QUESTIONNAIRE PRETESTING:
COMPUTER ASSISTED CODING OF CONCURRENT PROTOCOLS

Ruth N. Bolton
Tina M. Bronkhorst

Published In: Bolton, Ruth N. and Tina M. Bronkhorst (1995), "Questionnaire Pretesting: Computer Assisted Coding of Concurrent Protocols," in Answering Questions, Norbert Schwarz and Seymour Sudman (eds.), San Francisco: Jossey-Bass Publishers, 37-64.

Companies in many service industries regularly survey their customers to monitor their current performance, identify potential service enhancements, and to evaluate the effect of enhancements. For example, GTE's surveys ask customers about their recent telephone experiences, their perceptions of various service attributes, and their evaluations of overall satisfaction, service quality and value. A large service company, such as AT&T, Marriot Corporation, Federal Express, or GTE, may contact as many as 50,000 customers each month! Most companies closely scrutinize their survey data, and utilize the results in a variety of policy decisions, ranging from modifications of the service delivery process to investments in plant and equipment. In these circumstances, managers are vitally interested in the quality of the information that they use in their decisions -- and that implies a corresponding interest in minimizing response and non-response error in the design of their surveys. As a result, companies have a vested interest in understanding the question answering process.

Over the past six years, GTE has worked to enhance the quality of the information it obtains through its surveys -- focusing on improvements to its customer satisfaction questionnaires. Our experiences have led to the development of a new pretesting methodology that identifies respondents' cognitive difficulties as they form answers to survey questions. This methodology entails the computer assisted coding of concurrent verbal protocols elicited during pretest interviews. Our prior research has shown that this methodology can be used to identify and improve defective questions.

This paper elaborates on two aspects of our pretest methodology. It describes techniques for eliciting concurrent verbal protocols in pretest interviews. It also describes an extension to our automatic coding scheme, designed to measure respondents' cognitive difficulties answering questions about low involvement, low frequency events. The study context is a pretest of a customer satisfaction survey intended for administration to residential customers of local telephone service. The study focuses on questions about low involvement telephone experiences, such as

calls to directory assistance, that are likely to be relatively low frequency events for most customers.

The remainder of the paper is organized in the following way. The following section briefly describes existing pretesting methods, and summarizes our development of a pretesting method based on computer assisted coding of concurrent protocols. The next two sections explain how to elicit concurrent verbal protocols in a pretest interview situation, and how to automatically code them. Then, we describe a pretest involving survey questions about low involvement, low frequency telephone experiences. The concluding remarks discuss how this methodology can be best utilized to study the question answering process.

PERSPECTIVE

In industry, market researchers pretest questionnaires by a variety of informal methods. A draft questionnaire may be reviewed by "experts." Or, a convenience sample of customers may be administered a draft questionnaire (often a mail survey) and then debriefed in a focus group interview. Another common strategy is for experienced interviewers (who may or may not be part of the research team) to administer the questionnaire to a small number of respondents and then participate in a debriefing session. Sometimes this activity amounts to little more than a CATI (Computer Assisted Telephone Interviewing) debugging exercise designed to detect problems with the programmed questionnaire, such as incorrect skip patterns! These different pretesting practices allow for quick adjustments to the survey instrument, and they are relatively cheap and easy to implement. However, they do not focus on the question answering process, nor attempt to directly measure respondents' cognitive difficulties. Consequently, as Fowler (1992) points out, a common feature of industry practices is that "the criteria . . . and standards they should use for when a 'problem' exists are seldom made explicit."

Prior Research

Many survey researchers have adopted observational monitoring as a simple, low cost and effective way of formally pretesting questionnaires. Observational monitoring relies on a coding

scheme based on interviewer-respondent interactions (e.g., Bercini 1989; Cannell and Oksenberg 1988; Fowler 1989, 1992; Oksenberg, Cannell and Kalton 1991). The questionnaire is administered to a sample of respondents using the intended method of survey administration. Expert coders monitor the interviews (or listen to audiotapes) and apply the coding scheme to each questionnaire item. An important strength of observational monitoring is that it can be conducted in actual field situations, which are frequently characterized by low involvement and high distraction conditions.

Formal questionnaire pretesting methods have usually focused on directly identifying question defects. For example, Hunt, Sparkman and Wilcox (1982) subjectively coded (i.e., used experts to code) question defects, such as ambiguous questions. More recently, survey researchers have used a variety of cognitive research methods to identify respondents' difficulties as they form answers to survey questions. Unlike observational monitoring, these methods are designed for developmental pretests conducted in laboratory settings. They have included: concurrent and retrospective "think aloud" interviews, paraphrasing, intensive interviews, confidence ratings, vignettes, response latency measurement and sorting tasks (Campanelli, Martin and Creighton 1989; Jobe and Mingay 1990; Royston, Bercini, Sirken and Mingay 1986; Royston 1987; Willis, Royston and Bercini 1991).

Cognitive methods can be characterized by the extent to which they rely on the collection of verbal versus non-verbal data, and the extent to which these data are analyzed by qualitative versus quantitative methods (Table 1). The majority of pretesting methods have collected and analyzed verbal data. Among the quantitative methods for analyzing verbal reports, observational monitoring relies on expert interpretation of verbal interactions to identify defective questions.[1] In contrast, our approach utilizes computer assisted methods to code the content of respondents' verbal reports and identify cognitive difficulties.

Applications of Automatic Coding in Questionnaire Pretests

Bolton (1991; 1993) developed and applied eleven different automatic coding categories for questionnaire pretesting. Across the two studies, she used seven categories of verbal cues (i.e., words or word strings) and four categories of non-verbal cues. Principal components analyses of the intensity measures for these categories provide evidence of the convergent and discriminant validity of four underlying dimensions of respondents' cognitive difficulties, labelled: comprehension, retrieval, judgment and response. The four dimensions seem to be relatively stable across different respondents and different survey content domains.

