RESEARCH REPORT SERIES
(Survey Methodology #2005-02)

Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results?

Jennifer Rothgeb, Gordon Willis[1], Barbara Forsyth[2]

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

[1.] National Cancer Institute

[2.] Westat, Inc.

Report Issued: 03-21-05

Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results?[1]

Jennifer Rothgeb, U.S. Census Bureau

Gordon Willis, National Cancer Institute

Barbara Forsyth, Westat, Inc.

Annual Conference of American Association for Public Opinion Research

Montreal, May 2001

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

I. Introduction

During the past 15 years, in an effort to improve survey data quality, researchers and survey practitioners have significantly increased their use of an evolving set of questionnaire pretesting methods, including review by experts, cognitive interviewing, behavior coding, and the use of respondent debriefing. Several researchers have addressed issues related to questionnaire evaluation, and have attempted to determine the potential strengths and weaknesses of each (Campanelli, 1997; DeMaio, Mathiowetz, Rothgeb, Beach, and Durant,1993; Oksenberg Cannell, and Kalton, 1991; Presser and Blair,1994; Willis, 2001). Further, several empirical investigations have evaluated the effectiveness of core features of these techniques, especially the use of verbal probing within cognitive interviewing (Davis and DeMaio 1992; Foddy, 1996) and several evaluative studies have attempted to assess the effectiveness of cognitive interviews in ameliorating questionnaire problems (Fowler and Cosenza, 2000; Lessler, Tourangeau, and Salter, 1989; Presser and Blair; Willis and Schechter, 1996; Willis, Schechter, and Whitaker, 1999); these are reviewed in detail by Willis (2001).

Increasingly, evaluations have focused on the side-by-side comparison of survey pretesting techniques, in order to determine the degree to which the results obtained through use of these techniques agree, even if they cannot be directly validated. However, this research is complex, as

evaluation in practice must take into account the multi-faceted nature of each of the pretesting techniques, and of questionnaire design in general (see Willis, DeMaio, and Harris-Kojetin, 1999). Although two studies (Presser and Blair, 1994; Willis, 2001) have specifically compared the results of cognitive interviewing, expert evaluation, and behavior coding, when these have been applied to the same questionnaire, this research has generally not been conducted in a way that allows for the separation of the effects of pretesting method from those of the organization applying these methods.

For example, Presser and Blair used expert panels whose members were different individuals than those conducting cognitive interviews, and who were in turn different from the coders who applied behavior coding. Thus, their finding that the expert panel discovered the greatest

absolute number of problems, and cognitive interviewing the least, cannot be uniquely attributed to either pretesting technique or the individuals applying them. Similarly, Willis (2001) assessed cognitive interviewing at two survey organizations, as well as behavior coding, and individual-level (as opposed to group-based) expert review. Although this study obtained relatively good correspondence between pretesting techniques, in terms of identifying candidate questions that appeared problematic and in identifying the same qualitative categories of problems, the particular techniques were again confounded with the individuals using them.

The overall objective of the current study was to rectify this limitation, and to avoid an "apples and oranges" type of comparison. Overall the selected design balanced technique with organization, for the same set of questionnaires (see Lessler and Rothgeb, 1999; Rothgeb and Willis, 1999), to determine level of agreement among three question pretesting techniques, when applied by each of three survey research organizations (The Census Bureau, Westat, and Research Triangle Institute). Therefore, we would be able to investigate the independent effects of organization, and or techniques, under conditions of controlled questionnaire content. For this research, multiple staff members within each of these organizations utilized three pretesting methods: Informal expert review, Formal cognitive appraisal, and Cognitive Interviewing. A classification scheme was then developed to code problems identified through any of the three methods, and by each organization[2].

