

Pretesting Questionnaires:  
Content Analyses of Respondents'  
Concurrent Verbal Protocols

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

One Line Abstract: How to identify and revise defective survey questions by using computer-assisted coding to diagnose the cognitive difficulties that respondents experience.

Key words: Information Processing, Measurement, Survey Research.

\* The author gratefully acknowledges the research assistance of Tina M. Bronkhorst and the helpful comments and advice of the editors and anonymous reviewers of Marketing Science.

PRETESTING QUESTIONNAIRES:  
CONTENT ANALYSES OF RESPONDENTS' CONCURRENT VERBAL PROTOCOLS

Abstract

Conventional questionnaire pretesting methods focus on directly identifying question defects, such as an ambiguous question. This paper proposes a new method that identifies respondents' cognitive difficulties as they form answers to survey questions. It entails a content analysis of concurrent verbal protocols elicited during pretest interviews. The effectiveness of the methodology is illustrated with pretests of multiple versions of the same survey. The results are used to illustrate how this method yields diagnostic information about questionnaire problems and improvements. Then, the results are compared with the results of observational monitoring by managers. The findings indicate that a questionnaire pretesting methodology that quantifies respondents' cognitive difficulties is a useful enhancement for identifying and "improving" defective questions.

## INTRODUCTION

Certain cognitive processes operate when a respondent answers a survey question. A respondent must comprehend (i.e., encode) the question, retrieve information from memory, weigh the information and form a response. If a respondent experiences cognitive difficulties, the response to the question may contain some element of error. A pretest is a small, pilot study to determine how a questionnaire can be improved to minimize response errors (Converse and Presser 1986), such as a respondent misinterpreting a question. Questionnaire pretests can be very important because response and other nonsampling errors are the major contributors to total survey error (Assael and Keon 1982).

Individual questions can be pretested for an acceptable level of response variation, meaning, task difficulty, and respondent interest/attention. The overall questionnaire can be pretested to assess the "flow" and naturalness of the sections, the order of questions, skip patterns, timing, and respondent interest and attention. These tests enable the researcher to identify and change questionnaire design features, such as vocabulary, response alternatives, and skip patterns, to minimize response errors and non-response errors.

Conventional questionnaire pretesting methods focus on directly identifying question defects. For example, Hunt, Sparkman and Wilcox (1982) subjectively coded question defects, such as ambiguous questions. This paper proposes a new methodology that identifies respondents' cognitive difficulties as they form answers to survey questions. This method entails a content analysis of concurrent verbal protocols elicited during pretest interviews. The empirical portion of the paper addresses the following questions: Can a content analysis of verbal protocols elicited during questionnaire pretests be used to detect defective survey questions? Is an automatic coding scheme that traces respondents' thought processes useful in revising defective questions?

Can quantitative analyses of the coded data identify "improvements" in questionnaire phrasing? How do the results compare with a traditional pretesting method, observational monitoring?

The paper reviews relevant prior research concerning questionnaire pretesting. The next two sections develop a new questionnaire pretesting methodology. Then, iterative pretests of different versions of the same survey illustrate how this method yields diagnostic information about questionnaire problems and improvements. The results are compared with the results from a "traditional" pretest of the same survey that relied on observational monitoring by managers. The remaining sections discuss the situations in which the new questionnaire pretesting methodology will be useful and appropriate.

## BACKGROUND

### Pretesting Strategy

There is very little guidance about *pretesting* questionnaires. Converse and Presser (1986, p. 52-75) recommend iterative pretesting with small sample sizes. Most researchers agree that the first pretest should be administered by personal interview, even if the questionnaire will ultimately be administered by mail or telephone. In this "developmental" pretest, trained interviewers administer an outside version of the survey instrument to members of the target population. After the questionnaire has been revised and evaluated by experts, Converse and Presser recommend that a "polishing" pretest should be administered to members of the target population according to the intended survey administration method.

In 1983, researchers began collaborating on the application of cognitive psychology to survey research (Jadine, Straf, Tanur and Tourangeau 1984; Hippler, Schwarz and Sudman 1987). Subsequently, the National Center for Health Statistics (NCHS) initiated the use of process tracing methods to investigate how question characteristics may affect the recall and estimation strategies used by respondents (Lessler and Sirken 1985; Sirken et al 1988; Willis et al 1991). A variety of cognitive research methods have been used to pretest questionnaires: concurrent "think aloud" interviews, paraphrasing, retrospective "think aloud" interviews, confidence ratings, vignettes, response latency measurements, and sorting tasks (Campanelli, Martin and Creighton 1989; Jobe and Mingay 1990; Royston, Bercini, Sirken and Mingay 1986; Royston 1987). The cognitive research methods that are most frequently applied to questionnaire pretesting entail the elicitation of verbal reports.

#### Pretesting Tactics

Concurrent protocols can be a rich source of information about respondents' cognitive processes. However, protocol generation lengthens task time and it may also change respondents' primary processes (Biehal and Chakravarti 1989; Russo, Johnson and Stephens 1989). Research has not yet identified the task conditions under which the elicitation of concurrent protocols will not disrupt primary processes. Russo, Johnson and Stephens (1989) suggest that subjects should be trained in verbalization prior to data collection and instructed to preserve naturalness over completeness; non-directive prompting by the interviewer should also be minimized.<sup>1</sup> Previous authors have recommended retrospective protocols when they can be taken immediately and the processing episode is brief; and demand effects that would lead respondents to distort the process do not seem likely for most respondents (e.g., Blair and Burton 1987). However, Russo, Johnson and Stephen's (1989) study also raises questions about the veridicality of retrospective protocols.

Campanelli, Martin and Creighton (1989, p.372) comment about debriefing questions and retrospective probes in pretests: "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."

Although pretests have received little methodological attention, Hunt, Sparkman and Wilcox (1982) conducted an early test that evaluated traditional methods of pretesting. They found that subjectively coding respondents' verbalizations was not very successful in detecting defective questions. Detection rates given the presence of an error were low, and depended on the type of error (e.g., loaded question) and the method of pretest (i.e., face-to-face interviews or telephone interviews). In particular, their research evidence suggested that concurrent protocols were as ineffective as retrospective protocols. Surprisingly, more errors were detected in telephone interviews relative to face-to-face interviews.

#### Pretest Data Analysis

Content analyses of verbal protocols have been used to study cognitive processes and responses to survey questions (e.g., Belson 1981; Weber 1985), but they have seldom been used to pretest questionnaires. There are two possible approaches to a content analysis of pretest interviews: subjective coding and automatic coding. In subjective coding, a trained coder(s) or "expert" reads a verbalization or segment and assigns a pre-determined code on the basis of his/her understanding of the text. In automatic coding, a

computer finds and marks all the cues predetermined by the researcher. The computer ensures that the underlying vocabulary and inference rules are applied consistently. Note that subjective and automatic coding can be applied to precoded non-verbal cues, as well as verbal cues, that take place during the pretest interview.

Subjective coding is particularly appropriate for testing hypotheses about specific cognitive processes and how these process affect the formation of a response. For example, it can be used to study how respondents recall information (e.g., Bickart et al 1990), such as their use of a reference period or their handling of inaccessible information. Automatic coding is particularly suited to providing objective descriptions of the content of respondents' verbal reports. For example, automatic coding can be used to study respondents' ability to understand certain words and phrases or their ability to retrieve information when provided with certain cues. It is also useful for testing the robustness of a coding scheme to changes in vocabulary and rules (Ericsson and Simon 1984). Both coding methods have limitations. Subjective coding requires expert interpretation of respondents' verbal reports. Automatic coding typically relies on relatively simple coding rules that do not completely capture the richness of verbal reports.

Pretest interviews are typically evaluated by eliciting information from interviewers, observing the interviews, coding or tallying answers. When quantitative analyses of pretest data have been conducted, researchers have used subjective coding schemes based on the entire verbalization. These coding schemes are usually based on some sort of typology of defective questions. Willis et al (1991) detected 39 problems in an health survey questionnaire concerning assistive devices, and categorized 31% as structural problems and 69% as cognitive problems. Hunt, Sparkman and Wilcox's (1982) study subjectively encoded five types of faulty questions: loaded questions, double questions, ambiguous questions, inappropriate vocabulary, and missing alternatives. An error identification was scored if the respondent made comments "which could help in recognizing the error" (p. 272). Most pretests rely on observational monitoring -- i.e., coding the verbal interactions between the interviewer and the respondent *as the interview takes place* (e.g., Bercini 1989; Cannell and Oksenberg 1988; Morton-Williams and Young 1987).

#### METHODOLOGY

The available research evidence indicates that current questionnaire pretesting methods that attempt to directly identify question defects are effective in identifying some types of defective questions, but not others. An emerging theme in discussions of pretesting is that a *developmental pretest should use cognitive research methods to trace a respondent's thought processes as he/she forms an answer to a survey question*. This section proposes a new questionnaire pretesting method and discusses how it might be compared with traditional pretesting methods.

