From Impressions to Data: Increasing the Objectivity of Cognitive Interviews

Frederick Conrad, *Bureau of Labor Statistics* and Johnny Blair, *University of Maryland*[1]
Johnny Blair, Survey Research Center, 1103 Art-Sociology Bldg, College Park, MD 20742

Key Words:

I. Introduction

The most tangible result of the dialogue between survey methods research and cognitive psychology is the widespread use of think aloud methods for pre-testing questionnaires -- so-called cognitive interviews. In thinking aloud, people verbally report their mental activity while they are performing or immediately after they perform an experimental task (answering a survey question in the case of cognitive interviews). However the way the survey methods community has adapted these techniques may compromise their value for improving questionnaires. In particular, psychologists developed the methods out of a generally accepted theory of how people solve problems (Ericsson & Simon, 1992). This constrains the inferences about the data that researchers are licensed to make. In addition, the procedures that psychologists have developed for collecting and analyzing the data are quite systematic. In contrast, cognitive interviews are not especially grounded in theory, their administration varies widely among practitioners, and the way they are analyzed is often based on the practitioner's impressions.

This paper reports a method for analyzing think aloud data from cognitive interviews that requires coders to systematically consider a broad set of criteria in evaluating the verbal report for each question in a questionnaire. The crux of the method is a taxonomy of respondent problems which the analyst uses to classify verbal reports that seem to indicate trouble with a question. The problem categories are derived, in part, from a theory of survey responding to which many practitioners subscribe. By identifying the response stage at which a problem is likely to have occurred, certain solutions to the problem become more promising while others become less plausible.

In addition to the respondents' verbal reports, the analyst is provided with a relatively formal statement of the author's intentions when creating particular questions. By comparing the content of the verbal reports to the way the author intended the question to be answered, the analyst may identify problems that would otherwise have gone unnoticed and may also realize that behavior which seems to signal a problem is actually consistent with the question's design. The approach is intended to be usable by staff members with a range of experience and certainly should not require an advanced degree in psychology.

II. Toward more systematic cognitive interview procedures

There is a logical distinction between *collecting* think aloud data and *analyzing* them, although in cognitive interviews, this distinction is often blurred. Our concern with current cognitive interviewing practices extends to the data collection process, though our focus here is exclusively on analyzing respondents' verbal reports. We have temporally separated collection and analysis so that the analyst is free from the demands of conducting the interview and can devote full attention to the content of the reports. What's more, the analyst can exhaustively and repeatedly consider criteria about possible problems.

Our procedure requires analysts to consider an explicit set of criteria, in its entirety, when evaluating a verbal report for a particular question. Because these criteria are standard across both interviews and analysts, analysts are likely to identify respondent problems more reliably and objectively than when the criteria are unstated and developed by individual cognitive interviewers -- as is typical now. We have developed such a set of criteria and expressed it as a taxonomy of possible problems. The taxonomy is based on a generic theory of the response process, and so by assigning a problem to the taxonomy, one describes the information processing context in which the problem arises. The reasoning behind this is that such a description is a necessary step in resolving the problem and so the taxonomy will both help identify problems and promote solutions.

III. The Respondent Problem Matrix.

The taxonomy of possible respondent problems is represented as a matrix with three columns and five rows (see Table 1). The columns represent the major stages that a respondent is likely to pass through en route to answering a question. The rows correspond to five problem classes that, based on our experience, entail most of the problems for which respondents provide evidence in their think aloud protocols. The matrix representation stems from the idea that the different classes of problems can occur at each of the three response stages. Thus, each cell produced by crossing the rows and columns defines a specific problem category. The matrix in Table 1 contains 15 cells. We could have made finer distinctions within the rows

[1] The opinions expressed here are those of the authors and not necessarily those of the authors' institutions.

and columns creating more problem categories; however, this number of categories and their relatively coarse granularity seemed appropriate for use by relatively junior staff without extensive research experience, and appropriate for problem identification as opposed to hypothesis testing.

Variations on the Generic Response Model and Its Use

We accept the four stage response model proposed both by Cannell and his associates (e.g. Oksenberg & Cannell, 1977) and Tourangeau and his associates (Tourangeau, 1984; Tourangeau & Rasinski, 1988) as a kind of generic response theory which is cast at a high enough level that it must be, at least roughly, accurate. This model is specified at about the same level of detail as the view of problem solving that underlies the use of verbal reports in psychology (Ericsson & Simon, 1992). Just as the analysis of verbal reports of problem solving is guided by that theoretical perspective, so our analysis of survey response is guided by the generic response model.

