FROM IMPRESSIONS TO DATA: INCREASING THE OBJECTIVITY OF COGNITIVE INTERVIEWS

Frederick Conrad,*Bureau of Labor Statistics*and Johnny Blair,*University of Maryland*[1]
Frederick Conrad, BLS, 2 Massachusetts Ave. NE, Room 4915, Washington, DC 20212

Key Words: Cognitive pretesting methods, Verbal reports

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 (for example, see Willis, Royston & Bercini, 1991). 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, non-controversial theory of how people solve problems (Ericsson & Simon, 1992). The value of such a framework is that it constrains the inferences that researchers are licensed to make about think aloud data. Moreover, 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 drafting 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

The way cognitive interviews are typically used is as semi-structured, in-depth interviews, which enable the interviewer to form impressions about where the problems in a questionnaire lie. These impressions are usually enumerated in a written report and supported with examples. It is easy to see how different interviewers could reach different conclusions about the identity and locus of questionnaire problems depending on how they have conducted the interviews and the kind of evidence to which they are sensitive. In fact, one study has shown that reliability is fairly low for this implementation of cognitive interviews (Presser & Blair, 1993).

These reliability limitations could originate in *collecting* think aloud data, *analyzing* them, or both. Our focus here is exclusively on analyzing respondents' verbal reports, though the way they are collected certainly warrants extensive study. In our approach, collection and analysis are temporally separated 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.

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

[1] The opinions expressed here are those of the authors and not necessarily those of the authors' institutions.
We thank the following people for help and support: Atar Baer, Mick Couper, Jim Esposito, Sarah Jerstad, Yun-Chiao Kang, Dominic Perri, Stanley Presser, Linda Stinson, Deborah Stone, Tim Triplett, Clyde Tucker, Beth Webb and Gordon Willis.

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 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.

	RESPONSE STAGE		
PROBLEM TYPE	**Under-standing**	**Task Performance**	**Response Formatting**
Lexical			
Temporal			
Logical			
Compu-tational			
Omission/ Inclusion			

Table 1. Respondent Problem Matrix. Instances of each problem type can occur at each response stage.

Variations on 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. In general, the four stages are comprehension, retrieval, judgment and response formatting/selection. The model is specified at about the same level of detail as Ericsson and Simon's (1992) view of problem solving and like the Ericsson and Simon approach, it is not controversial. Just as the analysis of verbal reports of problem solving is guided by that Ericsson and Simon's 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.

Bolton (1993) has developed a scheme that enables a respondent's verbal reports to be automatically classified into a problem category that is associated with one of four response stages. The respondents' verbal reports are first transcribed into an electronic form and then the text for each utterance is automatically searched for keywords or inflectional cues that are indicative of a particular problem category. If a match is found, the utterance is classified accordingly. Removing a human coder from the loop leads to objective analyses of think aloud data. However, in the interest of objectivity, this approach

excludes the subtle judgments that (under current technology) only human coders can provide. We rely on such judgments in our approach.

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 (see Figure 1).

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 Cannell 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, and the respondent's relevant knowledge, for example concerning the questionnaire's topic; 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 as the response.

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 (stage2). Any problems in her protocol will have their source at some point after understanding.

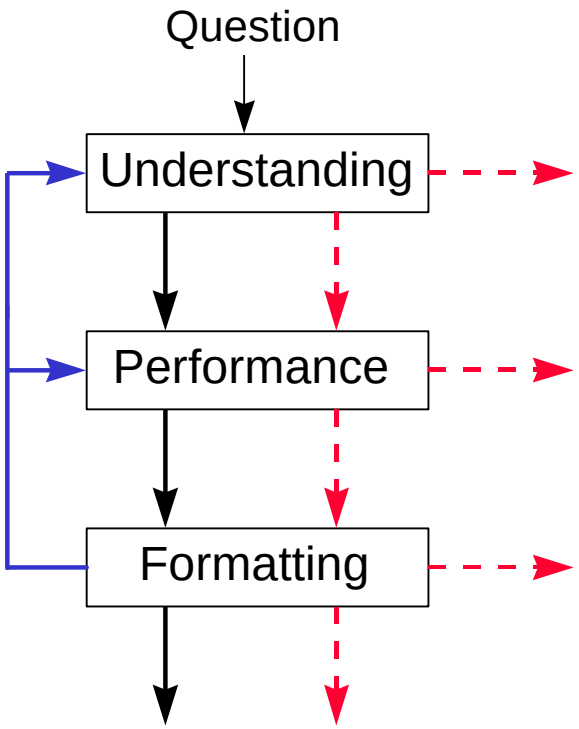

Figure 1. Revised response model for respondent problem coding scheme. Dashed arrows indicate inadequate input to next stage. Solid downward arrows indicate problem-free responding. Arrows on left indicate respondent can return to previous stages or current stage.