#### Identifying Cognitive Difficulties

The core of the proposed questionnaire pretesting methodology is a coding scheme that identifies respondents' cognitive difficulties arising during the response process. Researchers have proposed a variety of cognitive models of the response process (e.g., Kuncel 1981; Rogers 1973; Turner and Fiske 1965; Schwarz 1990; Willis 1991). The coding scheme used in this study is based on a conceptualization of the attitude formation process developed by Tourangeau (1984, 1987; Tourangeau and Rasinski 1988). Tourangeau proposed that the respondent's answer to an attitude question is the product of four stages or

macroprocesses: comprehension of the question, the retrieval of relevant beliefs and feelings from memory, the weighing of information to form a judgment, and the selection of an appropriate response alternative.

The objective of the coding scheme is to trace cognitive processes and responses as the respondent forms an answer, so the appropriate unit of analysis is a speech burst or utterance -- rather than the complete response to a question. Respondents' speech can be segmented into speech bursts using short pauses, intonation, and syntactical markers as cues. Speech bursts can be subjectively or automatically encoded depending on the goals of the pretest.

In general, the research questions addressed by a *developmental* pretest are relatively basic: Does the respondent understand the question? Is it difficult for the respondent to recall the appropriate information? Can he/she give an answer to the question? These research questions can be answered with descriptive information concerning respondents' verbal reports. Hence, this study codes respondents' speech bursts with an automatic, rather than a subjective, coding scheme.

#### Development of the Coding Scheme

The coding scheme was developed in two pilot studies because automatic coding schemes have not been used in prior pretests. The first pilot study developed seven coding categories that corresponded to response errors associated with three of the four macroprocesses: comprehension, retrieval, and evaluation/judgment (Bolton 1991). It involved face-to-face pretest interviews with 21 residential telephone company customers during 1989. The coding scheme was created from the domain of verbal protocols elicited during these interviews. Each coding category consisted of lists of word strings or non-verbal cues (e.g., a pause) postulated to be indicators of the underlying macroprocess. The coding scheme was iteratively revised by three investigators that reviewed the coded transcripts. Ultimately, the coding categories were used in a content analysis of respondents' verbalizations, and the results were used to identify defective questionnaire items.

A second (unpublished) pilot study was conducted to revise the coding scheme and to develop categories to reflect response difficulties (Bolton and Bronkhorst 1990) -- that is, respondents' inability to provide an answer. As before, face-to-face interviews were conducted with 50 residential telephone company customers to pretest a (different) survey. Again, an iterative process was used to revise and test the coding scheme. One of the seven categories used in the first pilot was dropped, three categories were added and two coding categories were revised by "narrowing" the domain of key words and strings (i.e., more restrictive strings).

The coding scheme ultimately consisted of nine categories: five verbal categories and four nonverbal categories. The five verbal categories are shown in Table 1; the four non-verbal categories are questions, pauses, broken utterances and unintelligible utterances. The coding scheme does not provide codes for all segments because it is not intended to completely describe the processing strategy of a respondent. There may be some overlap between codes that identify different information processing difficulties. However, codes do not have to unambiguously associated with a specific processing problems because the pretesting methodology relies on multiple codes to identify each problem. As described below, factor scores are created for each processing difficulty, so that each processing problem is identified on the basis of all nine coding categories.

-----

Table 1 here

Coding Categories

Comprehension. A question is defective if the respondent's interpretation does not match what the researcher intended (Belson 1981; Schwarz 1990). A respondent's ability to comprehend a survey question should be facilitated when the survey provides contextual information, such as bridging statements, a logical sequence of questions, groups of related questions and explicit (perhaps lengthy) questions (Tourangeau 1984). For example, oral presentation -- rather than written presentation -- of lengthy questions (i.e., with many qualifiers) may cause comprehension problems. The "best" measure of comprehension difficulty in the pilot studies was a nonverbal cue, namely an interrogative inflection or question by the respondent. A second measure of comprehension difficulty was provided by a category of verbal cues related to respondents' requests to the interviewer to repeat or clarify a question, such as "Repeat the question" and "Say that again." This category was newly developed for this study.

Retrieval. Surveys frequently elicit memory based (rather than stimulus based) judgments that necessitate the retrieval of information (rather than earlier judgments). A respondent's ability to retrieve information depends on whether the cues provided in the context of recall match the cues available during coding. The respondent's failure to retrieve information can occur when the relevant information was not stored in long term memory, it cannot be retrieved from available cues, or it is difficult to distinguish from related information (Tourangeau 1984; Schwarz 1990).

This study utilized three coding categories for retrieval difficulties. A pause (greater than 3 seconds) is a non-verbal cue that may indicate more than one cognitive processing difficulty. Since the pilot studies found that pauses were associated with coding categories that included verbal cues that indicated that respondents had difficulty recalling information, a pause was considered a coding category for retrieval difficulties. A second category, labelled forget, consisted of word strings such as "I forget" and "I can't recollect." In the pilot study, respondents uttered verbal cues from the forget category less often in response to a question that supplied contextual cues than in response to a similar question that did not (Bolton 1991). This study investigated the role a third coding category, the broken utterance. This non-verbal cue is characterized by a speaker stopping in mid-utterance and beginning a new utterance (e.g., "I've never -- oh, I remember"). A broken utterance may indicate changes in thought processes, such as sudden retrieval of information from memory or a shift from a retrieval to a judgment process. Survey questions that elicit broken utterances are likely to be questions for which the respondent's "top of mind" and "in depth" responses are different. They are usually more suitable for an in-depth interview or a mail questionnaire than for a telephone interview.

Judgment. There is considerable evidence that respondents' judgment heuristics and biases may lead to response errors (Schwarz 1990). In studies of the processing strategies used for behavioral frequency questions, researchers have observed that respondents' verbalizations are characterized by statements of uncertainty (e.g., Blair and Burton 1987). This observation suggests judgment difficulty may be captured by a category of verbal cues that indicate the respondent's lack of confidence in his/her answer to the question.

In both pilot studies, words such as "maybe," "perhaps," or "probably," were

cues that the respondent lacked confidence about how to answer the question. A second category of cues that may indicate a respondent is experiencing cognitive difficulty in forming a judgment is the unintelligible utterance (e.g., "um," "er"). However, they may also be associated with other non-verbalized processes (e.g., retrieval processes or internal computation) or speech particulars (i.e., vernacular).

Response. A respondent may understand a question and be able to retrieve and evaluate information from memory, but he/she may not be able to formulate a response. For example, double questions, ambiguous questions and questions with missing alternatives can lead to response difficulties (Payne 1951). The second pilot study indicated that respondents tended to utter verbal cues such as "can't say" or "don't know" to articulate such difficulties in formulating responses. In addition, respondents may experience multiple information processing difficulties that culminate in response difficulty.

#### Comparing with Traditional Pretesting Methods

The preceding paragraphs have proposed a new pretesting method based on the automatic coding of verbal and non-verbal cues. It is difficult to compare this method with traditional methods because "we have almost no rigorous knowledge about how to pretest" (Presser 1989, p. 35). Subjectively coded pretest interview data should be a natural basis for comparison with automatically coded pretest data. However, most subjective coding schemes have been developed to test hypotheses about processes (e.g., Blair and Burton 1987), not to pretest questionnaires. There are very few studies that describe a rigorous subjective coding scheme for pretesting and they are difficult to generalize to other study contexts. Willis et al's (1991) study relies on intensive verbal probing (e.g., "What does the term abdomen mean to you?"). Hunt, Sparkman and Wilcox's (1982) coding scheme is relatively narrow in scope because it was developed for five types of artificially generated, defective questions. Hence, observational monitoring -- that is, subjective coding in "real time" rather than using transcripts and/or audiotapes -- was selected as the basis of comparison.<sup>ii</sup>

Observational monitoring was originally used to monitor and evaluate interviewer performance (Cannell, Lawson and Hausser 1975), but it has subsequently been used to investigate respondent-interviewer verbal interactions (e.g., Oksenberg, Cannell and Kalton 1991). It is a "traditional" pretesting method with well-established coding schemes for verbal behavior. Observational monitoring has several desirable features in common with the proposed pretest methodology. It relies on concurrent verbal protocols and (typically) utilizes a coding scheme for respondents verbal behavior that is organized around cognitive processes. Hence, automatic coding and observational monitoring can be applied to identical pretest interview data to make inferences about cognitive difficulties and detect defective questions.

#### AN ILLUSTRATION

A study was designed to address the following research questions.

1. Is an automatic coding scheme based on verbalizations elicited during pretests useful in identifying the information processing difficulties that arise as respondents form answers to survey questions?
2. Is a coding scheme that traces information processing difficulties useful in detecting defective questions?
3. Can the information provided by a content analysis of respondents'

verbalizations be used to identify superior or inferior questions?