II. Design

The experimental design was developed in order to balance each major experimental factor, so as to render the analysis as unambiguous as possible. In particular, the overall requirement was to provide a form of balancing sufficient to enable a factorial combination of Technique, Organization, and Questionnaire; that is, each technique was applied by each of the organizations

[2]Throughout this paper we refer to the detection of "problems" in tested questions by the pretesting techniques that were evaluated. We recognize, however, that the presence or absence of actual problems is unknown, given the absence of validation data. Rather, we use this terminology for purposes of labeling; that is, to indicate that the result of pretesting has been to designate the question as potentially having a problem.

to each tested questionnaire. Further, it was decided that the use of three questionnaires on varied topics would, as well as making a Latin Square design possible, also increase generalizability of the results, with respect to the range of survey questions to which the results would meaningfully apply. The Latin Square design developed is represented in Table 1. Each organization selected three researchers, and each of these researchers applied one of the depicted sequences. It was decided that each of the three researchers would evaluate all three questionnaires, and each would use all three techniques. Further, the established sequences could be replicated across each of the three organizations, so that the design table was simply repeated a total of three times.

Table 1. Latin Square-based Experimental Design: Procedure used in each of the three organizations.

Within each Organization:	**Expert review**	**Forms appraisal**	**Cognitive Interviewing**
Researcher 1	(Questionnaire A)	(Questionnaire B)	(Questionnaire C)
Researcher 2	(Questionnaire C)	(Questionnaire A)	(Questionnaire B)
Researcher 3	(Questionnaire B)	(Questionnaire C)	(Questionnaire A)

Finally, each researcher applied an invariant ordering of techniques, starting with expert review, then forms appraisal, and finally, cognitive interviewing, rather than varying this ordering. This was done partly to reflect the ordering of techniques within usual survey pretesting practice. Further, we chose not to vary the ordering of pretesting techniques because this would, in some cases, present the forms appraisal system prior to expert review, producing a source of an undesirable carryover effect, as learning the (formal) forms appraisal system would very likely influence the evaluator's (informal) expert review activities, even when applied to a different questionnaire. On the other hand, this design resulted in the switching of the questionnaire content (between A, B, and C) for each evaluation trial, from the perspective of each evaluator, and therefore did not take advantage of the natural progression across techniques that evaluators

normally experience as they apply these techniques to a single questionnaire. However, this limitation was viewed as an acceptable compromise, as the design selected allowed for the control of Pretesting Technique and Organization as the main factors of interest, and in particular, retained an uncontaminated factorial combination of Technique, Organization, and Questionnaire in a relatively efficient manner.

III. Method

Staff participating in the research consisted of a lead senior methodologist at each organization along with two other researchers at each. All participating staff had previously conducted expert reviews and cognitive interviews for other questionnaire-design projects.

A. Survey Instruments

We selected a total of 83 items which were distributed among three questionnaire modules on different survey topics, deliberately choosing subject matter with which none of the participating researchers had substantial experience. A subset of questions about expenses for telephones and owned automobiles was extracted from the U.S. Census Bureau's 1998 Consumer Expenditure Survey. Questions on transportation were extracted from the U.S. Department of Transportation's 1995 National Public Transportation Survey. Finally, questions pertaining to attitudes about environmental issues were extracted from the U.S. Environmental Protection Agency's 1999 Urban Environmental Issues Survey. We selected topics which could be administered to the general population by telephone and which contained very few skip patterns so as to maximize the number of sample cases receiving each question.

B. Pretesting Techniques

We chose to evaluate questionnaire pretesting techniques that are commonly used following initial questionnaire drafting. Expert review and cognitive interviewing are very frequently applied in Federal cognitive laboratories, and we decided to also include the forms appraisal method, which is more systematic than an expert review, but less labor intensive than cognitive interviewing.

1. Expert Review

The first method used in evaluating the questionnaires was informal, individually-based expert review. Participating researchers each independently conducted an expert review on an assigned questionnaire (A, B, or C in Table 1), and determined whether he/she thought each questionnaire item was problematic. The questionnaire review form was designed so that each item was accompanied by a 'problem indicator box' which the researcher marked if he/she perceived a potential problem with the item, for either the interviewer or the respondent. Space was also provided under each question for the researcher to write specific notes about the suspected problem. No other specific instructions were provided to the researchers conducting the expert review, except for a short description of overall questionnaire goals. Each of the three researchers at each of the three organizations completed one expert review on one assigned questionnaire module.