However, if she just cannot grasp what she is being asked to do (comprehending the task), she has not derived the necessary information to perform the second stage. Her problem lies in understanding and she recognizes her difficulty. The dashed, horizontal arrow exiting from the comprehension stage in Figure 1 represents this situation.

It is possible for the respondent to lack the necessary information to begin a new stage without recognizing this. The result may be that he carries out the next stage without adequate input, increasing the chances that he will do so incorrectly. The dashed downward arrow between the stages in Figure 1 represents this situation. Suppose, for example, the respondent believes that the task is to provide rough estimates of some quantity when in fact the intended task is to provide as precise estimates as possible. The respondent's imprecise estimates could be interpreted as evidence that the task is too difficult to perform when, in fact, the source of the problem is in the input to the performance stage – an understanding of the question and associated task. This kind of analysis should help clarify these distinctions and help fix the problem more successfully.

The second assumption required to adapt the generic model to the content of verbal report data is that respondents can 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 herself, or reason about the task based on the content other questions, for example "I just answered this question about my *occupation* by giving them my *job title* but I now realize they already asked me about my *job title* so the current question about my *occupation* must have to do with my duties instead of my title."[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. The data from the primary task 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 home 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 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" to the available options.

[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 perform the task in the least demanding way that produce a plausible answer (Krosnick, 1991).

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. We now turn to the fifteen problem categories that result from crossing the rows and columns.

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. Examples of lexical/understanding problems include (1) not knowing what is meant by a word like "nitrogen" or "spatial" in "spatial abilities;" (2) being unfamiliar with idioms like "the lion's share;" and (3) despite being familiar with the meanings of a pair of words not understanding their combination in the question, such as "medical purchases."

As an example of a lexical/task performance problem consider a question that asks the respondent for the number of rooms in her house. She understands what is generally meant by the term "rooms" but is unsure whether to count her living/dining area as one or two rooms because there is only a partial wall separating the two spaces. She understands what task she is being asked to perform but has trouble using the words in the question to perform the task.

It is considered a lexical/response formatting problem if the respondent cannot easily or correctly assign the information 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 would be the case in the compact disk example given earlier where the respondent cannot map a numerical quantity to the qualitative response options.

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. These problems are sometimes special instances of lexical problems. Our experience has shown that they are sufficiently numerous to warrant their own category. For example, an inclusion/exclusion/understanding problem might occur when the respondent is asked a question about "doctors" and interprets this as including chiropractors when the author intended "doctors" to include only physicians. If this can be clarified, the respondent can then perform the task.

An inclusion/exclusion/task performance problem occurs when there is no explicit decision rule for including or excluding instances from a category and the respondent is required to make this decision as part of the task. For example, lets assume the respondent understands the phrase "religious groups" and can easily include items that are typical of the category like Catholics or Muslims. However, the respondent cannot decide whether to include or exclude a group like the Branch Davidians, which, if included, would certainly be less typical than Catholics or Muslims.

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 points provided on the response scale are 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. Like inclusion/exclusion problems, temporal problems are often a special case of lexical problems. In this case they involve trouble grasping the meaning of temporal terms or using them. As with inclusion/exclusion problems, we have created a distinct category for temporal problems because of the prevalence of questions involving time periods.

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.

A temporal/response formatting problem typically involves a response produced in the primary task that is somehow incompatible with the available response options. Much like the lexical/response formatting example, a respondent might produce a precise count in response to a question about frequency for some activity during a specific time period. Because the response options are qualitative, such as "not very often," "occasionally," etc., the respondent has trouble selecting an option.

Logical problems. There are several types of logical problems. Each can occur at any of the response stages, though we provide examples for each at primarily one of the stages.

One type of logical problem 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/understanding problem. "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.

A second type of logical problem involves false presuppositions in a question. 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 (a logical/task performance problem) because she has no information about her rate of monthly doctor visits.

A third 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. Let's assume the respondent interprets the primary task as an instruction to recall and count all of the freak accidents she has experienced over some time period. (Admittedly, this question could pose serious problems understanding the task but we won't consider them for this example). A lexical/response formatting problem could occur because the respondent is unsure could if the response options ("rarely, "often" etc.) are calibrated for rare events (e.g. "often for a rare event") or for events of all frequencies.

Contradictions and tautologies can also involve information exchanged in different questions or sections of the interview. So, for example, after the respondent has indicated that she approves of the president'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).

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.

A question whose syntax is particularly complicated, for example with many embedded clauses, could pose a computational/understanding problem if the respondent cannot parse it as it is spoken by the interviewer.

If the task involves recalling relatively detailed episodes from autobiographical memory, particularly over a long period of time, the respondent may be unable to comply with the instructions, leading to a computational/task performance problem. Similarly, the task may require holding too many partial responses in working memory to complete the task. For example, if the respondent is asked how many magazines she receives by mail, she may forget whether or not she has already counted a particular one.