4. How do the results produced by this method compare with the results of obtained from observational monitoring?

The first three questions are investigated by applying the proposed pretest methodology and analyzing the results to make inferences about respondents' cognitive difficulties and to identify defective questions. The fourth question is investigated by comparing these results with the results from observational monitoring. These questions are addressed by a study in which GTE pretested three different versions of its business customer survey. The following paragraphs describe the background of the survey, the design of the study, the method of collecting pretest interview data, and the coding procedures used for both automatic coding and observational monitoring.

## Background

GTE regularly conducts customer satisfaction surveys that are administered over the telephone. In 1989, as part of revisions and pretests of GTE's residential surveys, GTE conducted the two pilot studies that developed the automatic coding scheme. In 1990, the entire pretest methodology was evaluated in two pretests that evaluated different versions of GTE's business customer survey. The first pretest used the split-ballot method to compare the current version (CV) of the questionnaire with a proposed revision (RV). The CV of the questionnaire had been used by GTE for four years. The RV had been prepared to correct certain question defects identified by management. Based on the results of this pretest, additional question defects were identified and a new version (NV) of the questionnaire was prepared. The second pretest evaluated the NV and tested alternative wordings of certain questions using the split ballot method.

## Data Collection Procedures

Telecommunications decision-makers in small businesses were recruited to participate in personal interviews. In the first pretest, one-on-one interviews were conducted with 28 customers at focus group facilities in Dallas and Los Angeles. Thirteen customers were administered the CV and 15 customers were administered the RV. In the second pretest, one-on-one interviews were conducted with 30 customers at focus group facilities in Los Angeles. All customers were administered the NV. For certain questions, 15 customers were asked one wording and 15 were asked an alternate wording. The small sample of 58 customers (in total) is consistent with traditional pretesting practice.

Experienced interviewers administered the questionnaires after receiving one day of special training. They were trained to elicit concurrent protocols by strictly following the questionnaire, providing reinforcement, and backchannelling.<sup>iii</sup> The interviewers began by reading task instructions for the elicitation of concurrent protocols. These instructions are similar to the instructions used in other contexts (e.g., Ericsson and Simon 1984, p.376); they ask the respondent to "constantly think aloud while you are deciding about your answers." Then, the interviewer administered two practice ratings questions similar in format to the actual survey questions.

The introductory sentences of the survey provided the transition to the actual survey questions. An example of a survey question is the following: "Overall, how would you rate the quality of telecommunications GTE has provided to your company over the past three months? Would you say A for excellent, . . . or F for poor?" The interviewer used the phrases "Remember, I'm interested in what you are thinking," and "You're doing a good job of thinking aloud" prior to every second question to reinforce the respondents' behavior. The interviewer did *not* use any intensive probes. The interviews lasted 45 to 90 minutes, and they were audiotaped for subsequent analysis.

GTE executives watched the pretest interviews from behind one-way mirrors. After the first pretest, they became interested in formally recording their impressions. Consequently, two managers monitored the pretest interviews and recorded the occurrence of certain predefined verbal behaviors during the second pretest (i.e., of the NV) only. They were trained in observational monitoring techniques at a one-day practice session that included instructions about how to use an observational monitoring form, opportunities to code pre-recorded (audiotaped or videotaped) interviews, and feedback on their performance. Since practice sessions indicated that reliability was reasonably high, a single coder was assigned to monitor each interview.<sup>iv</sup>

### Coding Procedures

Observational Monitoring (OM). The OM techniques used at GTE are similar to traditional pretesting methods in that respondents' entire verbalizations are subjectively encoded. The coding scheme is similar to other OM coding schemes used in pretesting (e.g., Bercini 1989). Four codes are used to record the respondent's behavior; two are used to record the interviewer's behavior; and one code is used to record "managerial issues." Since this study concerns problems with questions (rather than interviewers), it focuses on the four respondent codes.

\*Comprehension difficulties are coded whenever a respondent experiences difficulty in understanding a question.

\*Retrieval difficulties are coded whenever a respondent says they can't remember or pauses for more than three seconds during speech.

\*Judgment difficulties are coded whenever a respondent makes comments about the question, such as remarking on the difficulty in answering a questions or on the similarity of two questions.

\*Response difficulties are coded whenever a respondent gives an inadequate answer or replies "don't know," "can't say," etc.

The NV pretest interviews were coded by managers as the interviews took place using a paper version of the questionnaire in which codes were listed in the left hand margin. This design allows the coder to follow the questions. The codes are quickly and easily tabulated. The percentage of respondents experiencing a particular cognitive difficulty for a given question is obtained by dividing the number of respondents for which the relevant category was coded by the total number of respondents.

Automatic Coding (AC). A subset of questions were selected for detailed analysis. Audiotapes of the customers' responses to these interview questions for each version of the questionnaire (i.e., CV, RV and NV) were transcribed into electronic form and the transcripts were segmented for analysis. The respondents' speech was segmented using short pauses, intonation, and syntactical markers (for complete phrases, clauses or sentences) as cues. Since the subjective precoding of simple non-verbal cues tends to be highly reliable, a single coder reviewed all audiotapes and inserted markers for questions, pauses (greater than three seconds), broken utterances and unintelligible utterances.<sup>v</sup> Then, Miller and Chapman's (1982) computer program, *Systematic Analysis of Language Transcripts* (SALT) was used to mark the occurrence of the nine coding categories. The mechanics of automatically coding interviews with SALT are described in detail elsewhere (Bolton and Bronkhorst 1991).

SALT determined whether each segment had any cues corresponding to each of the nine categories. This information provided an indicator of the occurrence of the processing problem (i.e., yes/no). It was also used to count the frequency of each category code for each answer by each respondent. This frequency was divided by the total number of segments spoken by the respondent to answer the question, yielding a measure of the intensity of the processing problem. This division creates a statistic that lies between zero and one. For example, a respondent might pause three times during 10 segments uttered in response to a particular question, yielding an occurrence measure of 1.0 and an

intensity measure of 0.30. Note that both the occurrence and intensity measures adjust for the fact that some respondents are more verbose than others.

#### Data

An observation described the respondent's identification number, the question answered, and the version of the question: CV, RV, or NV), the occurrence and intensity measures associated with each of the nine AC categories, the number of segments spoken by the respondent, and the OM codes (for the NV only). Table 2 shows the average values of each of the eighteen AC measures for respondents' answers to eight questions. Since skip patterns dictated that some customers were not asked certain questions, these average values are calculated from (58 x 8 - 48 =) 416 observations. The average customer spoke nine segments in response to a question from the interviewer, suggesting that customers had little difficult "thinking aloud." The occurrence measure for the "Questions" category shows that 19% of the respondents used an interrogative tone of voice (i.e., asked the interviewer a question). This percentage seems remarkably high -- indicating that respondents may be experiencing comprehension difficulties with certain questions. The intensity measure for the "Questions" category (3%) is relatively low in comparison with other categories -- indicating that respondents usually asked one question, rather than multiple questions. The lower half of Table 2 shows the percentage of respondents assigned each of the OM codes for the eight questions from the NV. For example, 35% of respondents' answers were coded as indicating response difficulties. This value is relatively high, suggesting that many respondents gave inadequate answers.

-----  
Table 2 here  
-----

#### PRELIMINARY ANALYSES

This section describes preliminary analyses of the automatically coded pretest data and explains how to interpret the results. An exploratory factor analysis is conducted to investigate the underlying dimensions of the coding categories. The resultant factor scores are used in an analysis of variance to provide diagnostic information about superior or inferior questions.

#### Factor Analysis

A factor analysis was conducted to determine whether respondents' coded verbalizations reflected a smaller subset of underlying cognitive difficulties.

The nine "intensity" variables were subjected to a principal components analysis with a varimax rotation. This analysis identified four independent factors with eigenvalues greater than one that explained 56% of the variance in the original nine categories.<sup>vi</sup> (See Table 3.) Consistent with the pilot studies (Bolton 1991), the questions and repeat categories load heavily on factor one. This factor was labelled COMPREHENSION difficulties. The new coding categories, can't say and don't know, load heavily on the second factor.

This factor was labelled RESPONSE difficulties. Consistent with the pilot studies, the confidence category loads heavily on a separate factor. A new category of non-verbal cues, unintelligible utterances, is also highly associated with this factor. Hence, the third factor was labelled JUDGMENT difficulties. The pause and forget categories load on the fourth factor, labelled RETRIEVAL difficulties. This result is similar to the pilot study, except that the forget loading is low here. Broken utterances are also highly associated with this factor. This result is consistent with the notion that

broken utterances can indicate the sudden retrieval of information from memory.