Other researchers have used the four stage response model for classifying respondent problems. Lessler and Forsythe (Forsythe, Lessler & Hubbard, 1992; Lessler & Forsythe, 1996) have structured a taxonomy of problems on the basis of response stages. Like our approach, theirs is a general taxonomy, applicable to most surveys. Theirs differs from ours in that it is designed for experts to directly appraise a questionnaire rather than for coders to classify respondents' verbal reports. Under Lessler and Forsythe's approach, the expert uses the taxonomy as a set of criteria to consider about each question. This can be done without the time and expense involved in laboratory testing of respondents. As with methods in other domains that rely on expert judgment rather than behavioral data (see, for example Nielsen, 1994 in the domain of evaluating software usability) there is no empirical evidence that the experts' judgments predict respondents' actual experience. If one has the time and resources to collect laboratory data on respondents' thinking, we believe they most accurately predict the kinds of problems likely to be encountered by actual respondents.

Bickart and Felcher (1996) have developed a taxonomy for coding verbal protocols that is also based on a four stage response model. Bickart and Felcher's approach differs from ours in several ways: First, theirs is specialized for verbal reports about answering behavioral frequency questions; ours is intended to be usable for various types of questions in various questionnaires. While specialized coding schemes, by definition, need to be developed for each new survey or study, ours is "ready-to-use" for each new study. Second, their taxonomy is designed to classify the *strategies* that respondents use in order to address detailed analytical questions; ours is aimed at the *problems* they experience when answering questions.

In our use of the generic response model, we are assuming that respondents execute the stages of the response process in a fixed sequence, though we recognize that stages can overlap: One stage may still be underway when the next is initiated. Nevertheless, the processes that define a stage are quite distinct and so if a respondent provides verbal evidence of a problem, it is usually possible to infer that it originated in one of the following stages: (1) understanding the question and the implied task, (2) performing the primary task, and (3) mapping the results of that task to the response categories presented in the question.

While the generally accepted response model has four stages, our model has three. This is because verbal reports are not sensitive to all of the distinctions that are implied by the four stage model. In particular, an analyst cannot distinguish between retrieving information from long term memory on the one hand and evaluating what has been retrieved on the other: Verbal reports are based on the content of working memory and not the retrieval operations that transfer information there in the first place (Ericsson & Simon, 1992). For this reason, we have combined the retrieval and judgment stages proposed by Cannel and Tourangeau into a single stage -- performing the primary task.

Our version of the model includes two additional assumptions in order to account for different types of common problems that can be indicated in verbal reports. First, in order for the response process to proceed smoothly, the information produced at one stage must serve as adequate input for the next stage. The input to the first stage is the words which comprise the question, including the response categories; the understanding that is produced at this stage serves as input to the task performance stage; the information that is yielded by performing the task, serves as the input to the response formatting stage; the output from the response formatting is articulated or otherwise indicated by the respondent.

This is relevant to diagnosing problems because the content of a verbal report can suggest the problem occurs at one stage when an "adequate input" analysis indicates it actually has its roots in another stage. For example, if the respondent's protocol indicates she understands the question and implied task (stage 1) then she has the necessary information to perform the primary task (stage 2). Any problems in her protocol will have their source at some point after understanding. However, if she just cannot grasp what she is being asked to do, she has not derived the necessary information to begin the second stage. Her problem lies in understanding and she recognizes her difficulty. A more complicated case, and one where the analysis of inputs is instructive, is the one in which she believes she has correctly interpreted the question and task when, in fact, she has not. Under these circumstances, she lacks the requisite information

to begin the performance stage but may still perform some task – it just happens to be the wrong task. While this might appear to be a performance problem, when viewed in these terms, it is clearly an understanding problem.

The second assumption extends the generic model by allowing respondents to re-execute previous stages. The response process advances sequentially through the three stages when it works flawlessly, but when the respondent has trouble and is aware of it, he may re-start the sequence at the point he believes his error or difficulty occurred. The evidence for this might be an explicit request for the question to be re-read, or for the definition of a term. Alternatively, it might be the respondent's attempt to re-represent the task to himself, or reason about the task based on the content other questions ("I have answered this question about my *occupation* by giving them my *job title*; they already asked me about my *job title* so the current question about my *occupation* must have to do with my duties") [2].

Response Stages and Problem Classification

Understanding. We view a survey question as a set of instructions to the respondent about the task he or she is to perform. This means that understanding a question involves both determining what information is being requested (a literal interpretation of the question) and recognizing an unstated directive about how that information is to be provided (what procedure the respondent is to use in order to satisfy the request). For example, in order to understand the question "During the past six months, how many times have you been to the doctor?" a respondent must represent the utterance as a request for the number of doctor visits over a particular time interval (a literal interpretation of the question) as well as an instruction, for example, to count all remembered doctor visits or an instruction to report a known rate of doctor visits, etc., or a more general instruction about the set of acceptable procedures for producing the requested information.

Respondents come to understand these often implicit task instructions through the same mental processes that they use to recognize indirect requests in ordinary conversation (e.g. Clark & Bly, 1995; Levinson, 1983; Searle, 1975). These processes work remarkably well in everyday language but listeners occasionally make inferences that differ from what speakers intend and sometimes fail to make an intended inference (e.g. Clark, 1979). Ideally, the questionnaire author has considered what response process is most likely to be implied by the question, and has chosen wording to encourage the desired process; whether or not the author has actually given this any thought, respondents will try to infer the process they are "intended" to use[3].

