of language, discourse, and cognition. More specifically, the critique identifies words that are unfamiliar to most respondents, vague predicates (verbs, adjectives, adverbs), ambiguous noun-phrases, questions with complex syntax, questions that overload working memory, and questions with incorrect or misleading presuppositions (Graesser, K. Wiemer-Hastings, Kreuz, & P. Wiemer-Hastings, 2000; K. Wiemer-Hastings, P. Wiemer-Hastings, Rajan, Graesser, Kreuz, & Karnavat, 2000). The identification of such problems should be useful to the survey methodologist if the computer tool can accurately flag 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. The computer aid would be even more useful if it also offered suggestions about the revision of problematic questions, but question revision was beyond the scope of QUAID.

The development of a computer aid, such as QUAID, does not need to be perfect in order to be useful. First, QUAID need not be exhaustive in its coverage of potential problems. It could offer advice about those components for which it can deliver accurate feedback. Second, QUAID need not perfectly match the judgments of human experts. Indeed, some components of question answering are so complex, technical, or subtle that they are invisible to the unassisted human eye, even the eye of an expert in questionnaire design or the eye of an accomplished computational linguist. For example, it would be impossible for these experts to catch all of the problems in sentence syntax and working memory load. Very few experts would have the time and patience to dissect each question at such a fine grain. Third, QUAID need not be perfectly accurate. It could be useful even if it produced occasional errors in diagnosis, such as identifying a misleading presupposition that really does not pose a problem to the respondent. Such faulty diagnoses would sometimes be eliminated when the human experts scrutinize the computer output. That is, 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.

Although an imperfect QUAID would have some utility, we cannot avoid the worry of providing adequate tests of its performance. Unfortunately, however, we have not yet found a good gold standard for evaluating question quality. The judgments of experts who critique questions might be faulty, so they are not the perfect gold standard. Answers to survey questions are rarely assessed on the dimension of validity (see Schober & Conrad, 1997), so such a gold standard is absent. Answers to survey questions are frequently assessed on reliability, but reliability alone is a seriously inadequate gold standard. In an ideal world, we would have access to the correct answers to questions for particular respondents, and then we would observe how well the articulated questions elicit responses that recover the correct answers. Questions with flaws should produce answers that deviate from the correct answers. Unfortunately, such analyses are virtually nonexistent in the survey world, or are prohibitively expensive to collect.

In this paper, we will not present the perfect gold standard for assessing questions on question quality. But what we do plan to do is to report some data on the judgments of experts, the output of QUAID, and comparisons between human experts and QUAID. The results of our analyses will underscore the difficulty of arriving a gold standard for question quality.

Common Problems with Questions

Graesser's previous research has identified 12 problems with questions that frequently occur in surveys (Graesser, Bommareddy, Swamer, & Golding, 1996; Graesser, Kennedy, Wiemer-Hastings, & Ottati, 1999). Many of these problems have been incorporated in various analytical coding schemes of survey methodologists, as discussed in these publications. At present, QUAID can handle 6 of these problems with some degree of correspondence with human experts, so we will focus on these 6 problems in the present paper. These problems are presented 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 well-formed formed syntactically.
(5) Working memory overload. Words, phrases, or lauses impose a high load on immediate memory.
(6) Misleading or incorrect presupposition. The truth value of a presupposed proposition is false or /inapplicable.

It is beyond the scope of this presentation to precisely define these six problems and to specify how the QUAID tool identifies them (see Graesser et al., 2000). Instead, we will illustrate the tool with a critique of one example question. This question appeared 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 many of the above 6 problems. It imposes working memory overload in at least two ways. First, the first noun-phrase is long and cumbersome; the respondent is forced to keep track of a long list of 6 or more family members. Second, the respondent is asked whether each of these family members has had a heart attack or myocardial infarction so there is a 6 x 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 working memory that has limited capacity. 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.

Can Respondents Identify the Problems with Questions?

One value of QUAID is that it would help the survey methodologist identify the problems with questions that would be missed by respondents during pretesting. Survey researchers have frequently advocated the collection of think aloud protocols from a sample of respondents during pretesting (Bickart & Felcher, 1991; Jobe & Mingay, 1991; Lessler & Sirken, 1985; Willis, Royston, & Bercini, 1991). Graesser et al. (1999) reported, however, that most of the six problems are completely missed by respondents who critique a survey during pretesting. The only problems that adult respondents can identify with any modicum of reliability are problems 1 (unfamiliar technical term) and 3 (vague or ambiguous noun-phrase).

Can Experts Identify the Problems with Questions?

Graesser et al. (1999) raised concerns that expert survey methodologists might miss many of the problems if they are not adequately trained in linguistics, discourse, and cognition. Indeed, it is conceivable that experts will miss problems even if they are highly trained in these areas. Graesser et al. (2000) conducted a study that assessed how well experts can identify the six problems. Experts evaluated a corpus of 550 questions on the six problems (3300 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 any of the 6 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 at the US Bureau of Census.

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.