Bolton's (1991) first study was a split ballot experiment that tested two versions of a residential customer survey concerning telecommunications services. Her experiment showed that it was possible to distinguish between superior and inferior versions of questions on the basis of the frequency and intensity of different automatic coding categories. In her second study, Bolton (1993) conducted a split ballot experiment to test three versions of a business customer survey concerning (different) telecommunications services. The study demonstrated how to identify defective survey items through statistical tests that compared respondents' cognitive difficulties with alternate versions of the same question or across different questions. Specifically, factor scores representing the four underlying dimensions were used as dependent variables in a series of one-way ANOVAs that compared different items. This study also provided some evidence supporting the external validity of this assessment of defective questions.

These two studies demonstrated the automatic coding of concurrent protocols elicited during pretest interviews yields explicit quantitative criteria for identifying defective questionnaire items on the basis of respondents' cognitive difficulties. In addition, Bolton and Bronkhorst (1991) provided a comparison observational monitoring, subjective coding and automatic coding of depth interview data on three criteria: ability to test propositions, flexibility and efficiency. In summary:

* Ability to Test Propositions. Automatic coding is particularly appropriate for testing propositions about the content of respondents' verbal reports. Since the content

characteristics or respondents' speech includes a wide range of cues indicative of cognitive processes -- including non-verbal cues such as broken utterances, pauses, unintelligible utterances, and tone of voice -- automatic coding can also be used to test some propositions about cognitive processes. However, it may be less suitable for tracing certain processing strategies, such as search or comparison strategies.

* Flexibility. Like other coding methods, considerable effort is required to develop a rich set of automatic coding categories. However, it is very easy for the researcher to revise and test alternative coding schemes. Furthermore, the application of any given coding scheme is always 100% reliable.

* Efficiency. Automatic coding is more efficient than subjective coding because the actual application of an automatic coding scheme is much faster. For example, 15 one-hour personal interviews require about 20 hours for transcription and precoding -- and 30 minutes for implementation of the actual automatic coding scheme. Naturally (like other coding methods), additional time is required to develop the coding scheme. Automatic coding is cheaper and faster than subjective coding when there are many interviews or interviews are lengthy.[2] Observational monitoring is more efficient than either method, but its coding schemes are necessarily simpler.

Observational monitoring and automatic coding seem to be natural complements. Both methods provide quantitative criteria for identifying defective questionnaire items. However, observational monitoring seems more suitable for polishing pretests under field conditions; whereas automatic coding seems more suitable for developmental pretests under laboratory conditions.[3]

Extending Automatic Coding to Address Other Research Objectives

Our research shows that automatic coding of verbal reports is a useful approach to studying how respondents answer survey questions. However, substantial research is necessary to develop coding categories that are suitable for different research objectives. In this study, we develop new

automatic coding categories to identify when respondents are experiencing cognitive difficulties retrieving information about low involvement, low frequency events. Many attitude and opinion surveys concern such "mundane" events, the results should be useful in a wide variety of survey contexts.

Numerous studies support for the notion that respondents experience cognitive difficulties responding to questions about low involvement, low frequency events. Since low involvement events are subject to less processing, respondents will have few associations in memory, which may create biases in their responses (Jonides and Naveh-Benjamin 1987). For example, an individual can create a belief, attitude or intention if the construct does not exist in long term memory -- sometimes using retrieved answers to earlier survey questions as inputs to the response generation strategy (Feldman and Lynch 1988). In addition, research shows that respondents use different processes to construct answers to frequency questions, including direct recall, enumeration or estimation strategies (Bickart et al 1990; Blair and Burton 1987; Felcher and Calder 1990; Schwarz 1990). For low frequency events, it seems likely that respondents will use estimation strategies that combine fragmented recollections concerning the event using an inferential heuristic. For example, an individual may estimate a frequency or probability by the ease with which instances or associations can be brought to mind -- even though availability may be affected by factors that are not related to actual frequency (Tversky and Kahneman 1973). Respondents' answers to survey questions about low frequency events may be characterized by similar biases in the question answering process. In summary, it seems likely that respondents will experience cognitive difficulties in forming answers to questions about low involvement, low frequency events that will lead to response errors.

COLLECTING CONCURRENT PROTOCOLS IN PRETEST INTERVIEWS

Our questionnaire pretesting methodology involves the elicitation of concurrent verbal protocols -- that is, asking respondents to think aloud as they answer the survey questions. This

method of eliciting verbal reports must be distinguished from cognitive interviews -- that is, intensive interviews in which an experienced interviewer (or the investigator) administers extensive probes and debriefing questions (e.g., Royston 1989). Unlike intensive interviews, think aloud interviews utilize identical data collection procedures so that responses can be compared across questions.[4] Each interviewer follows the same procedures so that each respondent hears the same task instructions. This feature allows problem questions to be identified.

The data collection procedure has four major components.

1. The interviewer reads the task instructions asking the respondent to think aloud.
2. The interviewer administers two or more practice questions and provides feedback to the respondent.
3. The interviewer administers the questionnaire,
   * following the questionnaire exactly,
   * allowing time for the respondent to complete his/her thoughts,
   * backchannelling, and
   * providing feedback and prompts to the respondent.
4. After completing the questionnaire, the interview administers several debriefing questions.

Interviewer Training.

Special interviewer training makes it possible to control how interviews are administered, as well as monitor the standard of interviewers (Bronkhorst 1991). In our experience, it is possible to use professional interviewers -- that is, interviewers with prior experience with personal interviews -- to administer think aloud interviews. The interviewers are given a full day of training on the techniques used to elicit concurrent protocols. The training exercises include watching and analyzing videotapes of think aloud interviews and role playing with each other. Near the end of the training period, the interviewer administers a pretest interview to an "expert" respondent (i.e., a respondent that mimics the characteristics of difficult respondents) to prepare an interviewer for a

"worst case" respondent and to enhance his/her basic skills.