2. Forms Appraisal

For the forms appraisal, we utilized the Questionnaire Appraisal System (QAS) developed by Research Triangle Institute (RTI) for evaluation of draft questions for the CDC Behavioral Risk Factor Surveillance System (BRFSS). The QAS is intended mainly as a teaching tool for relatively novice questionnaire designers, and as a resource to be used by more experienced individuals. Overall, it provides a guided, checklist-based means of identifying potential flaws in survey questions (See Attachment A for a copy of the QAS.) For each survey question to be evaluated, the researcher completes a QAS form that leads the user to consider specific characteristics of the question and the researcher decides whether the item may be problematic with respect to that characteristic. There are eight general dimensions on which each item is evaluated: Reading, Instructions, Clarity, Assumptions, Knowledge/ Memory, Sensitivity/Bias, Response Categories, and Other. Within each of the eight dimensions there are several sub-dimensions for which the researcher evaluates the item, for a total of 26 separate “checks” for each survey question. For each check, the researcher circles a Yes/No box indicating whether the item is perceived to be problematic. In addition, when a "yes" is marked, the researcher also enters notes about the nature of the potential problem. The QAS was developed in order to

provide multiple means for detecting problems, rather than minimizing overlap between coding categories.

Because most of the participating researchers did not have prior experience with the QAS forms appraisal, we provided each researcher with a self-study manual. In addition, researchers completed a few practice exercise test questions using the forms appraisal, and their completed work was reviewed by the project manager at that organization. Then researchers were given their assigned module, additional instructions, and QAS forms to complete. Each of the three researchers at the three organizations completed a QAS for each questionnaire item in their assigned module.

3. Cognitive Interviews

Our third pretesting method was cognitive interviewing. Each organization independently developed a cognitive interview protocol, after expert reviews and forms appraisals had been completed. Because there is variation between organizations in the degree of use of scripted versus unscripted probing, and in the manner in which results are summarized, we did not attempt to standardize these aspects of the research, as such differences between organizations were of interest. Each organization independently recruited research subjects. Each interview was expected to last approximately one hour. Cognitive interviews were conducted both in the organizations' cognitive laboratories and off-site at locations convenient to subjects. All laboratory subjects were either staff of the organizations, or members of the general population who were 18 years of age or older. Each of the three researchers within each organization conducted three cognitive interviews with their assigned modules. As with the other testing techniques, researchers marked a problem indicator box after each questionnaire item, for each interview, when they believed that a potential problem existed, and entered open-ended written comments for marked questions.

After the three cognitive interviews at each organizations were completed, the head researcher from each organization reviewed and summarized these results, by making a determination of

whether, for each tested item, significant problems had been detected for that item. We believed that this approach most closely replicates usual practice of cognitive interviewers, as results from all cognitive interviews (rather than each individual cognitive interview) conducted by a particular interviewer are typically evaluated in total to determine where in the questionnaire problems may exist. This practice also served to equate scores based on cognitive interview results with those from the expert review and forms appraisal, for which each questionnaire item was coded only once by each technique as a potential problem.

IV. Results

A. Item Summary Score computation

The initial level of analysis involved only the number of problems identified as problematic, and not the qualitative nature of problems. In order to determine whether pretesting techniques were consistent in their identification of individual problems as problematic, each item was given a dichotomous score (Problem versus No-Problem) by each researcher, for each of the three pretesting techniques (expert review, forms appraisal, and cognitive interviews). Then, for each of the 83 items across the three questionnaires, a Summary Score consisting of the total number of times a problem was assigned was assessed. Summary scores were computed both by assessing: a) the number of organizations that identified a problem, under each technique (e.g., Census, RTI, and/or Westat under cognitive interviewing); and b) the number of Techniques that identified that item as problematic, within each Organization (e.g., whether Expert Review, Forms Appraisal, and/or Cognitive Interviewing identified the item, when tested at the Census Bureau). Each of these scores could therefore range between 0 and 3.

B. Analysis of Summary Scores

The foundation of our analysis was based on the Summary Scores for each pretesting Technique, and for each research Organization. In our analysis we examined differences between mean item scores (through ANOVA), and correlations between item scores. Results of each are described below.