-----  
Table 3  
-----

In general, the factor analysis results are remarkably "clean." Pauses, unintelligible utterances and broken utterances might be expected to load highly on more than one factor because they could reflect more than one information processing problem, but they do not. The results for COMPREHENSION, RETRIEVAL and JUDGMENT factors are similar to the first pilot study results, despite the fact that this sample (business customers) is different from the pilot study sample (residential customers). This finding suggests that the coding categories are relatively generic and might be used in other studies.

-----  
Table 4 here  
-----

Interpretation

The next stage of the analysis investigated whether a content analysis based on the coding scheme could be used to detect defective questions and provide diagnostic information. It focused on four questions (shown in Table 4) in which the phrasing was altered between the three versions. The wording changes could be dramatic (Q3) or subtle (Q1). Since a pretest entails responses to multiple questions of varying difficulty, managers and researchers will be interested in comparing the results for a particular question with the results for some "base case" (such as all other questions in the survey). They can inspect the magnitudes of the AC occurrence and intensity measures, or they can factor analyze the intensity measures and compare the average factor scores.

-----  
Table 5 here  
-----

Customers' responses to different versions of the four questions can be described by factor scores generated to represent the four types of cognitive difficulty: COMPREHENSION, RETRIEVAL, JUDGMENT and RESPONSE. The factor scores for each cognitive difficulty were standardized to have a mean of zero and a standard deviation of one. The average factor scores for each version of each question are shown in Table 5. Higher values of the AC occurrence and intensity measures or the factor scores for a particular question indicate that respondents generated more verbal and non-verbal cues indicating difficulties with the question. As a rough heuristic, a value is "high" if it is two standard errors above the overall mean (i.e., above the mean value across all questions). Recall that the mean factor score across all questions is zero. Since the standard error of the mean is about 0.23 in this study,<sup>vii</sup> an average factor score of (2 x 0.23 =) 0.46 or higher indicates a potentially defective question. For example, the average COMPREHENSION factor score for the CV of Q3 is 0.9135 which is more than two standard errors above the mean of zero. Hence, these data indicate that respondents experienced relatively more COMPREHENSION difficulties for the CV of Q3 than for other questions. It is also possible to compare the average factor score for a particular question to another factor score for the same question. For example, since the average COMPREHENSION score for the CV of Q3 is 0.9135 and the average RESPONSE score is -0.3101, respondents seem to have experienced more COMPREHENSION

difficulties than RESPONSE difficulties for this question.

#### Analysis of Variance

Since this study tests alternative versions of the same question, it is also possible to compare the average factor score for a particular question to the average factor score for another version of the same question. Relatively large differences (greater than two standard errors) between factor means indicate that respondents experienced different levels of cognitive difficulties. For example, since the average COMPREHENSION score for the CV of Q3 is 0.9135 and the average COMPREHENSION score for the NV of Q3 is -0.2875, respondents are clearly experiencing more COMPREHENSION difficulties with the CV than the NV. An analysis of variance (ANOVA) can also be used to test the hypothesis that the factor means are equal for different versions of the same question.<sup>viii</sup> One-way ANOVAs were conducted for each set of factor score means for each question. The null hypothesis that the factor means are equal across versions (CV/RV/NV) was rejected by an F test for 69% (11/16) of the ANOVAs ( $p < 0.15$ ) -- implying that there are differences in the average factor scores across different versions of the four questions. This finding is rather surprising because: (1) the versions of each question were designed to produce similar responses, albeit with different phraseology; (2) the small business customer population was expected to be reasonably homogeneous; and, (3) the sample sizes are relatively small so that only large differences can be detected.

The statistics reported in Table 5 are interpreted in the following way.

For example, an F test of the null hypothesis that the CV and NV of Q3 (data communications) have equal factor means is rejected ( $p < 0.05$ ) for two of the four factors: COMPREHENSION, and RETRIEVAL. Inspecting the means for these two factors, it can be seen that respondents have higher average COMPREHENSION and RETRIEVAL scores with the CV than with the NV. In general, a lower factor score implies fewer difficulties associated with associated with that particular macroprocess. Hence, the NV of Q3 appears to be the "better" version in the sense that it yields significantly fewer cues indicating comprehension and retrieval difficulties.

When there are more than two versions of a particular question, Scheffe's multiple comparison test was used to determine which versions had significantly different factor means. The Scheffe test requires larger differences between means for significance than most other tests, and it is applicable to groups for which sample sizes are unequal. Since the sample sizes in this study are rather small, all Scheffe tests used a significance level of 0.15. The right hand column of Table 5 shows which versions of each question had significantly different factor score means according to the Scheffe test. For example, respondents have higher COMPREHENSION and RESPONSE scores with the RV of Q4 (installation) than the NV1. They also have higher COMPREHENSION and RETRIEVAL scores with the RV of Q4 than NV2. Hence, the test results suggest that NV1 and NV2 are "better" versions than the RV.

#### RESULTS

This section presents the results from automatic coding and observational monitoring. It describes how management selected the "best" version of each question using the AC results. Then, it compares the OM and AC results for the NV of the questionnaire. Lastly, the improvement in the survey is assessed.

#### Automatic Coding Results

Question One. Q1 (overall quality) was a key survey item intended to elicit a global evaluation of the company's service quality. In exploring

potential revisions to this question, the company wanted a simple question that implied a broad definition of service. Respondents' answers to the NV were characterized by higher JUDGMENT scores than their answers to the CV. The two versions of this question are identical except that the NV uses the broader term "telecommunications services" rather than "telecommunications." The broader term seems to create judgment difficulties for small business customers (unlike large business customers) that have not used many telecommunication services (e.g., data communications, maintenance).

Respondents' answers to the NV were characterized by higher RETRIEVAL scores and lower RESPONSE scores than their answers to the RV. Since the NV contains a reference period ("over the past three months") as well as the term "telecommunications services," it appears to shift respondents' efforts towards retrieval processes ("can't remember") and away from response processes ("don't know"). The reference period apparently encouraged customers to make memory based, rather than stimulus based, judgments for Q1. Management preferred to obtain a memory based judgment, so the current wording of the question that included the simple term "telecommunications" and a three-month reference period was retained.

Question Two. Local telephone service is a monopoly in which customers are unable to choose among suppliers. Management was particularly interested in alternative ways of asking about the hypothetical situation in which customers could choose among suppliers. Respondents' answers to the CV and the RV of Q2 (recommend) are characterized by lower JUDGMENT scores than their answers to the NV. This result is consistent with the general notion that short, simple questions are preferable. Respondents' answers to the RV have higher RESPONSE scores than their answers to the NV. This result seemed to be due to problematical wording in both the CV and the RV because they ask customers "would you recommend" rather than asking about choice. Management decided that none of the three versions of this question were acceptable.

Question Three. Managers had considered Q3 (data communications) to be very straightforward. However, after the first pretest, they realized that business customers were experiencing considerable difficulty with this question. The second pretest investigated a new version of the question that defined data communications using terms that customers had employed during the first pretest. Table 5 shows that respondents' answers to the NV had lower COMPREHENSION and RETRIEVAL scores than the CV. Respondents apparently experienced much less difficulty with the NV because it provided a richer set of retrieval cues. Management concluded that the NV was a superior question.

Question Four. Q4 was intended to determine whether the customer was sufficiently knowledgeable about installation services that he/she could answer a sequence of more detailed questions. Management was dissatisfied with the CV and experimented with alternative wordings that asked about knowledgeability without implying personal contact with the installer. The null hypotheses that the COMPREHENSION, RETRIEVAL, JUDGMENT and RESPONSE means were equal across versions were rejected ( $p < 0.15$ ). NV2's phrasing (i.e., "responsible for managing") seems dramatically different from the phrasing of the other three versions (i.e., "familiar with"). However, the Scheffe test results indicated that the RV's phrasing (i.e., "familiar with the more general aspects") is associated with cognitive difficulties. Respondents' answers to the RV had significantly higher COMPREHENSION and RESPONSE scores than their answers to the NV1. Respondents' answers to the RV had significantly higher COMPREHENSION scores and significantly lower RETRIEVAL scores than their answers to the NV2.

(Perhaps comprehension problems stop customers from attempting to retrieve information.) These results suggested that the RV and NV2 versions of the question were unacceptable. Although the average factor scores for the CV and the NV1 were not significantly different, the respondents' answers to the NV1 have lower COMPREHENSION, JUDGMENT and RESPONSE scores. Hence, this version of the question was selected by management.

#### Observational Monitoring Results

Table 6 displays the OM results for the NV. The right column shows the percentage of interviews that were coded with each type of difficulty for each of the four questions. For example, 20% of respondents experienced comprehension difficulties on the NV of Q1 (overall quality). For comparison purposes, the middle three columns of show describe the automatic coding (AC) categories (e.g., forget, pause, and broken), the percentage of respondents that were automatically assigned each code for the NV (e.g., 20% of the respondents were automatically assigned the pause code for Q1) and the associated factor score (e.g., 0.31 for the retrieval factor).