In think aloud interviews, unlike intensive interviews, the dual role of interviewer and observer is completely eliminated. The interviewers do not have to be familiar with the questionnaire objectives, nor have any knowledge about question defects. Furthermore, since all interviews are monitored, the interviewer is not asked to identify question defects to the investigators. After one day's training, about 30% of professional interviewers are able to appropriately elicit concurrent protocols. Rather than spend additional time on training, we generally choose to train more interviewers than are actually needed for the data collection process.

Task Instructions

The task instructions ask the respondent to think aloud and give him/her the opportunity to ask questions about the process. Interviewers are trained to read the task instructions verbatim (but not to memorize them), and to look at the respondent during natural pauses (e.g., at the end of paragraphs), giving the respondent an opportunity to ask a question or backchannel. Interviewers are trained to answer respondents' questions by finding and re-reading the relevant portion of text. The interviewer is not required (and is actually forbidden!) to use his/her own words in answering respondent questions. This rule ensures that each respondent receives the same information from the task instructions.

Sample task instructions used to pretest a GTE residential telephone customer survey are shown in Exhibit 1. The task instructions are similar to the instructions suggested by Ericcson and Simon (1984), with modifications related to the specifics of each questionnaire pretesting situation. For example, mail questionnaires (in which respondents write their answers) have different task instructions than phone questionnaires (in which respondent speak their answers).

Backchannelling.

To encourage the respondent to think aloud, the interviewer must seem attentive and interested in the respondent's thoughts. The interviewer creates an environment in which

respondents feel free to take their time and express all of their thoughts by making eye contact, waiting during natural pauses and "backchannelling" -- that is, nodding or saying "uh-uh" or "okay" during natural breaks in the conversation. Backchannelling enables the interviewer to signal attention/interest in the respondent's thoughts and control the pace of the interview. Pauses longer than 3-5 seconds may make the respondent uncomfortable, but natural pauses of 3-5 seconds are common. Although 3-5 second pauses may seem uncomfortably long to the interviewer, he/she should be trained not to interrupt them. The respondent may still have further thoughts to say aloud.

The interviewer must also be trained not to "lead" the respondent by indicating agreement or disagreement (either through tone of voice, words or nods), or signal the respondent to stop speaking by looking down or away or making marks on the survey. Interviewers should not use leading questions or comments, such as "Would that be good or excellent?", interrupt the respondent, or use a tone that signals the respondent has "said enough."

Feedback and Prompts

The use of positive feedback enhances the quality of think aloud interviews. However, the type and timing of feedback must be controlled. An interviewer should only give positive feedback when the respondent is performing the think aloud methodology correctly. These prompts are very powerful because they encourage the respondent to keep doing what he/she has been doing. If used at inappropriate times, feedback will encourage the respondent not to think-aloud. We train our interviewers to use only two positive feedback phrases (after a pause of 3-5 seconds):

"You're doing exactly what I want you to do. Keep thinking aloud."

"You're giving me really good input. Keep thinking aloud."

Interviewers are trained to use two prompts during the think aloud interview. Unlike positive feedback, prompts reinforce the idea of thinking aloud (without criticizing the respondent).

"Remember, I'm interested in what you are thinking."

"Keep telling me what you are thinking."

These prompts should be provided relatively frequently throughout the interview. In our pretests, the interviewer's copy of the questionnaire has a prompt inserted after every second question. In addition, the interviewer is trained to prompt the respondent if he/she is silent for longer than three seconds after the question is asked.

Interviewers are trained to use only the four phrases described above for two reasons: (1) It will be hard to compare interviews because different interviewers are behaving differently; and (2) The respondent may interpret these comments as encouragement about his/her answer -- rather than encouragement about his/her thinking aloud. Although this procedure may seem repetitive and mechanical, only a few respondents have displayed irritation -- in hundreds of pretest interviews. However, to be effective, the interviewer must have appropriate timing and tone of voice.

Practice Questions

At least two practice questions are administered to allow the respondent to become accustomed to the think aloud process. If the practice questions go well, the entire interview is more likely to go smoothly. The questions designed to be similar in format to the actual survey questions, but they ask about a different topic. They provide the respondent with an opportunity to see if he/she correctly understood the task instructions, and to ask the interviewer questions. The practice questions allow the interviewer an opportunity to clarify the task instructions and provide feedback to the respondent about his/her performance. An example of a practice question that we use in pretesting our residential customer questionnaire is:

Overall, how would you rate the quality of the local newspaper's coverage of world events during the past three months? Would you say poor, below average, average, good or excellent?

(The response format is the same as the one used in the actual questionnaire.) We allow a special prompt for practice questions: If the respondent answers the question without thinking aloud, the interviewer may ask: "How did you arrive at that answer?" This prompt forces the respondent to

recall other thoughts that he did not say aloud. In addition, it encourages the respondent to fully express his thoughts in responding to subsequent questions.

AUTOMATIC CODING OF VERBAL REPORTS FROM PRETEST INTERVIEWS

Automatic coding is a computer-assisted procedure for counting the occurrence of pre-specified verbal and non-verbal cues contained within each respondent's answer to a particular survey item.[5] A coding category consists of a list of words, word strings, or non-verbal cues. Automatic coding has been thoroughly described in Bolton and Bronkhorst (1991), but we will briefly summarize it here.

Transcript Preparation.