1. Analysis of differences between Pretesting Techniques

The mean item scores (given a possible minimum of 0 and a maximum of 3) for each pretesting technique were as follows: a) Expert Review: 1.55; b) QAS: 2.93; c) Cognitive Interviews: 1.46. These results suggest that the Question Appraisal System was the most productive in identifying potential questionnaire problems (on average, it found a problem in 2.93 of 3 possible opportunities, or 97.7%). Although the forms appraisal is very sensitive in detecting potential problems, one might question the specificity of this method: The fact that there is very little variation (basically every item was found to have one or more problems) seems to represent the "promiscuous use" of coding with this method. On the other hand, the means of the items scores for the expert review and cognitive interviews indicate that they both identified potential problems about half the time, on average.

To determine whether the pretesting Techniques found significantly different numbers of problems, and whether they found different number of problems in each of the three questionnaire modules, analysis of variance (ANOVA) was conducted. The unit of analysis, or "case," was the questionnaire item; the independent variables were questionnaire module and pretesting technique; and the dependent variable was the Summary Score, or number of times each item was flagged as having a problem (0-3). The Questionnaire (A, B, or C) was equivalent to the 'between-subject' factor and pretesting technique the 'within-subject' or repeated measures factor. The ANOVA results indicated that questionnaire module had no overall effect on problem identification frequency, but there was a large difference by pretesting technique ($F=92.8$, $p<.001$). There was no significant interaction between questionnaire module and pretesting technique ($F=1.8$, $p<.13$).

To determine where differences were within the overall pretesting technique effect, a two-way ANOVA was conducted among the pairs of pretesting techniques. ANOVA results for *expert review versus cognitive interviewing* indicated no significant differences, and a marginal interaction between questionnaire module and pretesting technique ($F=2.78$, $p<.07$). ANOVA results for *expert review versus forms appraisal* indicated a large difference ($F=157.60$, $p<.001$)

between item scores for expert review and the forms appraisal, independent of the questionnaire module ($F=1.98$, $p<.14$). Similarly, ANOVA results comparing items scores between *forms appraisal versus cognitive interviewing* revealed a large difference ($F=153.03$, $p<.001$) between the two techniques, independent of questionnaire modules ($F=.23$, $p<.4$).

Spearman correlation analyses were then conducted to determine the degree to which the different pretesting techniques determined the same questionnaire items to be problematic. Because of ceiling effects (and resultant restriction in range) of the item scores for the forms appraisal, only the expert review and cognitive interviews could be meaningfully correlated. The correlation coefficient for Spearman's r between the summary scores for expert review and cognitive interviews was .27 ($p<.02$), demonstrating positive, but low correlation between the two methods in the items identified as problematic.

2. Analysis of Differences Between Research Organizations

Similar to the test of differences as a function of Technique, the mean scores (range of 0-3) for each research organization were as follows: a) Census: 1.95; b) RTI: 2.02; c) Westat: 1.96. The similarity in the mean scores demonstrates that a comparable criterion level in identifying problems was adopted, overall, across organizations. Analysis of variance conducted to determine whether the research organizations obtained different number of problems and whether they found the same or different number in each of the three questionnaire modules revealed no significant effect of questionnaire module, organization, or interaction between module and organization.

Spearman correlations between the item Summary Scores produced by different organizations (across all pretesting techniques) were very similar, and all low-moderate in magnitude: a) Census - RTI: .38 b) Census - Westat: .34, c) RTI - Westat: .38, all $p < .001$. However, because Spearman correlation may not itself be a sufficient measure, a number of other measures were computed to assess the key issue of level of agreement between organizations: In a set of pairwise comparisons (the three organizations compared two at a time), Kappa statistics averaged

approximately .3, Yule's Q statistics averaged .6, tetrachoric correlations averaged .5, Robinson's A averaged .7, and the Intraclass Correlation Coefficient (ICC) two-way random effects model revealed single-measure (pairwise) values of approximately .4. For analysis of the consistency between Organizations taken altogether (rather than pairwise), the meaned ICC value was .62, and the coefficient alpha reliability level was .62. Altogether, although these measures vary greatly in magnitude, and it is difficult to know which applies best, it appears that there was an overall moderate level of agreement between the Organizations conducting the pretesting.