The percentage of customers experiencing difficulty with specific questions can be compared to identify superior or inferior questions. As a rough heuristic, managers considered OM percentages greater than 15% to be "high." Small percentages (0-5%) of respondents experienced comprehension difficulties with the NV of questions three and four, whereas larger percentages (14-20%) of respondents experienced comprehension difficulties with the NV of questions one and two. Apparently, questions about specific aspects of service are easier to comprehend than questions designed to elicit global evaluations. Although a high percentage of respondents (34-43%) gave inadequate answers to the first three questions, these results may be an artifact of the concurrent protocol technique. The elicitation of concurrent protocols may be encouraging open-ended responses.

-----  
Table 6 here  
-----

There is some agreement between the AC and OM results.<sup>ix</sup> For example, AC results show that Q2 (recommend) has high levels of judgment difficulties (i.e., unintelligible utterances and cues that respondents lacked confidence); OM results also indicate high levels of judgment difficulties. However, overall consistency between AC and OM results seems to be low. For example, although AC counted a significant percentages of pauses (20%) and broken utterances (23%) in respondents' verbalizations for Q1 (overall quality), OM did not identify any retrieval difficulties (0%). As shown in Table 7, the correlation between the AC factor scores and the OM codes for comprehension difficulties is 0.48 ( $p < 0.01$ ) across the four questions. However, other statistically significant correlations seem to arise from the coders' tendency to record any respondent difficulty as a comprehension difficulty.

-----  
Table 7 here  
-----

Correlations between the AC factor scores and the OM codes can be low even when the raw percentages are similar. For example, consider the OM results for the installation question. NV2 seems to slightly dominate NV1 because the latter has a higher percentage of judgment difficulties. This result is consistent with the raw percentages obtained from AC. However, the JUDGMENT factor scores give the opposite result, indicating that NV1 is

superior to NV2 (although the difference is not statistically significant at  $p > 0.15$ ). This apparent contradiction between the raw percentages and the JUDGMENT factor scores (both produced by AC) is due to the way that the factor scores are calculated. The factor scores are weighted combinations of all nine AC categories. Hence, AC results reflect more subtle differences in customers' responses.

#### Assessing Survey Improvement

Improvement in the survey can be quantified as follows. GTE had originally intended to replace the current version (CV) of the survey with the revised version (RV). Customers' cognitive difficulties with the revised version of the survey can be measured by the factor scores. The left side of Table 8 shows the average factor scores for the revised versions of the following three questions: overall quality, data communications and installation. After the two pretests, GTE decided to use the CV of Q1, the NV of Q3 (data communications) and the NV1 of Q4 (installation). All versions of Q2 were rejected, so this question is excluded from the analysis. Customers' cognitive difficulties with the "final" versions of the questions are measured by the factor scores shown on the right side of Table 8. The average COMPREHENSION, JUDGMENT and RESPONSE scores are almost one standard deviation lower in the final version of the survey. (Recall that the standard deviation of each factor score is one.) The RETRIEVAL score is slightly higher because GTE elected to retain the current version of Q1 (overall quality) which includes a reference period. (Recall that managers wanted to encourage customers to make memory based judgments for Q1.) Overall, the results show that the final version produces fewer cognitive difficulties for respondent.

-----  
 Table 8 here  
 -----

Another way of assessing survey improvement is to examine the accuracy of customers' responses. Managers could verify the accuracy of customers' answers to Q3 by examining the internal consistency of their answers or by checking company records to determine whether a customer has subscribed to data communications services from GTE. The following transcript of a customer's answer to the CV of Q3 illustrates inconsistency within a response. (Each new line denotes a new segment; ">" denotes a broken utterance; and "(uh)" denotes an unintelligible utterance.)

No.  
 You mean data services.  
 we do have data services but not from GTE.  
 GTE provides the line.  
 the telephone line.  
 We use the line for data services, that's it.  
 we don't.>  
 it's not.>  
 GTE does not provide us with any (uh) like modems.  
 computer modems or anything like that.  
 They provide us with a line.  
 we use that line for a computer modem to communicate.  
 That's it.

The customer reported that her company uses a computer modem to send data to another location (implying that her firm uses GTE's data communication

services/lines) and yet she answered Q3 in the negative. Hence, she made a response error. Several customers gave internally inconsistent answers that indicated a response error when responding to the CV of Q3, whereas no customers did so in responding to the NV of Q3. Hence, managers concluded that the NV produced more accurate responses -- confirming the AC results. Since the results from checking the internal consistency of answers to Q3 were very clear-cut, GTE did not check business customers' records. Other questions in the survey elicit opinions, so they are not objectively verifiable from checking internal consistency or company records.

#### DISCUSSION

*Is an automatic coding scheme based on verbalizations elicited during pretests useful in identifying the information processing difficulties that arise as respondents form answers to survey questions?*

The results of the factor analysis demonstrate that the coding scheme was useful in identifying cognitive difficulties that arise as respondents form answers to survey questions. The factor analysis identified four independent dimensions of cognitive difficulties that were labelled COMPREHENSION, RETRIEVAL, JUDGMENT and RESPONSE. It isn't possible to unambiguously link these four factors with the cognitive processes underlying answers to survey questions that have been proposed by Tourangeau and others. However, these four factors have good face validity and, since eight of the nine coding categories loaded heavily on one of the four factors, there is evidence of convergent and discriminant validity. Furthermore, these macroprocesses are relatively stable across different respondents (i.e., business versus residential customers) and content domains (i.e., different surveys with different questions). Although further work is necessary to develop a richer AC coding scheme (e.g., to better discriminate among recall and judgment difficulties), this information is much more detailed than the information provided by conventional pretesting methods.

*Is a coding scheme that traces information processing difficulties useful in detecting defective questions?*

The ANOVA results demonstrate that the coding scheme can be used to distinguish among different versions of the same question. It is possible to identify and improve defective items by replacing questions with which respondents experienced higher levels of cognitive difficulty with less difficult versions. It is also possible to make general statements about the cognitive effort associated with different questions. For example, respondents seem to have higher RESPONSE scores with generally worded questions and higher JUDGMENT scores with specific questions. This finding suggests that more specific questions may cause respondents to be uncertain about the accuracy of the information that they have recalled or about how to weigh the information that they have recalled or both.

*Can the information provided by a content analysis of respondents' verbalizations be used to identify superior or inferior questions?*

This study demonstrates that superior or inferior questions can be identified from raw AC codes or from summary factor scores. For example, the ANOVA results for Q3 clearly show that the NV is superior to the CV. The magnitude of the reduction in cognitive difficulty can be quantified as shown in Table 8. Furthermore, evidence external to the AC results also indicated that the NV of Q3 leads to reduced response errors. Customers did not give internally inconsistent answers indicating a response error when responding to the NV of Q3.

Subsequent to the pretests, GTE obtained objective, independent evidence that the AC pretest results had yielded improved surveys that gathered better information. GTE compared the responses of two national probability samples of small business customers. The CV was administered to one sample and the newly revised survey was administered to the other. Interviewers reported significantly fewer problems administering the new survey. More importantly, the responses were significantly different for some questions -- indicating decreases in response error. In a dramatic example, 20% of customers responded affirmatively to the CV of Q3, whereas 86% of customers responded affirmatively to the NV of Q3. (Interestingly, the proportion of "don't know" responses was constant across the two versions.) Apparently, most customers answered "no" to the CV because they did not recognize the term "data communication services" as referring to GTE's provision of data lines for sending information via modems, credit card readers, and fax machines.

How do the results produced by this method compare with the results of observational monitoring?

The results indicate that OM and AC work equally well at detecting question defects that lead to comprehension difficulties. OM seems better at detecting response difficulties -- probably because human coders are better able to determine whether a respondent has adequately answered a fixed format question. Key word lists cannot adequately identify all situations in which a respondent ultimately gives an inadequate answer. For example, if a respondent does not give an alternative from the fixed response categories, he/she will not necessarily indicate this fact by using a word string such as "can't say."

In general, OM seems to perform particularly well when the pretesting task requires expert interpretation. For example, GTE's OM coding scheme includes a separate code to identify questions that respondents did not interpret as GTE intended and questions for which the interviewer (as opposed to the respondent) experienced difficulties. In one question, some respondents confused a yellow pages salesperson with a telephone operations salesperson -- but this confusion was not evident to a naive listener. This type of unanticipated problem can't be picked up by the AC codes.

AC detects cognitive difficulties -- particularly retrieval and judgment difficulties -- that OM does not detect. For example, AC results show that almost one-third of respondents experienced retrieval and judgment difficulties in answering Q1 (overall quality) -- but the OM percentages are substantially lower. OM's failure to detect retrieval and judgment difficulties is partially due to the fact that coders did not seem to recognize the non-verbal cues that indicate these types of cognitive difficulties. However, AC's use of non-verbal cues does not entirely explain this failure. For example, the AC occurrence measure indicates that almost 60% of respondents expressed a lack of confidence in their answers to Q2, but the OM occurrence measure indicates that only 17% of respondents experienced such difficulties.