The verbal protocols are transcribed into an electronic format. Then, to utilize our pretesting method, the protocols are segmented and precoded. A coder segments an individual's response to a question by listening to an audiotape of the interview and noting cues that indicate the end of one thought or the beginning of another. These cues include short pauses and changes in intonation. (An individual's response to a simple survey question usually consists of about ten separate "segments" or thoughts.) The coder also "precodes" the transcript by inserting ASCII markers for certain non-verbal cues, such as pauses, into the transcript. For example, pauses could be assigned the ":" marker, and broken utterances (dropping one thought for another) could be assigned the ">" marker. Coder reliability for segmenting and precoding simple non-verbal cues tends to be very high.

Coding Procedure

To analyze the segmented and precoded transcripts, we use a computer software package entitled Systematic Analyses of Language Transcripts, or SALT (Miller and Chapman 1986). However, other software packages have similar coding capabilities, such as the Oxford Concordance Program (Hockey and Marriott 1982), and the General Enquirer System (Kelly and Stone 1975; Stone, Dunphy and Smith 1966). The investigator inputs the transcripts and the coding

categories. SALT finds and marks the occurrence of each coding category for each segment. It also counts the frequency of each coding category for each questionnaire item for each respondent. This frequency can be divided by the number of segments spoken by the respondent to obtain a measure of the "intensity" of a coding category for a particular survey item. This division adjusts for the fact that some respondents are more verbose than others.

PRETESTING A SURVEY ABOUT TELEPHONE EXPERIENCES

As part of its quality efforts, GTE regularly conducts customer satisfaction surveys that are administered over the telephone. The "GSQ" survey elicits detailed reports from residential customers concerning certain telephone experiences, such as calls to directory assistance and long distance calls. For example, the following items are part of a sequence of questions about long distance calls:

a) Thinking only about the long distance calls you made, did you encounter noise or static on the line? (yes/no)

b) What kind of noise or static problems did you encounter? (crackling, frying, clicking, humming, buzzing, ...)

c) How often did this problem occur in the last 30 days? Would you say seldom, sometimes, often or almost always?

Most telephone experiences are low involvement events for residential customers, and -- as these examples suggest -- these particular experiences are typically low frequency events, as well. Consequently, there was some managerial concern that the GSQ survey might be eliciting inaccurate responses and providing poor quality information for policy decisions. Hence, GTE undertook a program of enhancements to the GSQ survey, including a questionnaire pretest. This paper focuses on selected portions of the GSQ: the overall quality question usually asked at the beginning of the questionnaire, a sequence of questions about long distance calls, and a sequence of questions about operator assistance (shown in Table 2). The overall quality question is an

important and useful "reference case" because many companies ask for a similar global evaluation in their customer satisfaction surveys. The questions about long distance calls and operator assistance attempt to elicit information about low frequency events that are (nevertheless) of interest to the telephone company.

The pretest used the split-ballot method to compare two versions of the GSQ. In this pretest, residential customers were recruited to participate in personal interviews at focus group facilities in Dallas, Texas. Experienced interviewers administered the questionnaires after one day of special instruction. Twenty two customers were administered one version (GSQ1) and 24 customers were administered another version (GSQ2). The interviews lasted 45 to 90 minutes, and they were audiotaped for subsequent analysis.

The audiotapes were transcribed into electronic form and the transcripts were segmented on the basis of short pauses, intonation and syntactical markers. The automatic coding scheme included sixteen verbal and non-verbal coding categories. It included modified versions of seven categories of verbal cues that had been used in prior research, denoted: repeat, similar, forget, no experience, confidence, can't say, and don't know. It also included seven new coding categories to measure respondents' cognitive difficulties forming answers to questions about low involvement, low frequency events. Two of the new categories were intended to capture respondents' reports about the frequency of events: low frequency and high frequency. We hypothesized that these categories would reflect respondents' usage of different retrieval strategies. Five of the new categories were intended to capture respondents' reports about their judgment processes: difficult, certain, no problems, problems, and expect. We hypothesized that the no problems, problems, and expect categories would reflect respondents' use of heuristics in their judgment processes. For example, casual observation had suggested that customers frequently answer ratings questions with statements such as, "I haven't had any problems, so I'd have to say 'excellent'" and the customer satisfaction literature had suggested that customers utilize predictive expectations. We also

hypothesized that the difficult and certain categories would measure their judgment difficulties. Altogether, SALT was used to mark the occurrence of 14 categories of verbal cues, displayed in Table 3. In addition, a single coder inserted markers for two categories of non-verbal cues: questions and pauses.

Tables 4, 5 and 6 show the average intensity measures associated with each of the sixteen coding categories for each question or question sequence.[6] The intensity measures are interpreted in the following way. Consider the first column of Table 4. On average, across all 22 respondents' answers to the overall quality question in GSQ1, 0.71% of the segments were questions and 3.06% of the segments were pauses. GSQ1 had been in the field for many years, so that most of our "standard" coding categories have low intensity measures, indicating that "major" question defects (e.g., comprehension difficulties) have been successfully eliminated.

In prior research, we have used an intensity level of 5% as a cut-off to indicate a potentially severe question defect because -- since the typical respondent speaks about ten segments in reply a survey questions -- it implies that about 50% of respondents uttered a single verbal cue in this category. Using observational monitoring, Fowler (1992) applies a cut-off criterion of 15% to identify defective questions. However, he is coding the entire response, rather than segments.

RESULTS

Statistical Tests

Overall Quality Question. The GSQ asks customers to rate the overall quality of services provided by GTE. It is positioned as the first question in both GSQ1 and GSQ2 to obtain a top-of-mind response, and it is also asked again as the last question in GSQ2 to obtain an in-depth response. GSQ1 asks the customer "would you evaluate . . . your local telephone company" whereas GSQ2 asks "would you rate . . . GTE." Table 4 provides a comparison of all three versions. Customers' responses to the GSQ2 versions have fewer pauses ($p < 0.10$) and more statements about high frequency events ($p < 0.05$). This result suggests that GSQ2 is eliciting a

memory based assessment by providing the GTE cue. It is interesting to note that there are significantly ($p < 0.15$) more verbal cues indicating lack of confidence when the overall quality question appears at the end of the survey. This finding suggests that the end placement of the question creates judgment difficulties because it encourages the customer to take into consideration all the issues that have been addressed by the survey.