Scoring the Experts' Ratings of Problems

Table 1 presents a summary of the problem evaluation ratings by the experts. Three measures are reported in the table. The **problem incidence** is the proportion of questions in which at least 1 expert had a rating of 3 or 4. The **problem score** is a value that varies from 0 to 1: (sum of expert ratings - 3) / 9. The **interjudge reliability score** is the proportion of agreements among pairs of experts (1-2 versus 3-4 split). A number of conclusions can be drawn from the data in Table 1. First, the six problems were not rare occurrences in the corpus of questions, even though the questions had been pretested and scrutinized by personnel at the US Census Bureau. Second, the 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 .85, which is rather modest. Other measures of reliability (i.e., correlations among ratings, 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 3 judges weighted the various criteria differently when they made the judgments. Second, the judges may have experienced some problems of fatigue while making thousands of decisions. Third, the detection of some problems is 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.

Table 1: Problems identified by human experts.

	Problem incidence	Problem score	Interjudge reliability
(1) Unfamiliar technical term	.238	.131	.83
(2) Vague or imprecise relative term	.403	.184	.73
(3) Vague or ambiguous noun-phrase	.486	.184	.69
(4) Complex syntax	.328	.151	.77
(5) Working memory overload	.274	.147	.81
(6) Misleading presupposition	.186	.007	.85

QUAID (Question Understanding Aid)

This section briefly describes the QUAID computer tool. QUAID has interface options that correspond to the 6 problems with questions. The computer user can turn each of the 6 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 critiques the question on the 6 different components (or as many of the 6 that the user desires). QUAID currently runs on a Pentium computer with a Linux operating system. The software was developed in the LISP programming language. QUAID will be available to the public on the Web starting in January, 2001.

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 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 6 categories, QUAID would print out 6 short summary messages that point out the particular problems (as illustrated earlier). In addition to this short feedback, there is a HELP facility that defines each problem more completely and gives examples of particular problems.

It is beyond the scope of this paper to describe the mechanisms that QUAID used to critique the questions. It suffices to say that there was a combination of empirical tests and theoretical developments in computational linguistics.

Comparison of QUAID and Human Experts

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 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. For the present purposes, we will report data at threshold values that yielded good performance scores.

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. 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)

Table 2 presents the different performance measures for the 6 categories of problems with questions. These include the hit rates, false alarm rates, d' scores, precision scores and problem likelihood scores.

A number of conclusions are supported by the data in Table 2. First, the QUAID tool was able to discriminate problematic questions because the *d'* scores were above zero. All of these d' scores are statistically significant when we analyzed frequency tables and computed chi-squares. That is, a chi-square test of association was computed on each 2 by 2 frequency table that includes the frequency of hits, misses, false alarms, and correct rejections. 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. 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 up problematic questions. The recall scores and precision scores, measures that are standard in computational linguistic, are compatible with these conclusions. That is, there is a tradeoff between recall scores and precision scores. For the first 3 problem categories, recall scores are more impressive than the precision scores; for problems 4, 5, and 6, recall scores are less impressive than the precision scores. These analyses provide some informative guidance in modifying QUAID in the future.

So What Should be the Gold Standard for Question Quality?

The persistent question remains as to what the appropriate gold standard should be. Feedback from respondents is problematic because their judgments are sometimes insensitive to problems that allegedly exit. The judgments of experts in language, cognition, and world knowledge are problematic because the experts have only a modest level of agreement. The modest interjudge reliability scores can perhaps be explained by the variability in their research background, to the subtlety of the theoretical components, or to fatigue. The validity of the computer output from QUAID is indeterminate because there is no criterion reference.

In the future, we plan on pursuing two approaches to testing the accuracy of QUAID. First, we plan on administering the surveys to respondents and measuring the incidence of clarification questions in a conversational interview format (Schober & Conrad, 1997). These are questions that the respondents ask in order to clarify the meaning of the question (e.g., What do you mean by infarction?). The incidence of clarification questions should be positively correlated with the problems identified by QUAID.

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

	Hit rate (recall)	False alarm rate	d' score	Precision score	Problem likelihood
(1) Unfamiliar technical term	.86	.41	1.31	.17	.09
(2) Vague or imprecise relative term	.94	.53	1.48	.17	.10
(3) Vague or ambiguous noun-phrase	.95	.61	1.37	.06	.04
(4) Complex syntax	.29	.03	1.33	.40	.07
(5) Working memory overload	.29	.04	1.20	.34	.08
(6) Misleading presupposition	.62	.31	.51	.74	--[a]

a Misleading presuppositions were analyzed on a restricted subset of the data because of the extremely low incidence score.

Second, we plan on assessing the test-retest reliability of questions that are prepared in three conditions: (1) original questions on survey, (2) questions revised by survey methodologists, and (3) questions revised by survey methodologists who use QUAID. QUAID will be validated to the extent that condition 3 yields higher test-retest reliability scores than conditions 1 and 2. Nevertheless, it is important to be cautious and acknowledge that these two approaches alone do not provide a precise gold standard for assessing question quality. So we remain in the hunt for the ideal gold standard.

References

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

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, CA: Jossey-Bass.

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

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. (2000). QUAID: A questionnaire evaluation aid for survey methodologists. Behavior Research Methods, Instruments, and Computers, 32, 254-262.

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.

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. 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.

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.

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.

Wiemer-Hastings, K., Wiemer-Hastings, P., Rajan, S., Graesser, A.C., Kreuz, R.J., & Karnavat, A. (2000). DP-A detector for presuppositions in survey questions. Proceedings of the Sixth Applied Natural Language Processing conference.

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.

Willis, G.B., DeMaio, T.J., & Harris-Kojetin, B. (1999). Is the bandwagon headed to the methodological promised 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.