If the respondent and author differ in their understanding of the task, the respondent may provide data that are entirely inaccurate from the author's perspective – though this is likely to go undetected. Note that it is possible for the respondent to understand all the words in the question as they were intended and still incorrectly interpret the task. In any event, evidence that the respondent has misinterpreted the task will be apparent in the verbal reports produced during the second stage, primary task performance.

Performing the primary task. Assuming the respondent has managed to interpret the instructions, it becomes possible to execute the second stage, that is, to perform the primary task. By "primary task" we mean the particular mental operations used to produce the "raw data" on which the response is ultimately based. These data are then converted into an acceptable response format, which is a secondary task, and the third stage of our model. These data can be a collection of autobiographical events used to answer a frequency question, a retrieved or computed opinion, facts about ones own circumstances like the number of rooms in ones homes or the highest level of schooling achieved, facts about the world like "people do not use air conditioners in the winter" to support an inference about ones utility expenditures, and so on.

The primary task varies extensively depending on the question and the associated task but the kinds of processes required to answer most questions are retrieval, comparison, deduction, mental arithmetic and evaluation, among others. It is possible that the primary task will involve combinations of various processes. For a problem to be associated with this stage, the respondent must be trying to perform the intended task, but finding it difficult or impossible to execute the required processes. For example, a question may require a comparison of two quantities that are expressed in non-comparable units, "Which has more fiber, an apple or a cup of apple juice?".

Response Formatting. Assuming the respondent is able to perform the primary task, it is still possible he will have problems producing an acceptable response because the data yielded by the primary task processes do not easily map to the explicit response options. Suppose the respondent is asked how many compact disks he owns. He performs the primary task

[2] Allowing respondents to use their knowledge of other questions diverges from earlier versions of the generic model which were defined for individual questions in isolation.

[3] This characterization assumes an ideally cooperative respondent. In practice, respondents may be more likely to treat the task as the least demanding way to produce a plausible answer (Krosnick, 1991).

and the result turns out to be 46. The response categories are "very few", "an average amount", and "quite a few". The respondent does not know how to map "46" the available options.

Note, in the above example, the respondent knows the meanings of the words in the response options. In contrast, a respondent who does not know what the words in a response option mean is considered to have an understanding problem, not a response formatting problem because the response options are considered part of the question. Suppose the respondent is asked to check any skills that his job requires and is presented with a list of skills preceded by check boxes. One of the options is "spatial abilities" and he simply doesn't know this phrase. By our view, he has not succeeded in interpreting the literal question. It may be that if he knew what the phrase meant he would have no trouble mapping the information he has retrieved about his job to this category.

Problem Classes

The rows in the matrix correspond to five problem classes that, based on our experience, entail most of the problems for which respondents provide evidence in their think aloud protocols: (1) lexical problems, (2) inclusion/exclusion problems, (3) temporal problems, (4) logical problems, and (5) computational problems. In order to make the set of problem classes exhaustive, we treat the computational problem class, in part, as a residual category. The fifteen categories that result from crossing the rows and columns, and the way they are to be used in problem coding, are discussed in the coding instructions in the appendix. An overview follows.

Lexical problems. The first of these classes, lexical problems, has to do with not knowing the meanings of words or how to use them. What we have in mind by *meaning* is the "core" or "central" meaning of a word or phrase, not the subtleties of its scope. So problems like not knowing what is meant by a word like "nitrogen" or "spatial" in "spatial abilities"; they may not be familiar with idioms like a "the lion's share"; and they may be familiar with the meanings of a pair of words but do not understand their combination in the question, such as "medical purchases". These are all considered lexical/understanding problems. The apples and apple juice example above could be a lexical/task performance problem in that the respondent may know what these terms mean, but she cannot compare their fiber content because her understanding of apple juice does not include information about fiber and it does include information about the number of apples required to produce a fixed amount of apple juice.

Inclusion/exclusion problems. The second problem class, inclusion/exclusion problems, also involves word meanings but the problem lies in determining whether certain concepts are to be considered within the scope of a word in the question. For example, assume a respondent understands the phrase "religious groups" and can easily include items that are typical of the category like Catholics or Muslims. The respondent has trouble knowing whether to include or exclude a group like the Branch Davidians, which, if included, would certainly be less typical than Catholics or Muslims. We consider this to be an inclusion/exclusion/task performance problem because the respondent understands the conventional sense of "religious groups" but has trouble using it to classify certain instances that come to mind.

An example of an inclusion/exclusion/response formatting problem involves using a response option that was not explicitly provided such as "7.5" when the legal points on the response scale are presented as whole numbers. One interpretation is that the respondent has supplemented the set of response options because the whole numbers in the scale map ambiguously to a concept the respondent needs to quantify.