Long Distance Sequence. In GSQ1, there are two separate sequences of questions about long distance service that distinguish between major carrier (e.g., AT&T, MCI) long distance service and GTE long distance service. In GSQ2, there is one sequence of questions that asks about all long distance service (i.e., both carrier and GTE). Table 5 shows a comparison of all three versions. Customers are more likely to report no experience, no problems, and low frequency in response to the sequence about GTE long distance. This result suggests that respondents have more difficulty retrieving information about GTE long distance service.

Operator Service Sequence. The operator service sequence asks about calls to directory assistance and toll operators. Virtually identical questions appear in both versions of the GSQ, but the bridging statements are different. The GSQ2 explains to the customer that he/she will be asked about both directory and operator services. Customers replied with consistently more no experience and low frequency verbal cues when answering GSQ1. This result indicates that customers are having trouble calling to mind experiences with operator services when inadequate retrieval cues are provided.

Discussion

In general, the no experience, low frequency, problems and no problems coding categories seem useful in measuring cognitive difficulties. For all three questions/sequences, they seem to detect the effect of improved retrieval cues in facilitating memory based (rather than stimulus based) judgments.

The low frequency and lack of confidence categories are high (> 5%) for both the long

distance and operator service sequences -- indicating retrieval and judgment difficulties. Since these findings occur across a number of different questions, and different versions of the same questions, these difficulties probably arise from the nature of the topic rather than the design of the particular survey items. It seems possible that (at least some) respondents are omitting or telescoping events, or using some other retrieval strategy to respond to questions on these topics. For example, respondents' answers to earlier questions about other telephone experiences are readily accessible in memory, and they may be used to form responses. It's also possible that respondents are using a variety of judgment heuristics. For example, respondents may generate evaluations by extrapolating from service experiences from other providers, such as electric utilities. Consequently, the study results show that the long distance and operator service sequences are likely to be subject to response errors.

Respondents generated very few verbal cues from the expect, difficult and certain coding categories. In retrospect, there seem to be good reasons for these results. Although there is an extensive literature that postulates that customers form assessments of service quality by comparing perceptions of performance with expectations (e.g., Parsuraman, Zeithaml and Berry 1985), expectations may be passive for low involvement services (e.g., Oliver 1989). This theory would explain why respondents' generated few cues from the expect category. It seems likely that few verbal cues were generated in the difficult category because respondents are unwilling to say that the survey task is difficult for self-esteem reasons. Lastly, although the certain category detected few cues, its opposite -- the lack of confidence category -- performed well. In particular, it indicates relatively high levels of evaluation difficulties (> 5%) for the operator service sequence.

CONCLUDING REMARKS

This paper has described the methodology that GTE has used in pretesting its various customer satisfaction surveys. The goal of our pretesting efforts is to improve questionnaire design in a low cost and timely fashion. Our experience indicates that pretesting costs are about 3% of

total survey costs -- and that these costs are more than offset by the cost savings due to improved questionnaire design (Bolton 1993). The cost implications are likely to be similar for other large scale survey efforts.

Our methodology is based on automatically coding concurrent verbal protocols collected in personal interviews. Unlike intensive interviews, the think aloud interviews are administered in a standardized fashion across respondents -- yet they provide a rich set of diagnostic information about how respondents answer survey questions. In contrast with observational monitoring, the pretest is administered under laboratory -- rather than field -- conditions. However, in common with observational monitoring, automatic coding provides clear-cut criteria for identifying question defects on the basis of quantitative measures of respondents' cognitive difficulties. These measures are particularly important for surveys of attitudes and opinions, for which there is no "objective" way to measure the extent of response bias.

The elicitation of concurrent protocols may create demand effects (e.g., Biehal and Chakravarthi 1989). Consequently, it isn't possible to assess how a questionnaire will perform under field conditions -- when subjects are not highly motivated and there may be many distractions. Hence, the elicitation of concurrent protocols is most appropriate for developmental pretests or for studies designed to detect subtle question defects. Some market research studies include very complex tasks, such as conjoint or discrete choice experiments, that are very different from conventional surveys administered by mail, phone or personal interview. In these studies, the elicitation of concurrent protocols may interfere with respondents' processing and it may be preferable to elicit retrospective protocols at natural stopping points or after the task is complete. Whether the investigator uses concurrent or retrospective protocols in the developmental pretest, a subsequent polishing pretest under field conditions will always be necessary. At GTE, we have generally used observational monitoring.

At this point, we have published four studies that show how our pretest methodology

provides diagnostic information about respondents' cognitive difficulties that identifies defective questions. The automatic coding categories have been transferable from one study to another, and the four underlying dimensions of cognitive difficulties have remained stable. In all four studies, it has been possible to apply quantitative criteria to identify superior questions. For example, the first GSQ2 wording of the overall quality question is superior to the GSQ1 wording in this study.

The unusual feature of this study is presence of detailed questions about low involvement, low frequency experiences -- such as long distance calls and operator service. The results show that customers utter certain categories of verbal cues -- specifically, low frequency and confidence categories -- in pretest interviews to indicate that they are experiencing cognitive difficulties forming responses questions about low involvement, low frequency telephone experiences. It seems likely that respondents are basing their (error prone) responses on partial recollections and evaluation heuristics.