Difficult mental arithmetic could be coded as a (computational/response formatting) problem if, for example, it required 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. Analyzing verbal protocols

The approach to analyzing verbal reports that we are advocating has two parts: using the coding categories and eliciting author intent to inform coding decisions.

Using the coding scheme

Coders must first understand the problem categories. In order to train them in these concepts, their first task is to create at least one example of each problem category in the taxonomy. Other staff members who are already well versed in using the problem categories review the example problems and give feedback to the coders. The

coders revise their examples on the basis of the feedback and submit them for a subsequent review. This process continues until their examples are judged to illustrate a category's central concepts. In our experience this occur after two or three cycles.

The coders are then 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. A particular question may have more than one problem. The coders are given descriptions of the problem categories and, if they also conducted the interviews themselves, they are encouraged to consider their interview notes when classifying problems.

Author intent

When the coder or analyst's understanding of the intent of a question has to be inferred from the question text, that understanding may differ from that of the author. This is true if only because first drafts of questions have imperfections. It stands to reason, therefore, 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 questions in the draft instrument that were flagged in the review. The author is then questioned about these points. Finally, the author's responses are summarized and embedded next to the questions in the draft instrument. This document is given to the coders.

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.

Two interviewers each conducted ten cognitive interviews to ostensibly pretest a draft survey instrument. This instrument was 50 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.

The interviewers then 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 implementation of the matrix which prompted the coder for problems in each category (cell) 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 question about cognitive interviews in general, and not our analysis technique per se, is whether one can be confident that a small number of interviews can expose most of the notable problems in a questionnaire. The method has been advocated as a small sample alternative to traditional pretesting (Lessler, Tourangeau & Salter, 1989), though just how many interviews are required for thorough pretesting is not yet clear. One indication that a set of cognitive interviews has uncovered most of the problems in a questionnaire is that the same problems are identified in multiple interviews by different analysts. By this criterion, our two sets of ten cognitive interviews have not turned up all of the notable problems: While the two coders, each analyzing their own interviews, identified about 1.9 problems per question, they identified the same problems (assigned the same code) for only seven out of 50 questions (14%).

This strongly implies that a larger number of interviews is required in order to exhaustively identify problems. Because the approach we are advocating is designed to be usable by junior staff, an organization's cost for these additional interviews is considerably smaller than it would be if graduate level psychologists conducted the interviews, as is common practice.

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. (The previous analysis involved different analysts coding different interviews.) 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 these additional coders. That strikes us as moderately reliable performance – especially given the poor reliability found by Presser & Blair (1994) for conventional analysis of cognitive interviews. Nonetheless, we should be able to 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 psycholinguistic literature that participating in a conversation leads to qualitatively different comprehension of a speaker's references than does overhearing that same conversation (Schober, 1989). In fact, there was no evidence of such an effect in our study. The pairs of coders who did no interviewing, and therefore accrued none of the special insight that might have come from also interviewing the respondent, identified 78% of the same problems as each other – the same proportion of overlap as was found for coder pairs where one member had also conducted the interviews. If there had been any effect on coding from also having conducted the interview, then variation in interviewing duties should have affected which problems were identified, thus lowering overlap. But there was no effect of interviewing on overlap.

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.

After the coders had identified and classified the problems in the interviews that they had conducted, we asked them to revisit the think aloud data with some knowledge of the author's intentions behind particular questions. The coders were provided with a copy of the questionnaire that was annotated with information about the intent behind the first 39 questions. (The remaining 11 questions were written by a different author and their intent was not considered in the current study.) The additional information led the coders to revise their original codes, presumably sharpening their problem detection. One coder identified seven additional problems; the other coded one new problem, deleted nine problems and revised the code for four.

One lesson from this exercise is that authors can be intentionally imprecise about the goals of a question. A case in point is the ambiguity surrounding the way respondents are intended to answer "How long have you been employed at your current job?". It was not clear if respondents were to discount time spent away from the job such as for maternity leave or not. By interviewing the author, it became clear that the question was not intended to provide this degree of precision but to provide coarse distinctions between experienced and less experienced workers. Knowing this could help direct resources to refining questions that genuinely do not function as intended.

Another lesson is that furnishing analysts with knowledge of the author's intentions ultimately leads the analysts and authors to converge in their understanding of particular questions. Sometimes authors are themselves unclear on their goals for a question and certainly analysts are often in the dark. Several cycles of this approach should bring both parties closer to a mutual understanding of the question. Moreover, by iteratively refining questions to make the author's intent clear, it is increasingly likely that respondents will understand the question as its author intended it to be understood

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 largely impressionistic 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.

VII. 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.

Bolton, R. N. (1993) Pretesting questionnaires: Content analyses of respondents' concurrent verbal protocols. *Marketing Science, 12, 280-303.*

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.

Lessler, J. T., Tourangeau, R. & Salter, W. (1989). Questionnaire design in the cognitive research laboratory. *Vital Health Statistics,*Series 6, No. 1.

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.

Presser, S. & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology,*

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.

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