Overall, the pattern of results portrayed above showed that different Organizations behaved fairly consistently with respect to how often they selected questions as problematic (they used similar overall criterion levels). However, they agreed only to a moderate degree with respect to which particular items were problematic. To some degree, it may be unrealistic, under the design used, to expect a large degree of item-specific agreement. Most importantly, only three interviewers were used at each organization, and each interviewer conducted only three interviews; hence, variability with respect to both interviewer and subject characteristics could have been very high. However, these parameters are fairly representative of common practice within cognitive laboratories (e.g., nine total subjects per single round of interviewing), so that the restriction imposed in this experiment is not artificial in nature.

C. Qualitative coding of problems

Although it is useful to determine whether different techniques and organizations produce different number of problems identified, we are most interested in determining whether the *types* of problems uncovered by various techniques and organizations are similar or different. To determine the source of the identified problems, we developed a qualitative coding system (described in the next section) which could be applied to the results of all three pretesting techniques. However, because of resource constraints we decided to qualitatively code only the 15 items which were identified as the most problematic, based on the total summary scores computed for the previous analysis (see Attachment B for question wordings of these 15 items).

1. Classification Coding Scheme

The Classification Coding Scheme (CCS) (Attachment C) was developed to reflect several categories, as well as sub-categories, of question problems. The 28 CCS codes are grouped, at the highest level, under the familiar headings of the four-stage cognitive response model: a) comprehension and communication, b) retrieval, c) judgement and evaluation, and d) response selection. Within each of the four stages were mid-level categories, and at the lowest level, the most detailed description of the problem; for example -- *undefined technical term; complex estimation; complex or awkward syntax*. It was important that the CCS codes be independent of one another and that rules be established on the use of any codes which may be ambiguous. In contrast to the QAS, the CCS was developed in order to attempt to maximize inter-rater agreement, with respect to assignment of individual codes[3].

2. Application of CCS Scheme to Questions

The three lead researchers worked together to assign CCS codes to the 15 selected items, by reviewing the open-ended researcher notes concerning the problems that had been identified through each of the three pretest methods by each or the three organization (hence, each item received nine evaluations). Each item received as many codes as the researchers agreed were found to apply to that item, based on the written comments only[4].

3. Results of coding scheme application

Table 3 illustrates the frequency with each of the 27 CCS codes assigned, overall, to the 15 selected items. Collectively, the lead researchers identified a total of 338 problems, across nine separate (Technique X Organization) evaluations, for an average of 2.5 codes per question. From Table 3, it is clear that a small number of codes accounted for a large proportion of problems identified. Six codes *(Difficult for interviewer to administer, Vague topic/unclear*

[3]Note that the lowest-level CCS codes are in fact very similar to those used in the QAS. This similarity may reflect a tendency for question coding systems to converge on a key set of problems that are relatively standard across questionnaires.

[4]Although the QAS system provided its own coding system, only the QAS written notes were coded, in order to maintain consistency across pretesting techniques.

question, Undefined/vague term,
Undefined reference period, High detail required/information unavailable, and Erroneous assumption) accounted for 69.9 percent of all identified problems

TABLE 3. Frequency of CCR Codes assigned to 15 most problematic items

Code	Problem label		Frequency	Percent
Comprehension and Communication -				
Interviewer difficulties:				
1	Inaccurate Instruction		3	0.9
2	Complicated Instruction		2	0.6
3	Difficult for interviewer to administer		23	6.8
Question content:				
4	Vague topic/Unclear question		48	14.2
5	Complex topic		1	0.3
6	Topic carried over from previous question		0	0.0
7	Undefined/vague term		58	17.2
Question structure:				
8	Transition needed	1	0.3	
9	Unclear instruction to respondent		10	3.0
10	Question too long	8	2.4	
11	Complex or awkward syntax		16	4.7
12	Erroneous assumption		33	9.8
13	Several questions		4	1.2
Reference period:				
14	Reference period carried over		2	0.6
15	Undefined reference period		28	8.3
16	Unanchored or rolling reference period		0	0.0
Retrieval from memory -				
17	Shortage of memory cues		3	0.9
18	High detail required/information unavailable		46	13.6
19	Long recall/reference period		3	0.9
Judgment/Evaluation -				
20	Complex estimation		8	2.4
21	Potentially sensitive/biasing		6	1.8
Response Selection -				
Response terminology:				
22	Undefined term in response category		4	1.2
23	Vague terms in response categories		0	0.0
Response units				
24	Response categories contain wrong/mismatching units		4	1.2
25	Unclear to respondent what response categories are		3	0.9
Response structure				
26	Overlapping response categories		0	0.0
27	Missing response categories		13	3.8
28 ***OTHER (uncodable)*** --			11	3.3
TOTAL			338	100.0