GTE originally intended to streamline the GSQ by revising, collapsing or eliminating the most problematical questions. However, the entire survey was eventually suspended -- and it has yet to be reinstated. Management is considering alternative strategies for collecting this type of information. For example, directory assistance can be evaluated by "secret shoppers" -- that is, expert evaluators that call the operator and then evaluate the subsequent service contact.

The new coding categories developed for this study were somewhat successful. In particular, the low frequency and no problems categories should be useful in future research. They could be applied in conjunction with the eleven categories developed in earlier work (Bolton 1993), or in conjunction with completely new coding schemes. For example, Bickart and Felcher (elsewhere in this volume) have shown how content-based coding categories can be applied -- in conjunction with process-based coding categories -- to study retrieval strategies. There may be additional opportunities to use automatic coding to study the question answering process in other research contexts.

Exhibit 1

Task Instructions

[Respondent's name], we are conducting interviews with telephone company customers to test the design of a survey questionnaire. We plan to conduct a telephone survey of GTE residential telephone customers with this questionnaire, but before we do we want to gain a better understanding of how customers like you understand and respond to the questions in the survey.

Everything that we say here today is confidential and will only be used to help us test the questionnaire. We will be recording our conversation so that we can study this interview in more detail later, and for no other purpose.

I will be reading questions to you from the survey. However, I want you to answer these questions somewhat differently that you would in a regular survey. I want you to constantly think aloud while you are deciding about your answers. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you hear the question until you have given your final answer to the question. Thinking aloud will help us understand your thoughts about the question.

When I say think aloud, I mean say aloud everything that goes through your mind. You shouldn't worry if you sometimes feel that what you are thinking is not relevant to the question. I am interested in all your thoughts.

Do you understand what I am asking you to do? [If necessary, clarify by re-reading the sentences.]

Just act as if you are alone in the room talking to yourself. If you are silent for any length of time, I will remind you to keep talking. Thinking aloud may seem a little difficult at first, but then it should become very easy.

During the survey, if you want to go back to a question I asked earlier, either to change your answer or give more information, please feel free to do so.

Table 1

Questionnaire Pretesting Methods

Data Analysis Method	Data Collection Method	
	Verbal Reports	Non-Verbal Data
Qualitative	Intensive interviews Concurrent or retrospective protocols Paraphrasing Vignettes	Observational and Informational Methods
Quantitative	Observational monitoring Subjective Coding	Response latency measurement Card sorting tasks
	Automatic coding	

Table 2

Survey Questions

GSQ1	GSQ2-first	GSQ2-end
Single question asked at the beginning of the survey: How would you evaluate the overall quality of services provided by your local telephone company?	*Single question asked at the beginning of the survey:* How would you rate the overall quality of services provided by GTE?	*Single question asked at the end of survey:* Same as version two.
GSQ1a	GSQ1b	GSQ2
Sequence of questions about "long distance calls, that is, calls out of your local area" , including: How would you rate the quality of those long distance calls that you personally dialed from your telephone number during the past 30 days?	*Sequence of questions about "any long distance calls through your GTE Telephone Company*," including: How would you rate the quality of those GTE long distance calls that you personally dialed from this telephone number during the past 30 days?	*Sequence of questions about "long distance calls that you make out of your local area ... requiring 1+", including* : How would you rate the overall quality of long distance calls that you personally dialed from your telephone number during the past 30 days?
GSQ1	GSQ2	
Sequence of questions, beginning with: My last questions are concerning your directory assistance and operator service. Have you, personally, called a directory assistance or information operator from your telephone number to ask for a local number within the past 30 days.	*Sequence of questions, beginning with:* My next questions are services provided by telephone operators. I will be asking your opinions about the handling of directory assistance requests and the handling of calls requiring a long distance operator. Have you personally called a directory assistance operator for help	

	with telephone numbers or information in the past 30 days?	

Table 3

Coding Categories

REPEAT	HIGH FREQUENCY	NO PROBLEMS
repeat	always	no:problem
say:that:again	occasionally	no:problems
listen:to&again	everyday	don't:have&problem
hear:that:again	frequently	only:problem
ask:that	repeatedly	not&problem
what&mean	daily	not&problems
explain	hourly	never&problem
interpret	every:day	haven't&problem
define	perpetually	haven't&problems
comment	continually	few&problems
I:don't:think:I:got:you	constantly	didn't:have:problem
what&looking:for	incessantly	didn't:have:problems
what:was&again	at:all:times	no:complaint
do:you:mean	night:and:day	no:complaints
are: you:talking	commonly	don't&complaint
are&talking:about	habitually	don't&complaints
asking	in:general	only:complaint
about:the:question	several	never&complaint
that:word	normally	never&complaints
is:that:what:you're:asking	usually	few&complaints
problem&question	all:the:time	not&complaints
in:terms:of	every:time	didn't&complaint
so:it:says	often	didn't&complaints
in:other:words		haven't&complaint
how&evaluate		haven't&complaints
how&rate		not&complaint
I:misunderstood		nothing:wrong
I:misunderstand		anything:wrong
I:thought:you:said		something:wrong
don't:understand		didn't&wrong
		nothing:wrong
		never&wrong
		not&wrong
		hasn't&wrong
		no:mistake
		no:mistakes
		didn't&mistake
SIMILAR	LOW FREQUENCY	CAN'T SAY
identical&question	seldom	I:can't:say