AC relies on a fairly complex coding scheme, that includes a variety of subtle verbal and non-verbal cues, so it produces a rich set of diagnostic information. For example, AC distinguished between verbal cues that the respondent lacked confidence about how to answer the question and non-verbal cues (unintelligible utterances) that suggested non-verbalized processing. In contrast, OM coding schemes typically do not operate at this level of detail. A more complex subjective encoding scheme could provide a richer description of respondents' cognitive strategies as they form answers to the survey questions.

For example, it could describe how respondents make comparisons in forming judgments or how they use retrieval cues. However, existing pretest methods have not examined these questions.

Do the incremental benefits of the diagnostic information provided by AC outweigh the incremental costs of automatically coding the data? Bolton and Bronkhorst (1991, p. 286-7) provide some comparative information for automatically coding fifteen one-hour interviews. If a researcher is developing a new coding scheme, both OM and AC may require as much as 40-50 hours of development work. After the pretest interviews have been conducted and transcribed, experts typically spend about 15 hours reviewing the interview transcripts and analyzing OM codes. In contrast, it takes about 20 hours to segment and precode the transcripts, plus a few hours to tabulate the AC results. In other words, OM requires about 15 hours work by an expert whereas AC requires 22-25 hours work -- mostly by someone with a much lower level of expertise. As the number or length of the interviews increases, AC becomes more attractive because the fixed costs (including the time to develop and iteratively test new codes for special situations) are spread over more interviews. Also, as the investigator gains experience, automatic coding becomes increasingly time and cost-effective. Hence, AC would be particularly attractive to organizations (e.g., market research firms, large corporations and government) that pretest questionnaires frequently and desire standardized procedures.

#### CONCLUDING REMARKS

Prior research suggests that observational monitoring is a simple, low-cost, flexible and effective system for pretesting. However, this study shows that observational monitoring does not uncover all question defects or adequately diagnose all the question defects that it identifies. Automatic coding uncovers question defects that lead to retrieval and judgment difficulties that observational monitoring does not detect. It also provides more detailed diagnostic information, such as distinguishing between respondents' statements of lack of confidence and subtle cues indicating non-verbalized processing.

These findings are consistent with Oksenberg, Cannell and Kalton's (1991) experience with observational monitoring. They concluded that observational monitoring "does not always identify the sources of the problems it uncovers" (p. 362). They have used special probes and interviewer ratings to uncover additional problems and they debrief coders and interviewers to obtain additional insights concerning the source of problems. However, they report that special probes work well for comprehension difficulties but not for other difficulties.

Automatic coding seems particularly suited to identifying comprehension, retrieval and judgment difficulties that arise from question phrasing defects, whereas observational monitoring seems better suited to situations requiring expert judgment of response difficulties arising from other questionnaire design factors, such as question order. For example, GTE uses observational monitoring techniques to uncover respondent difficulties due to question order and to investigate instances where respondents' interpretations of a question does not match managerial intent. Hence, these findings suggest that a content analysis that relies on computer-assisted coding can be a potentially important complement to conventional analyses. A researcher may decide to use both observational monitoring and automatic coding to evaluate different aspects of a survey.

Researchers can use the proposed method to quantify respondents' cognitive difficulties with specific survey items and to relate these difficulties to specific question characteristics. This information can be used to identify question defects and to make revisions -- with measurable improvements in respondents' ability to respond to the questions. Furthermore, in this study, subsequent survey results provided independent evidence that the revised questionnaire yielded "better" information for decision-makers to use in policy decisions. Hence, the proposed questionnaire pretesting methodology seems to be a useful step towards measuring and controlling non-sampling error.

As a general rule of thumb, survey researchers are advised to allocate five to ten percent of their budgets to survey pretesting. (This rule is typically honored in the breach!) The resultant questionnaire revisions may entail structural changes in the questionnaire (e.g., question order), as well as changes in phrasing (e.g., request for cooperation). In addition to decreasing response errors, these changes can make the interviews more pleasant for respondents and make the survey instrument easier to administer -- leading to higher cooperation rates, more complete information, shorter interview times -- and substantial cost savings. The magnitude of the cost savings due to pretesting depend on the nature of the survey and the size of the sample. However, the following example provides some "ballpark figures." These calculations are based on GTE's experience with a survey of 12,800 respondents that cost about \$282,000. The market research supplier reported an eight percent reduction in costs due to questionnaire revisions: \$15,100 (5%) due to changes in cooperation rates, \$2,100 (1%) due to specific changes in questions, and \$4,800 (2%) due to a reduction in average interview time of three minutes. Since pretesting costs run about three percent of the total survey costs,<sup>x</sup> the estimated net cost saving was about \$13,540 (5%).

Further research is necessary to more fully explore the situations in which it is beneficial to apply content analysis to pretest interviews. Theoretical work is necessary to understand respondents' cognitive processes and responses to survey questions. Empirical work is required to develop and refine appropriate coding schemes for different interview topics and characteristics. For example, it would be desirable to develop additional codes to identify judgment difficulties. Finally, it would be desirable to conduct experiments to evaluate this questionnaire pretesting methodology by applying it to survey questions that have been designed to elicit certain types of problems.

FOOTNOTES

Table 1  
Coding Scheme

<u>Repeat</u>	<u>Forget</u>
repeat	don't:remember
interpret	do:not:remember
define	forget
explain	can't:think
comment	cannot:think
how&rate	I'm:trying:to:think
do:you:mean	can't:recall
ask:that	cannot:recall
asking	no:memory
I&misunderstand	can't:remember
I&misunderstood	cannot:remember
I:don't:think:I:got:you	no:recollection
I:thought:you:said	no:remembrance
say:that:again	doesn't:call:to:mind
are&talking:about	does:not:call:to:mind
in:terms:of	won't:call:to:mind
so:it:says	will:not:call:to:mind
in:other:words	doesn't:jog&memory
about:the:question	does:not:jog&memory
that:word	doesn't:remind:me
don't:understand	does:not:remind:me
listen:to&again	don't:recollect
are:you:talking	do:not:recollect
hear:that:again	
is:that:what:you're:asking	
problem&question	
what&looking:for	
what:was&again	
what&mean	
I:take:it:that	
I'm:sorry	
pardon:me	
wondering	
confused	
you:mean	
ask:that:question	
say&one:more:time	
one:more:time&the:question	
I'm:lost	
I:lost:you	
listen:to:the:question	
read:the:question	
state:the:question	
hear:the:question	
listen:to:the:question	
ask:the:question	
read:that:question	
state:that:question	
hear:that:question	
	<u>Confidence</u>
	probably

	approximately maybe perhaps I:guess kind:of unless somewhere:in:there I:reckon not:certain I&imagine depends mostly sort:of not:sure whatever or:something
--	---

Table 1 (cont'd)

Coding Scheme

<u>Can't Say</u>	<u>Don't Know</u>
I:can't:say	I:don't:know
I:can't:tell	I:wouldn't:know
I:can't:rate	not:know
I:can't:evaluate	I:don't:think:I:know
I:can't:judge	I:have:no:idea
tough&rate	don't:have&any:idea
not:easy&rate	no:experience
difficult&rate	never:experienced
hard&rate	not:experienced
tough&evaluate	any:experience
not:easy&evaluate	never:experience
difficult&evaluate	not:experience
hard&evaluate	haven't:experienced
tough&to:say	not:familiar:with
not:easy&to:say	no:need:for&service
difficult&to:say	no:need:for&option
hard&to:say	never:done
tough&judge	never:used
not:easy&judge	never:use
difficult&judge	don't:use
hard&judge	haven't:used
tough&to:tell	not:familiar
not:easy&to:tell	
difficult&to:tell	
hard&to:tell	

Table 2

Means\*

Coding Category	Intensity (%)	Occurrence (%)
Automatic Coding		
Questions	3.31	19.47
Repeat	2.27	16.35
Pause	1.38	11.30
Forget	0.22	2.40
Broken	1.61	12.50
Unintelligible	12.82	50.48
Confidence	5.13	32.69
Don't Know	2.10	15.63
Can't Say	0.34	3.13
Observational Monitoring		
Comprehension	11.54	na
Retrieval	1.92	na
Judgment	11.54	na
Response	34.62	na

\* Lower scores imply fewer cues indicating cognitive difficulties. On average each respondent uttered nine segments in response to a question. Sample size is 416 observations for automatic coding and 104 observations for observational monitoring.