Temporal problems. Temporal problems involve the time period to which the question applies or the amount of time spent on an activity described in the question. Temporal problems are often a special case of lexical problems -- trouble grasping the meaning of temporal terms or using them -- but due to the prevalence of questions involving time periods, we have dedicated an entire problem class to difficulties with concepts of time. A respondent would have a temporal/understanding problem if he interpreted the phrase "in the last year" to mean "in the previous calendar year" instead of "in the last 12 months" as was intended. As an example of a temporal/task performance problem, imagine that a question involves the phrase "the current month" but because the interview occurs early in a new month, the respondent forgets about the change of month and considers the phrase to refer to what is actually the previous month. This is a performance and not an understanding issue because the respondent perfectly well understands the phrase "the current month" but assigns it an incorrect reference.

Logical problems. There are several types of problems in this category (which can occur at any stage). One type involves the devices used to connect concepts: logical connectives like "and" and "or", and other devices such as negation and complementarity (e.g. "infectious diseases other than hepatitis"). Consider the following logical problem with question and task understanding. "In the last week have you purchased or had expenses for meats and poultry." The phrase "meats and poultry" is intended to describe a category of foods and the respondent is intended to answer "yes" if he has purchased any items from that category, whether a meat product or a poultry product. However, the respondent

interprets the question as an instruction to respond "yes" if he has purchased both meat and poultry products. This particular problem involves understanding the task.

Another type of logical problem involves contradictions and tautologies. For example, "Do you experience freak accidents rarely, sometimes or often?" By definition freak accidents happen rarely, so the options are not logical. This problem could have its effect in the first stage by preventing the respondent from understanding the task; it could also have its effect in response formatting since it could be unclear if the response options were calibrated for rare events (e.g. "often for a rare event") or for events of all frequencies. Contradictions and tautologies can involve information exchanged in different questions or sections of the interview. So, for example, after the respondent has indicated that she approves of Clinton's "foreign policy" she is asked to rate his performance on "international affairs." While the question author may have intended the two questions to tap different opinions, the respondent believes she is being asked the same question twice and finds this baffling (and a violation of conversational norms).

A third type of logical problem involves the relationship between information in a question and its relationship to the respondent's circumstances. Suppose the respondent is asked "How many times a month do you visit a doctor?" and the respondent is a healthy, 25 year old. The presupposition in the question is that the respondent visits the doctor more than once a month but for this respondent the presupposition is false. The respondent understands the question but has trouble performing the task because she has no information about her rate of monthly doctor visits.

Computational problems. All of the problems in our coding scheme involve respondents' difficulty processing and manipulating information, so they are all computational in some sense. The current class of problem, which we have specifically called computational, functions as a residual category, because respondents have significant types of problems that do not fall into the other categories. Coders are instructed to assign problems to this category after all others have been considered. Many of the problems that are appropriately assigned to this category involve memory of one kind or another, but other problems involving language processing and mental arithmetic belong in the category as well.

Examples involving memory are forgetting information that was conveyed by the question (computational/task performance), difficulty recalling autobiographical facts or experiences (computational/task performance), and difficulty maintaining information in working memory such as a list or response options (computational/task performance). Language processing problems include difficulty integrating the various clauses of a particularly complex question (computational/understanding). A mental arithmetic problem (computational/response formatting) could involve difficulty converting a count of some kind -- yielded by the primary task -- into a percentage because the response categories are percentages; while the respondent understands what he needs to do, the division is too hard for him to do in his head.

IV. Using the coding scheme

Coders are asked to listen to tapes or read transcripts of the cognitive interviews and assign the problems that they perceive in the verbal protocols to one of the 15 problem categories in the coding scheme. They are given the descriptions of the problem categories that appear in the appendix and if they also conducted the interviews themselves, they are encouraged to consider their interview notes when classifying problems.

Author intent

The coding decisions demanded when using the respondent problem matrix may require coders to guess what the questionnaire author had in mind at the time he or she developed each question. It stands to reason that if coders had access to some of this information they would more accurately detect problems. In particular coders would make fewer false alarms and would classify legitimate problems more knowledgeably. As a result, the way the coders characterize and classify these problems may contribute more to solutions than if they are not exposed to author intent information. In addition, knowing what the author intends allows the evaluators to craft probes prior to the interview for places they think respondent performance may differ from author intent.

Therefore, in addition to the category descriptions, we advocate giving the coders a written summary of an interview with the author, conducted to elucidate the rationale behind each question, the intended interpretation of each question and the processes that respondents are intended to use in arriving at an answer. We have adopted the following procedure for developing the author intent document. First, the draft questionnaire is reviewed by several people knowledgeable about questionnaire design. Based on this review, a set of questions is formulated about any question in the draft instrument that were flagged in the review. The author is then questioned about these points. Finally, the authors responses are summarized and embedded next to the questions in the draft instrument. This final document is given to the coders. In the next section we describe a study conducted to evaluate our method. Some aspects of the evaluation are still underway. One of these is the use of author intent information, so we will not discuss it further in the current paper.

V. Evaluation of the method

Before a method such as the one we have developed can be recommended over the conventional use of cognitive interviews, there are several questions about its coverage and reliability that need to be addressed. Toward that end we conducted an evaluation study that provides some preliminary, empirical support. It is a case study: The number of participants is small and the interviews involve a single survey instrument. Therefore the results are mostly suggestive at this point.