same&question similar&question identical&answer same&answer similar&answer answer&again sound&alike question&close sound&identical I&answered&that&questi on like:I:said like:I:said&before experience:again opinion:again there:again repetitious	never sometimes at:times every:now:and:then once:in:a:while now:and:then time:to:time often:enough every:so:often few rarely little one:or:two once one:time one:instance one:incident	I:can't:tell I:can't:rate I:can't:evaluate I:can't:judge tough&rate not:easy&rate difficult&rate hard&rate tough&evaluate not:easy&evaluate difficult&evaluate hard&evaluate tough:to:say not:easy&to:say difficult:to:say hard:to:say rough&judge not:easy&judge difficult&judge hard&judge tough&to:tell not:easy&to:tell difficult&to:tell hard&to:tell
FORGET forget don't:remember can't:think I'm:trying:to:think	EXPECT I&assume I&expect assumption anticipate predict expectation	DON'T KNOW I:don't:know I:wouldn't:know not:know
NO EXPERIENCE no:experience never:experienced not:experienced any:experience never:experience not:experience haven't:experienced not:familiar:with no:need:for&service no:need:for&option never:use never:done	CONFIDENCE probably approximately maybe perhaps I:guess kind:of assume unless somewhere:in:there I:reckon not:certain I&imagine	CERTAIN I'm:certain definitely I'm:sure exactly

never:used don't:use haven't:used	depends mostly sort:of not:sure whatever or:something	
PROBLEMS problem problems complaint complaints mistake mistakes qualm qualms discrepancy discrepancies wrong	DIFFICULT difficult&question hard&question tough&question that's:a:hard:one that's:a:tough:one hard:to:deal:with not:easy&question isn't:easy&question	

Table 4

Overall Quality Question

Category	GSQ1	GSQ2-first	GSQ2-end	F-test
COMPREHENSION				
Questions	0.71	1.64	3.68	
Repeat	0.00	0.13	0.78	
Similar	0.00	0.34	0.78	
RETRIEVAL				
Forget	0.49	0.45	0.43	
Pause	3.06	0.91	0.00	$p < 0.10$
No Experience	0.74	0.13	0.00	
Low Frequency	7.07	5.63	4.71	
High Frequency	1.10	3.88	0.57	$p < 0.05$
JUDGMENT				
Confidence	4.18	3.94	8.21	$p < 0.15$
Difficult	0.00	0.25	0.00	
Certain	0.00	0.44	1.60	
Problems	4.20	4.18	3.47	
No Problems	6.56	9.17	5.31	
Expect	0.30	0.59	0.46	
RESPONSE				
Can't Say	0.00	0.99	0.00	
Don't Know	2.52	0.66	1.25	

Table 5

Long Distance Sequence

Category	GSQ1a	GSQ1b	GSQ2	F-test
COMPREHENSION				
Questions	4.03	2.55	3.55	
Repeat	1.42	1.43	1.68	
Similar	0.83	0.71	0.70	
RETRIEVAL				
Forget	0.90	0.43	0.92	
Pause	1.53	1.38	1.91	
No Experience	0.21	0.61	0.05	$p < 0.10$
Low Frequency	7.49	9.18	4.80	$p < 0.10$
High Frequency	3.64	4.76	2.70	
JUDGMENT				
Confidence	5.67	7.05	7.86	
Difficult	0.00	0.02	6.00	
Certain	0.17	0.52	1.00	
Problems	3.08	4.73	1.85	$p < 0.05$
No Problems	2.95	4.98	2.23	$p < 0.05$
Expect	0.17	0.07	0.81	
RESPONSE				
Can't Say	0.00	0.06	0.18	
Don't Know	2.00	2.45	4.23	

Table 6

Operator Service Sequence

Category	GSQ1	GSQ2	F-test
COMPREHENSION			
Questions	2.28	3.74	
Repeat	0.57	1.18	
Similar	0.58	0.13	
RETRIEVAL			
Forget	0.41	0.12	
Pause	1.33	1.11	
No Experience	1.20	0.00	$p < 0.10$
Low Frequency	8.46	4.76	$p < 0.10$
High Frequency	4.32	2.94	
JUDGMENT			
Confidence	7.19	6.94	
Difficult	0.00	0.00	
Certain	0.94	0.51	
Problems	2.09	2.17	
No Problems	1.44	2.81	
Expect	0.00	0.06	
RESPONSE			
Can't Say	0.00	0.00	
Don't Know	1.24	1.25	

REFERENCES

Bercini, Deborah (1989), "Observation and monitoring of interviews," Quirk's Marketing Research Review (May).

Bickart, Barbara A., Johnny Blair, Geeta Menon and Seymour Sudman (1990), "Cognitive Aspects of Proxy Reporting of Behavior," Advances in Consumer, 17, 198-206

Biehal, Gabriel and Dipankar Chakravarthi (1989), "The Effects of Concurrent Verbalization on Choice Processing," Journal of Marketing Research, 26, 84-96.

Blair, Edward and Scot Burton (1987), "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions," Journal of Consumer Research, 14 (September), 280-288.

Bolton, Ruth N. (1991), "An Exploratory Investigation of Questionnaire Pretesting with Verbal Protocol Analysis," Advances in Consumer Research, 18, 558-65.

_____ (1993), "Pretesting Questionnaires: Content Analyses of Respondents' Concurrent Verbal Protocols," Marketing Science, 12 (3), 280-303.

_____ and Tina M. Bronkhorst (1991), "Quantitative Analyses of Depth Interviews," Psychology and Marketing, 8 (4), 275-97.

Bronkhorst, Tina M. (1991), "Questionnaire Pretesting Methodology: Interviewer and Supervisor Training," GTE Laboratories Technical Memorandum 0464-10-91-420.

Cannell, Charles and Lois Oksenberg (1988), "Observation of Behavior in Telephone Interviews," Telephone Survey Methodology, Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg, eds., New York: John Wiley and Sons, 475-96.

Campanelli, P., E. Martin and K. Creighton (1989), "Respondents Understanding of Labor Force Concepts: Insights from Debriefing Studies," Proceedings of the Fifth Annual Census Bureau Research Conference.

Ericsson, K. Anders and Herbert A. Simon (1984), Protocol Analysis: Verbal Reports as Data, Cambridge, MA: MIT Press.