Table 3

Rotated Factor Loadings\*

	<u>FACTOR 1</u>	<u>FACTOR 2</u>	<u>FACTOR 3</u>	<u>FACTOR 4</u>
	(Comprehension)	(Response)	(Judgment)	(Retrieval)
Question	<b>.79145</b>	-.06748	-.09436	.21099
Repeat	<b>.84066</b>	.06873	.07090	-.07529
Pause	.15820	-.21737	.23077	<b>.63399</b>
Forget	.01068	.02804	-.09030	<b>.36401</b>
Broken	-.02313	.19954	-.01863	<b>.63907</b>
Unintell.	-.22708	.00641	<b>.76396</b>	.13931
Confidence	.28146	.12146	<b>.69347</b>	-.25739
Can't Say	-.01013	<b>.80387</b>	-.10354	-.00348
Don't Know	.02394	<b>.67899</b>	.21685	.09792
Eigenvalues	1.52	1.31	1.10	1.07
Variance Explained	0.17	0.15	0.12	0.12

\* Sample size is 416 observations.

Table Four

Selected Questions\*

Current (CV)	Revised (RV)	New (NV)
Overall, how would you rate the quality of telecommunications GTE has provided to your company over the past three months? Would you say A for Excellent, B for Good, C for Average, D for Below Average or F for Poor?	Overall, how would you rate the quality of telecommunications GTE has provided to your company?	Overall, how would you rate the quality of telecommunications services that GTE has been providing to your company over the past three months?
Based on your overall experience during the past three months, would you recommend GTE to someone at another business who has a choice between GTE and alternative suppliers? [Open-ended question. Response categories: would/would not/don't have a choice/don't know.]	If someone at another business had a choice between GTE and alternative suppliers, based on your overall experience, would you recommend GTE?	Suppose conditions existed where your company had a choice among GTE and alternative local telephone companies. Based on your overall experience with GTE, would you choose them to be your local telephone company?
Does your company have any data communications services from GTE? [Open-ended question. Response categories: yes/no/don't know.]	Does your company have any data communications services from GTE? [Same as current version.]	Do you send information between different locations using any of the following data equipment: computer, fax, modem or credit card reader?
Are you generally familiar with these installation services? [Open-ended question. Response categories: yes/no/don't know.]	Are you familiar with the more general aspects of these installation services?	(1) Are you personally familiar with the installation service provided to your company by GTE? (2) Are you personally responsible for managing installation with GTE?

\* Response categories are the same across all versions of a given question. GTE investigates response category issues using different research techniques than those described in this paper.

Table 5

ANOVA and Scheffe Test Results<sup>a</sup>

1. CV: Overall, how would you rate the quality of telecommunications GTE has provided to your company over the past three months?						
Difficulty	CV	RV	NV		F	Scheffe
Comprehension	-0.0332	-0.0033	-0.1304		0.19	
Retrieval	0.0702	-0.4598	0.3131		2.97**	RV/NV
Judgment	-0.4161	0.2179	0.3324		2.58**	CV/NV
Response	-0.0014	0.1644	-0.1909		2.13*	RV/NV
Sample Size	13	15	30			
2. CV: Based on your overall experience during the past three months, would you recommend GTE to someone at another business who has a choice between GTE and alternative suppliers?						
Difficulty	CV	RV	NV		F	Scheffe
Comprehension	-0.1997	0.0210	-0.0452		0.49	
Retrieval	-0.1956	-0.2646	0.1083		1.32	
Judgment	-0.4765	-0.4522	0.3645		7.27**	CV/NV
Response	0.3020	1.3812	-0.1279		4.10***	RV/NV
Sample Size	13	15	29			
3. CV: Does your company have any data communications services from GTE?						
Difficulty	CV	RV	NV		F	Paired
Comprehension	0.9135	NA	-0.2875		10.68***	CV/NV
Retrieval	0.5980	NA	-0.1087		4.97***	CV/NV
Judgment	-0.1875	NA	0.1092		1.94	
Response	-0.3101	NA	-0.2565		0.12	
Sample Size	28		30			
4. CV: Are you generally familiar with these installation services?						
Difficulty	CV	RV	NV1	NV2	F	Scheffe
Comprehension	0.5008	1.5180	-0.0839	0.4360	3.87***	RV/NV1
Retrieval	-0.5421	-0.6826	-0.2130	0.2804	3.49***	RV/NV2
Judgment	0.7042	0.4954	-0.2841	-0.4599	1.90*	
Response	-0.0278	0.5542	-0.3495	-0.2425	2.63**	RV/NV1
Sample Size	5	12	7	8		

<sup>a</sup> Significance level for the F tests are indicated by asterisks: \*\*\* p < 0.05, \*\* p < 0.10, \* p < 0.15. All Scheffe tests used a significance level of 0.15.

**Table 6**  
**Comparison With Observational Monitoring**

Difficulty	Code	AC Percentage	Factor Score	OM Percentage
<b>1. NV: Overall, how would you rate the quality of telecommunications services that GTE ...</b>				
<b>Comprehension</b>	<b>Questions</b>	<b>23.33</b>	<b>-0.1304</b>	<b>20.00</b>
	<b>Repeat</b>	<b>13.33</b>		
<b>Retrieval</b>	<b>Forget</b>	<b>0.00</b>	<b>0.3131</b>	<b>0.00</b>
	<b>Pause</b>	<b>20.00</b>		
	<b>Broken</b>	<b>23.33</b>		
<b>Judgment</b>	<b>Unintelligible</b>	<b>73.33</b>	<b>0.3324</b>	<b>10.00</b>
	<b>Confidence</b>	<b>36.67</b>		
<b>Response</b>	<b>Can't Say</b>	<b>0.00</b>	<b>-0.1909</b>	<b>36.67</b>
	<b>Don't Know</b>	<b>20.00</b>		
<b>Q2. NV: Suppose conditions existed where your company had a choice among GTE and alternative...</b>				
<b>Comprehension</b>	<b>Questions</b>	<b>27.59</b>	<b>-0.0452</b>	<b>13.79</b>
	<b>Repeat</b>	<b>24.14</b>		
<b>Retrieval</b>	<b>Forget</b>	<b>0.00</b>	<b>0.1083</b>	<b>6.90</b>
	<b>Pause</b>	<b>20.69</b>		
	<b>Broken</b>	<b>20.69</b>		
<b>Judgment</b>	<b>Unintelligible</b>	<b>75.86</b>	<b>0.3645</b>	<b>17.24</b>
	<b>Confidence</b>	<b>58.62</b>		
<b>Response</b>	<b>Can't Say</b>	<b>0.00</b>	<b>-0.1279</b>	<b>34.48</b>
	<b>Don't Know</b>	<b>10.34</b>		
<b>Q3. NV: Do you send information between different locations using any of the following data ...</b>				
<b>Comprehension</b>	<b>Questions</b>	<b>20.00</b>	<b>-0.2875</b>	<b>6.67</b>
	<b>Repeat</b>	<b>13.33</b>		
<b>Retrieval</b>	<b>Forget</b>	<b>0.00</b>	<b>-0.1087</b>	<b>0.00</b>
	<b>Pause</b>	<b>0.00</b>		
	<b>Broken</b>	<b>3.33</b>		
<b>Judgment</b>	<b>Unintelligible</b>	<b>63.33</b>	<b>0.1092</b>	<b>10.00</b>
	<b>Confidence</b>	<b>16.67</b>		
<b>Response</b>	<b>Can't Say</b>	<b>0.00</b>	<b>-0.2565</b>	<b>43.33</b>
	<b>Don't Know</b>	<b>10.00</b>		
<b>4. NV1/2: Are you personally familiar with the installation service ... / responsible for managing ...</b>				
<b>Comprehension</b>	<b>Questions</b>	<b>14.3 / 0.0</b>	<b>NV1: -0.0839</b>	<b>NV1: 0.00</b>
	<b>Repeat</b>	<b>14.3 / 0.0</b>	<b>NV2: -0.4360</b>	<b>NV2: 0.00</b>
<b>Retrieval</b>	<b>Forget</b>	<b>0.0 / 0.0</b>	<b>NV1: -0.2128</b>	<b>NV1: 0.00</b>
	<b>Pause</b>	<b>14.3 / 12.5</b>	<b>NV2: 0.2803</b>	<b>NV2: 0.00</b>
	<b>Broken</b>	<b>0.0 / 25.0</b>		
<b>Judgment</b>	<b>Unintelligible</b>	<b>57.1 / 37.5</b>	<b>NV1: -0.2841</b>	<b>NV1: 14.29</b>
	<b>Confidence</b>	<b>14.3 / 12.5</b>	<b>NV2: -0.4599</b>	<b>NV2: 0.00</b>
<b>Response</b>	<b>Can't Say</b>	<b>0.0 / 0.0</b>	<b>NV1: -0.3495</b>	<b>NV1: 14.29</b>
	<b>Don't Know</b>	<b>0.0 / 0.0</b>	<b>NV2: -0.2425</b>	<b>NV2: 12.50</b>

Table 7

Correlations Among Automatic Codes and Observational Monitoring Codes

Observational Monitoring	Automatic Coding			
	Comprehension	Retrieval	Judgment	Response
Comprehension	0.4783***	0.1978**	0.0316	-0.0594
Retrieval	0.0511	0.1039	0.0869	-0.1463*
Judgment	0.1880**	0.0818	0.1085	0.0262
Response	0.1026	0.0468	0.0970	0.0582

One tailed test: \*\*\*  $p \leq 0.01$  \*\*  $p \leq 0.05$  \*  $p \leq 0.10$

Sample size = 104

Table 8

Reduction in Cognitive Difficulties As Measured by Factor Scores

Difficulty	RV*	Selected Version	Difference	Change
Comprehension	+0.8094	-0.1349	-0.9443	better
Retrieval	-0.1815	-0.0838	+0.0977	worse
Judgment	0.5258	-0.1970	-0.7228	better
Response	0.4085	-0.2025	-0.6110	better

\* Factor means are based on three of the four questions.