Four interviewers each conducted ten cognitive interviews to ostensibly pretest a draft survey instrument. This instrument was 41 questions in length and concerned jobs, skills and use of time. The data collection procedure was modeled after what, in our view, is the prototypical approach to conducting cognitive interviews: The respondents were asked to provide concurrent protocols but if they did not do so, the interviewers were instructed to elicit a retrospective report; interviewers were given license to probe as they deemed necessary and explore possible problems with respondents. There were also several structured probes leading to uniformity across the interviews. These were derived from an earlier round of pretesting.

Two of the four interviewers prepared conventional, question by question, written reports on the problems they identified in their ten interviews. The interviewers were instructed to write a single narrative for each question, collapsing across individual interviews, but for any problems reported, they were to indicate the respondents by whom it was encountered.

The other two interviewers used the respondent problem matrix to classify the problems they identified in the verbal reports. They registered their coding decisions by interacting with a software version of the matrix which prompted the coder for problems in each category for each question. When prompted with a particular category name, the coder indicated whether or not she had detected a problem (or problems) of this sort and if so, entered a short textual description of the problem(s). The program wrote the results for each question for each interview to a file. One kind of report that can be produced from the coding results is mean number of problems per question for the ten interviews with a list of the problem descriptions.

By keeping the interview technique essentially the same for the four interviewers but varying the analysis and reporting of problems, we able to compare our approach to the conventional method in just the place they differ. In particular we would like to know which approach identifies a larger number of problems and which leads to greater overlap between interviewers in the problems they identify.

The two written problem reports described a total of 60 problems of which 43 (71%) were identified by only one interviewer. These discrepancies between the problems each interviewer identified cannot be attributed to a gross asymmetry in the number of problems reported by each: They each identified roughly equal proportions of the total set of problems (46% and 54%). The discrepancies arose because they each found different problems in their respondents' protocols. On the one hand, the 29% of the problems identified by both interviewers is not terribly impressive. But the problem reports were based on interviews with different sets of respondents. It may well be that traditional cognitive interviewing provides the best coverage of likely problems when several interviewers each collect data from several respondents. This would take advantage of the natural variation among respondents and interviewers' individual approaches. It would be important for subsequent research to identify the optimal range of interviewers and respondents, below which problems tend to be overlooked, above which there are relatively few additional problems detected.

[a paragraph presenting the comparable data for our technique will be inserted here; the data are still being analyzed; should be ready later today, 7/31]

Until now we have considered the amount of overlap in problems found in different interviews conducted by different interviewers. An important indication of how much stock one should put in the problems turned up with the respondent problem matrix is the amount of overlap in problems identified for the same set of interviews, coded by two people. To compute this kind of overlap, we trained two additional coders to use the method and asked them to code the taped interviews conducted by our two interviewer-coders. These coders did not conduct any interviews themselves. We considered there to be overlap if a pair of coders placed a problem in the same category or no problem in a category. On average, 77% of the problems identified by the interviewer-coders were also identified by the "pure" coders. That strikes us as moderately reliable performance, though we still need to conduct a detailed analysis of the discrepancies to see if we can increase the overlap, possibly by improving the coding instructions.

A related question is whether coders who have also conducted interviews detect different sorts of problems than coders who have not. There is evidence in the psycho linguistic literature that participating in a conversation leads to qualitatively different comprehension of a speaker's references than does overhearing that same conversation (Schober, 1989). If that is the case in our study, one would expect lower measures of overlap between interview-coders and pure coders than between the pairs of pure coders. In fact, there was no evidence of such a difference in our study. The pairs of pure coders identified 77% of the same problems – exactly the same proportion of

overlap as was found for coder pairs where one member had also conducted the interviews.

While this is a preliminary result, it could mean that a survey organization could separate the conduct of cognitive interviews from their analysis. Personnel who are best suited for eliciting verbal protocols can be given the data collection task and staff who are best able to use the coding system can be assigned analysis duties.

VI. Conclusions

The cognitive revolution in survey research was fueled by the success of cognitive psychology in characterizing human thinking, reasoning, comprehension and so on. That success is due in part to the development of compelling theories specified in computational terms. It is also attributable to the use of rigorous experimental methods, that rely on objective, quantifiable data wherever possible. It is ironic, therefore, that the way the survey methods community has adapted cognitive psychology is as a set of qualitative methods. Our work is an attempt to increase the consistency and objectivity of one "cognitive method," think aloud protocols, and in the process, to facilitate quantifying respondents' problems. Our method requires extensive evaluation before it can be widely recommended, though the preliminary evaluation suggests we are on the right path.

IV. References

Bickart, B. & Felcher, M. (1996). Expanding and enhancing the use of verbal protocols in survey research. In N. Schwarz and S. Sudman. (Eds.). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp. 115 - 142). San Francisco: Jossey-Bass Publishers.

Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive Psychology, 11*, 430 - 477.

Clark, H. H. & Bly, B.(1995). Pragmatics and discourse. In J. L. Miller and P. S. Eimas, (Eds.), *Speech, Language and Communication* (pp. 371-410). New York: Academic Press.

Forsythe, B., Lessler, J. & Hubbard, M. (1992). Cognitive evaluation of the questionnaire. In C. F. Turner, J. T. Lessler, and J. C. Gfroerer, (Eds.), *Survey Measurement of Drug Use: Methodological Studies* (pp. 13-52). Rockville, MD.: U.S. Department of Health and Human Services.

Krosnick, J. A. (1991). Response strategies for coping with cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213 - 236.

Lessler, J. T. & Forsythe, 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.259-292). San Francisco: Jossey-Bass Publishers.

Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Nielsen, J. (1994) Heuristic evaluation. In J. Nielsen and L. Mack (Eds.). *Usability Inspection Methods* (pp. 25-62). New York: John Wiley & Sons, Inc.

Oksenberg, L. & Cannell, C. F. (1977). Some factors underlying the validity of response in self-report *International Statistical Bulletin, 48*, 325 - 346.

Searle, J. R. (1975). Indirect speech acts. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics : Vol. 3. Speech Acts* (pp. 59 - 82). New York: Seminar Press.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*, 211-232.

Tourangeau, R. Cognitive sciences and survey methods (1984). In T. B. Jabine, M. L. Straf, J. M. Tanur and R. Tourangeau (Eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines* (pp. 73-100). Washington, D.C.: National Academy of Sciences Press.

Tourangeau, R. & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement, *Psychological Bulletin, 103*, 299- 314.

Table 1. Respondent Problems Matrix

STAGE

Problem Type	Understanding	Task Performance	Response Formatting
Lexical	the term "servings" is too vague	could not compare units in question (apples and cups of juice); confused about whether should convert one to the other	could not express opinion on scale with current labels: anchors are "too frequently" and "appropriately" and respondent wanted to register "not frequently enough"
Temporal	treated "the last year" as 1994 instead of the 12 months prior OT the interview	Respondent was not willing to recall the hours worked for every week in the last year.	seemed to treat middle frequency as typical and position own behavior accordingly.

Appendix: Problem code descriptions and coding instructions.
Problem Code Definitions

Introduction: We view the process by which respondents answer survey questions as involving three, roughly sequential stages, represented by the columns in the matrix: understanding the question, performing the primary task, and formatting the response. Within each of these stages, there can be different types of problems, represented by the rows in the matrix: lexical, temporal, logical, computational and omission and inclusion problems.

We view the question (including the response options) as a set of instructions about the task the respondent is to perform. Some aspects of the instructions are explicit in the question and others are implicit. Understanding (the first stage in the question answering process) includes making sense of the words in the question as well as of the instructions (that may go beyond the specific words in the question). The task that respondents are instructed to perform is referred to as the "primary task" to distinguish it from the response formatting task. This secondary task generally involves mapping the results of the primary task to the response options given with the question.

If a problem can be assigned to more than one cell that's okay, but please indicate your first choice.

LEXICAL PROBLEMS involve individual words and their meanings.

Understanding: The respondent has trouble understanding the question and/or the task because she is not familiar with a particular word or does not know what it means in the current context. It is possible that more than one word in the question will result in this kind of problem. In that case, code each problem separately. Problems of this type can involve words in the body of the question or words used to convey the response options.

Primary Task Performance: The respondent has trouble performing the primary task because it is difficult to use one or more of the words in the question. The emphasis is on using words as opposed to understanding their meaning. *Examples*: (1) The task might require the respondent to recall events from a category named by a word in the question but the respondent is uncertain whether what he has recalled actually belongs in the category; (2) the task might require respondents to compare quantities of two categories named by words in the question but the categories are not expressed in comparable units, complicating the comparison (e.g. the amount of fiber in an apple versus a cup of apple juice).

Response Formatting: The respondent cannot easily or correctly assign the information he produced in the primary task to an explicit response category because it is not clear how the meanings of the "raw" response and the category label interrelate. This type of problem is not related to understanding the meanings of words in the response categories but rather using those words to register a response. *Examples*: (1) The respondent is asked how many compact disks he owns and he determines the exact number; the response options are qualitative, for example, "very few", "an average amount", "quite a few", and the respondent does not know how to classify the number he has produced; (2) the respondent is asked about the a concept that the author believes most people disapprove of, for example, negative advertisements in political campaigns, but in fact the respondent approves of the concept; as a result the response options of "far more than necessary" "slightly more than necessary" and "just the right amount" do not capture the respondent's actual feeling.

TEMPORAL PROBLEMS involve the time period to which the question applies or the amount of time spent on an activity described in the question. Problems involving respondent's treatment of the concepts "usually" or "typically" are considered temporal because they imply a certain time period.