Felcher, E. Marla and Bobby Calder (1990), "Cognitive Models for

Behavioral Frequency Survey Questions, Advances in Consumer Research, 17, p. 207-211.

Feldman, Jack M. and John G. Lynch, Jr. (1988), "Self-Generated Validity and Other Effects of Measurement on Belief, Attitude, Intention and Behavior," Journal of Applied Psychology, 73 (3), 421-435.

Fowler, Floyd Jackson (1989), "Coding Behavior in Pretests to Identify Unclear Questions," Health Survey Research Methods, Department of Health and Human Services Publication No. 89-3447, p. 9-12.

_____ (1992), "How Unclear Terms Affect Survey Data," Public Opinion Quarterly, 56, 218-231.

Hockey, S. and Marriott, I. (1982), Oxford Concordance Program Version 1.0 Users' Manual, Oxford: Oxford University Computing Service.

Hunt, Shelby D., Richard D. Sparkman, Jr., and James B. Wilcox (1982), "The Pretest in Survey Research: Issues and Preliminary Findings," Journal of Marketing Research, 19 (May), 269-73.

Jobe, Jared B. and David J. Mingay (1990), "Cognitive Laboratory Approach to Designing Questionnaires for Surveys of the Elderly," Public Health Reports, (5), 518-24.

Jonides, John and Moshe Naveh-Benjamin (1987), "Estimating Frequency of Occurrence," Journal of Experimental Psychology: Learning, Memory and Cognition, 13 (2), 230-40.

Kelly, E. F. and P. J. Stone (1975), "Computer recognition of English Word Senses," Amsterdam: North Holland.

McFarland, Sam G. (1981), "Effects of Question Order on Survey Responses," Public Opinion Quarterly, 45, 208-15.

Miller, J. and Chapman, R. (1982), "Systematic Analysis of Language Transcripts (SALT)," Unpublished manuscript, University of Wisconsin.

Oksenberg, Lois, Charles Cannell and Graham Kalton (1991), "New Strategies for Pretesting Survey Questions," Journal of Official Statistics (7), 349-365.

Oliver, Richard L. (1989), "Processing of the Satisfaction Response in Consumption: A Suggested Framework and Research Propositions," Journal of Satisfaction, Dissatisfaction and Complaining Behavior, H. K. Hunt and R. L. Day, eds. Bloomington: School of Business, Indiana University.

Parasuraman, A. Valarie A. Zeithaml and Leonard L. Berry (1985), "A Conceptual Model of Service Quality and Its Implications for Future Research," Journal of Marketing, 49 (Fall), 41-50.

Royston, Patricia (1987), "Application of Cognitive Research Methods to Questionnaire Design, Paper Presented at the Society for Epidemiological Research Twentieth Annual Meeting.

_____ (1989), "Using Intensive Interviews to Evaluate Questions," Conference Proceedings: Health Survey Research Methods.

_____, Deborah Bercini, Monroe Sirken and David Mingay (1986), "Questionnaire Design Research Laboratory," Paper Presented at the Meetings of the American Statistical Association.

Schwarz, Norbert (1990), "Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction," Research Methods in Personality and Social Psychology, Clyde Hendrick and Margaret S. Clark (eds), Newberry Park, CA: Sage Publications.

Stone, P.J., D.C. Dunphy and M.S. Smith (1966), The General Inquirer: A Computer Approach to Content Analysis, Cambridge, MA: MIT.

Tversky, Amos and Daniel Kahneman (1973), "Availability: A Heuristic for Judging Frequency and Probability," Cognitive Psychology, 5, 207-232.

Willis, Gordon B., Patricia Royston and Deborah Bercini (1991), "The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires," Applied Cognitive Psychology, 5, 251-267.

1. Subjective coding also relies on *expert interpretation* of verbal reports to trace cognitive *processes*, but it has not been used in a pretesting context.

2. Very detailed information on the time required to automatically code interviews is provided by Bolton and Bronkhorst (1991). For example, it took about 40-50 hours to develop our first, simple automatic coding scheme of six categories. However, these six categories were used -- with minor modifications -- to pretest many different questionnaires over the next two years.

3. Bolton and Bronkhorst (1991) and Bolton (1993) applied both observational monitoring and automatic coding techniques similar to those described by Cannell and Oksenberg (1988) to analyze concurrent protocols. Their results indicated that automatic coding generally agrees with observational monitoring in identifying comprehension difficulties ($p < 0.01$). However, Bolton (1993) also showed that automatic coding detects retrieval and judgment difficulties that observational monitoring fails to detect because it has a richer set of codes and exploits non-verbal cues (which human coders have difficulty applying in real time interviews). Observational monitoring is better at identifying response difficulties because it requires an expert assessment of whether an answer is "adequate".

4. Probes and follow up questions may create unpredictable order effects (McFarland 1981). In addition, Campanelli, Martin and Creighton (1989, p. 372) make an important point: "It is somewhat ironic that we have documented the lack of standardization of meaning using standardized questions. At the very least, this seems to require us to acknowledge that our debriefing questions themselves may be subject to various unintended interpretations."

5. It is worthwhile to point out that automatic coding methods can be applied to any verbal data. In a pretesting context, there are likely to be some instances in which the researcher will choose to elicit retrospective protocols to study the question answering process. Retrospective protocols may be appropriate if taken immediately after a brief processing episode and instill no demand effects. A benefit of retrospective protocols is that they do not alter the natural survey condition (Blair and Burton 1987).

6. The statistic reported for the *problems* category has been adjusted. It was calculated by subtracting the average intensity for the *no problems* category from the (unadjusted) average intensity for the *problems* category. In other words, the adjusted statistic indicates the intensity with which respondents talked about problems, rather than both problems and lack of problems.