References

- Assael, Henry and John Keon (1982), "Nonsampling vs. Sampling Errors in Survey Research," Journal of Marketing, 45 (Spring), 114-123.
- Belson, William A. (1981), The Design and Understanding of Survey Questions, Aldershot, England: Gower Publishing Company.
- 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.
- \_\_\_\_\_ and Tina M. Bronkhorst (1990), "General Services Questionnaire: Pretest Results," GTE Laboratories Technical Memorandum 0253-03-90-420.
- \_\_\_\_\_ and Tina M. Bronkhorst (1991), "Quantitative Analyses of Depth Interviews," Psychology and Marketing, 8 (4), 275-97.
- 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.
- Converse, Jean M. and Stanley Presser (1986), "Survey Questions: Handcrafting the Standardized Questionnaire," Sage University Paper series on Quantitative Applications in the Social Sciences, 07-063, Beverly Hills: Sage Publications.
- Ericsson, K. Anders and Herbert A. Simon (1980), "Verbal Reports as Data," Psychological Review, 87 (May) 215-250.
- \_\_\_\_\_ and Herbert A. Simon (1984), Protocol Analysis: Verbal Reports as Data, Cambridge, MA: MIT Press.

- Hippler, Hans-J., Norbert Schwarz and Seymour Sudman (1987), Social Information Processing and Survey Methodology, New York: Springer-Verlag.
- 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.
- Jadine, Thomas B., Miron L. Straf, Judith M. Tanur, Roger Tourangeau (1984), Cognitive Aspects of Survey Methodology, Washington: National Academy Press.
- 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.
- Kuncel, Ruth Boutin (1981), "Reducing Diversity in Subject Interpretation of Items," in New Directions for Methodology of Social and Behavioral Science (9): Problems with Language Imprecision, D. Fiske (ed.), San Francisco: Jossey-Bass.
- Lessler, Judith T. and Monroe G. Sirken (1985), "Laboratory Based Research on the Cognitive Aspects of Survey Methodology: The Goals and Methods of the National Center for Health Statistics Study," Health and Society, 63 (3), 565-581.
- 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.
- Perreault, Jr., William D. and Laurence E. Leigh (1989), "Reliability of Nominal Data Based on Qualitative Judgments," Journal of Marketing Research, 26 (May), 135-148.
- Presser, Stanley (1989), "Pretesting: A Neglected Aspect of Survey Research," in Health Survey Research Methods, Floyd J. Fowler, Jr. (ed.), Department of Health and Human Services Publication No. 89-3447.
- Rogers, T. B. (1973), "Toward a Definition of the Difficulty of a Personality Item," Psychological Reports, 33, 159-166.
- Royston, Patricia (1987), "Application of Cognitive Research Methods to Questionnaire Design, Paper Presented at the Society for Epidemiological Research Twentieth Annual Meeting.
- \_\_\_\_\_, 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.
- Sirken, M. G., D. J. Mingay, P. N. Royston, D. H. Bercini and J. B. Jobe (1988), Interdisciplinary Research in Cognition and Survey Measurement, " Practical Aspects of Memory: Current Research and Issues, Vol. 1, M. M. Gruneberg, P. E. Morris and R. N. Sykes, eds., Chichester, England: John Wiley & Sons.
- Tourangeau, Roger (1984), "Cognitive Science and Survey Methods," in T. Jabine, M. Straf, J. Tanur and R. Tourangeau (Eds.), Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, Washington, DC, pp. 73-100.
- \_\_\_\_\_ (1987), "Attitude Measurement: A Cognitive Perspective," in H. Hippler, N. Schwarz, and S. Sudman (Eds.), Social Information Processing and Survey Methodology, New York: Springer-Verlag, 149-162.
- \_\_\_\_\_ and Kenneth A. Rasinski (1988), "Cognitive Processes Underlying Context Effects in Attitude Measurement," Psychological Bulletin, 103 (3), 299-314.
- Turner, C. B. and D. W. Fiske (1968), "Item Quality and Appropriateness of Response Processes," Educational and Psychological Measurement, 28, 297-315.
- Weber, Robert Philip (1985), Basic Content Analysis, Sage University Series on Quantitative Applications in the Social Sciences, 07-049, Beverly Hills and London: Sage Publications.
- 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 [in press].

FOOTNOTES

i.. Despite these precautions, the elicitation of concurrent verbal protocols may create a situation that is significantly different from "real world survey conditions" so that respondents' primary processes are different. Whether or not these differences are important depends on which processes are being studied. For example, a developmental pretest might use automatic encoding to focus on the comprehension process whereas a polishing pretest or a field pretest might employ subjective encoding to focus on evaluation and response processes.

ii.. "Real time" coding may not be particularly constraining. Oksenberg, Cannell and Kalton (1991, p. 362) report that "when tape recordings are used [for observational monitoring], it is rarely necessary to stop the tape in performing the coding."

iii.. Backchannelling is the occurrence of a speaking turn of one word (e.g., "uh-uh") that does not follow a question. This activity is the equivalent of a nod or other indication that the interviewer "hears" the respondent.

iv.. Reliability checks were made during the one-day training/practice session. Managers agreed about 90% of the time. In addition, there was a reliability check during the pretest. Both managers coded 45 questions/items. The percentage of the items for which the two judges agreed was 82% for the comprehension category, 100% for the retrieval category, 91% for the judgment category and response 96% for the response category. Perreault and Leigh's (1989) measure of reliability for judgment coded data (which adjusts for expected agreement) was calculated to be 0.80 for the comprehension category, 1.00 for the retrieval category, 0.91 for the judgment category and 0.91 for the response category. Perreault and Leigh (1989) suggest that, if the reliability of the coding process is high (greater than 0.90), it may be reasonable to complete the coding process with only one judge per response. Since reliability was high and managerial time was a scarce resource, only one manager was assigned to code each pretest interview.

v.. Perreault and Leigh's (1989) measure of reliability for judgment coded data was calculated for a sample of pretest interview data. 250 utterances were coded by two judges to identify non-verbal cues. The reliability measure was 0.99 for pauses, 0.94 for broken utterances and 0.86 for unintelligible utterances. In a personal communication, Dr. Judy Hall (Editor of the Journal of Non-Verbal Behavior) noted that reliability tends to be very high for these particular types of non-verbal cues. Since reliability is typically high and it was shown to be high in our study, it seemed reasonable to use a single coder.

vi.. The factor analysis results are substantially the same when the raw scores are normalized within a respondent across questions (Dillon, Frederick, Tangpanichdee 1985).

To investigate the stability of the factors, a random sample of 50% of the observations was drawn and subjected to an identical analysis. The same factors with approximately the same loadings were found.

vii.. The standard errors of the factor means are not reported in this paper due to space limitations. They are different for each factor, version and question combination, ranging from 0.04 to 0.83. The average standard error across all factors and questions is 0.23. This value is used for purposes of discussion in this section of the paper.

viii.. In general, residuals analysis indicated that the standard assumptions of normality and homoscedasticity were not violated. Since there was some evidence of heteroscedasticity for a few equations, the factors scores for these equations were subjected to a variance-stabilizing transformation and the regression analyses re-run. The results of the various statistical tests did not change. Hence, this paper reports the results using untransformed factor scores.

ix.. The consistency between OM and AC codes was cursorily investigated in one of the pilot studies (Bolton and Bronkhorst 1991). The correlations among three AC codes and

two OM codes were calculated and they showed some agreement. In particular, the AC questions category and the OM comprehension category were correlated at 0.72 ( $p < 0.001$ ).

x.. These costs are primarily data collection costs -- that is, the costs of recruiting, interviewing and compensating respondents. Automatic coding is not particularly laborious, time-consuming or costly. Subjective coding is equally (if not more) costly - - because of the time required to hire and/or train expert coders, plus encode the interviews. Furthermore, pretesting is much less costly and time-consuming than field experiments that measure the actual magnitude of non-sampling error.