Understanding: The respondent incorrectly interprets or has trouble interpreting the temporal characteristics of the question. *Examples*: (1) The respondent interprets the phrase "in the last year" to mean "in the previous calendar year" instead of "in the last 12 months" as was intended; (2) the respondent is unsure about how to interpret "usually;" her life has changed in the recent past and as a result her "usual" activities have also changed; she will respond differently depending on whether she answers about her life before or after the change. Example 2 could be viewed as a lexical problem because it involves a particular word in a particular context; however, problems with words that involve time should be treated as temporal problems.

Primary Task Performance: The respondent has trouble performing the primary task because it is difficult to use the temporal information in the question. *Examples*: (1) The task might ask the respondent about her activities "in the last five weeks" which she finds to be an unnatural unit of time, and therefore hard to use; this type of problem could, in principle, be treated as a Computational-Task Performance problem but all such problems involving the question's temporal properties should be assigned to the Temporal-Task Performance category; (2) the question concerns "the current month" but the respondent has forgotten that it is now a new month and performs the task as if it were still last month; this is not an issue of understanding the temporal properties of the task but rather one of using them because one can perfectly well understand "the current month" in an abstract way but part of the task is to make the phrase concrete with the correct month.

Response Formatting: The respondent has trouble selecting a response category because the temporal information produced in the primary task is somehow incompatible with the temporal information in the response categories. *Example*: The respondent is asked about the frequency with which he performs a particular activity, and so he performs the primary task by counting up all times he performed that activity during the reference period; in other words the primary task has produced numerical information; the response categories are qualitative (e.g. "extremely rarely") and the respondent does not know how to interpret the numerical information ("is 8 times 'somewhat rarely' or 'somewhat often'?").

LOGICAL PROBLEMS: These problems involve the logical relations in the question. They may be signaled by certain terms, such as logical connectives like "and" and "or", terms of negation such as "not" and "never", and terms that describe set theoretic ideas, like "*other* than those already mentioned." However, there may also be logical problems that do not use special terms but are inherent in the question, for example, contradictory information in the question.

Understanding: The respondent assigns the wrong logical interpretation to the question. *Examples*: (1) The respondent is asked if in the last week he has "Purchased or had expenses for meats and poultry." The respondent interprets this as instructions to respond affirmatively if he has purchased both meat and poultry products -- for example, pork chops and chicken; in fact the author conceived of the task as something like "have you purchased any products from the category meats and poultry?" so if the respondent had purchased either pork chops or chicken, he should respond "yes"; (2) In the question, "The homeless should not be allowed to collect welfare payments – Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree" the respondent overlooks the "not" in the question stem; this oversight seems to be related to the difficulty of understanding double negatives such as the one created by the "not" and "disagree" response categories.

Primary Task Performance: The logical relations expressed in the question are difficult to use in performing the primary task. While the logic may be understood, the respondent does not successfully (or easily) carry out the task implied by the logic. *Examples*: (1) The respondent is asked if in the last month he has purchased any products from the category of *Non-Powered Tools, and Miscellaneous Hardware and Supplies*. The respondent finds that he may need to think about each component category separately until he either recalls a relevant purchase or decides he has purchased no products in any of the component categories; however, he responds "No" after considering only the first category, *Non-Powered Tools*; (2) The respondent is asked "Have you been examined by any physicians not already mentioned?" but finds he cannot remember who has already been mentioned and that it is hard to think about the category "physicians not already mentioned"; while it is similar to a computational problem, it is also a logical problem because it involves categories and their complements (of all possible physicians, the ones that have not already been mentioned); it should be coded as a logical problem.

Response Formatting: It is hard to respond because the logic of the response task conflicts with the type of information produced by the primary task or there is a logical inconsistency among the response categories for the particular information produced through the primary task. *Examples*: (1) A response task requiring the respondent to select a single option when the respondent would prefer to check all that apply is a logical problem: the response task requires an "or" decision but the respondent would like to make an "and" decision. (2) if response categories are not mutually exclusive, a respondent

might produce information in the primary task that can be assigned to more than one category; this is a case in which the respondent would like to make an "or" decision but the response could be assigned to more than one category.

COMPUTATIONAL PROBLEMS: These problems involve the difficulties respondents have processing and manipulating information. This includes forgetting information (either explicit or implicit) conveyed by the question, forgetting information produced as part of answering the question that is needed to accurately complete the task, difficulty recalling facts or experiences, and, in general, difficulty or unwillingness to accurately do the "mental work" that respondents are asked to do. Computational problems that involve temporal or logical relations should be assigned to temporal and logical categories.

Understanding: The respondent has trouble understanding the question and/or task because it is too long or too complicated. This may involve forgetting part of the question or never processing it completely because if its grammatical complexity.

Primary Task Performance: The respondent cannot easily or accurately perform the primary task because it involves more components than can be kept in mind or more steps than the respondent can carry out. *Examples:* (1) the respondent is read a list of attributes and asked to indicate which apply to her; she asks for the list to be reread because she cannot remember all of the items; (2) the respondent is asked for the address at which she lived at a particular point in her life, but she cannot remember it; (3) the respondent is asked for total household income and cannot recall which member's income she has added to the sum of incomes at any one point.

Response Formatting: The respondent has trouble converting information produced in the primary task into an acceptable response format because the conversion involves more components than can be kept in mind or more steps than the respondent can carry out. The conversion process generally involves mental calculations or transformations that tax the respondent's abilities. *Example*: The primary task yields counts of some kind but the response categories are expressed in terms of percentages; the respondent must divide the counts by the totals and this turns out to be burdensome or error prone.

OMISSION AND INCLUSION PROBLEMS: These problems involve overlooking information or inappropriately incorporating information into the question answering process. Typically, problems caused by forgetting are assigned to computational categories, but information can be omitted because respondents are not aware of its relevance. In addition, information can be excluded or inappropriately included because of an erroneous inference.

Understanding: The respondent reaches an interpretation of the question that is either incomplete or overly elaborate. This kind of problem can arise because the respondent misunderstands the scope of a concept in the question. *Examples*: (1) A respondent is asked if she has any family members with a particular disease, and she decides that "family members" does not include brother-in-law, although the author did intend for in-laws to be included; (2) the same respondent decides that the disease, respiratory illnesses, includes allergies but the author intends for allergies to be excluded.

Primary Task Performance: The respondent performs the primary task without taking account of information in the question or information produced in the course of answering the question, or includes such information although it is unwarranted. *Example*: The respondent is asked how many people are in the household; she has a daughter who she considers to be part of the household but who lives away at college; however, she does not think this fact is relevant when, in fact, the authors would have excluded the daughter.

Response Formatting: The respondent assigns the response to a category that was not explicitly offered or does not consider an option that is provided. *Examples*: (1) The response task involves circling one whole number between 0 and 10 but the respondent writes in 6.5 because it better reflects his judgment; he has essentially added a response option; (2) the respondent believes the extreme values of a scale are never applicable and restricts the assignment task to the other values; he has essentially eliminated two options.

Addendum: The Problem Types

Lexical:

Lexical problems have to do with the 'core' meaning of words or idioms. So if a respondent does not know [or is uncertain] of the core meaning of a word or idiom, it is likely to lead to some type of lexical problem. e.g. not knowing what "nitrogen" means or not understanding the phrase "the lion's share."
Inclusion/Exclusion:
If the respondent understands the meaning of a word [e.g. "doctor"] but does not use it in the way the survey question intends-- which may be indicated explicitly in the question, in a transition statement or implicitly by the context the question appears in-- [e.g. to exclude anyone who is not an M.D. from the category "doctor"] then this is some type of inclusion/exclusion problem.
Another variation is when a respondent understands a general category [e.g. religious groups] and can easily include items that are typical [central to] the category [e.g. Catholics] but has difficulty handling items that are part of the category, but are less typical [or debatable] category members [e.g. Branch Davidians].
So this type of exclusion/inclusion is almost a sub-category [albeit a very specific one] under lexical. But note that there are other types of inclusion/exclusion problems as well.

Temporal:

Logical:

There are three general types of logical problems:

The first is a formal sense of logic, such as misusing logical connectors like the word 'and' or the word 'or', other grammar problems that cause logical confusion. Tautologies such as Do freak accidents happen rarely, sometimes or often? would also present "logical" difficulties. Since by definition freak accidents happen rarely, the question [which includes the response categories] does not seem a logical thing to ask. These 'formal' senses of logical are basically structural problems.

The second area has to do with the logical connection between the survey question [or its implied assumptions] and relation to the respondent's world, e.g. asking about how many times a month do you visit a doctor, when most respondents do not have behavior patterns that allow them to make logical sense of what the question asks. So for most respondents, who visit a doctor a perhaps couple of times a year, the question does not seem a logical way to ask about doctor visits because it does not match their view of the world [based on their own experience]. For someone with a serious health condition that requires weekly monitoring [as an outpatient] the question does make 'logical' sense.

The third has to do with the logical connection between different parts of the survey. For example, having first asked the respondent's sex, the survey asks a question inappropriate to that gender. Or having elicited the respondent's opinion about an issue, the survey then asks a question that does not take into account the information the respondent has already provided. In both instances, the questions strike the respondent as "illogical," though the respondent may still try to make some sense of the question. An example might be asking a redundant question. It seems to the respondent that one would not [logically] ask the same question again and therefore assumes the second question must mean something slightly different. The respondent then goes about trying to construct a sensible meaning to the question, given the context. [Of course, respondents may think a question is redundant when it's really not.]

Computational

All task performance is, to some extent, computational, i.e. it's information processing. We want to use our computational category as a "residual" for types of performance problems that do not fit into the other problem types, lexical, temporal etc. Treat Computational as if it were at the bottom of the list.

An important part of this category is memory problems. These generally take one of two forms. First, keeping something in mind [in working memory], like a list of response categories or all the parts and conditions of a very long question. Second, recalling some past event or behavior [from long term memory], like the grades the respondent got in elementary school.