Note that all of these codes were classed by the CCS system as comprehension/communication and retrieval problems, and none of these codes were from the judgement stage or response stage. Further, two codes (vague topic/unclear question and undefined/vague term) account for 31.4 percent of all problems. These results are consistent with findings from Presser and Blair (1997) and Willis (2001), who found vagueness and inclarity to dominate their qualitative coding results as well.

Analysis of Mid-Level CCS Categories

Because of the preponderance of small cell sizes at the lowest level of coding, the 28 CCS codes were recoded up to 11 "mid-level"categories in Attachment C. Further, the category *Erroneous Assumptions* was separated out from Question Structure as a separate category, as this appeared upon reflection to constitute a qualitatively separate category relating more to underlying logical structure than to Comprehension and Communication, which relate more directly to the communication of an underlying structure. Table 6 illustrates the resultant categories, as applied by each Pretesting technique.

From Table 6, it seems that Techniques differed in terms of the problems they identify. For example, cognitive interviews did not appear to detect interviewer difficulties, but were apparently sensitive in detecting potential problems with question content. Presser and Blair (1994) also found that cognitive interviews did not serve to detect interviewer problems. It further appears that expert review might be more sensitive in detecting problems with question structure than forms appraisal or cognitive interviews. However, these observations are somewhat speculative; again due to the small sizes of some cells, we were unable to conduct summary Chi-square tests.

Table 6. Mid-level CCS categories by pretesting technique

"Mid-level" CCS category	Expert review	QAS	Cognitive Interviews	Total number of codes assigned
Interviewer difficulties	9.2%	11.1%	0.0%	28
Question content	29.2%	25.3%	50.7%	107
Question structure	21.5%	11.1%	4.0%	39
Reference period	4.6%	11.1%	6.7%	30
Retrieval	7.7%	16.7%	18.7%	52
Judgment/Evaluation	3.1%	5.1%	2.7%	14
Response Terminology	0.0%	1.5%	1.3%	4
Response Units	1.5%	2.0%	2.7%	7
Response Structure	4.6%	4.6%	1.3%	13
Erroneous Assumption	9.2%	10.1%	9.3%	33
Other	9.2%	1.5%	2.7%	11
Total	100.0%	100.0%	100.0%	--
(n)	65	198	75	338

Table 7 presents the distribution of mid-level CCS categories by Organization. Few differences appear across techniques, so that it again appears that the distribution of problem types is more similar across organizations than it is across pretesting technique.

Table 7. Mid-level CCS categories by research organization.

Mid-level CCS category	Census	RTI	Westat	Total number of codes assigned
Interviewer difficulties	6.6%	10.8%	7.8%	28
Question content	31.4%	33.3%	30.4%	107
Question structure	13.2%	10.8%	10.4%	39
Reference period	9.9%	8.8%	7.8%	30
Retrieval	17.4%	12.8%	15.7%	52
Judgment/Evaluation	5.0%	1.0%	6.1%	14
Response Terminology	0.0%	2.0%	1.7%	4
Response Units	1.7%	2.0%	2.6%	7
Response Structure	3.3%	2.9%	5.2%	3
Erroneous Assumptions	9.9%	10.8%	8.7%	33
Other	1.7%	4.9%	3.5%	11
Total	100.0%	100.0%	100.0%	--
n	121	102	115	338

Analysis of CCS Categories at Highest Coding Level (Cognitive Processing Model)

Finally, the data were further collapsed according to each of the stages in a four-stage cognitive response model. Note that the problems identified as "erroneous assumptions" and "something else" are excluded from Tables 8 and 9. In addition, due to small cell sizes, the 'judgment and evaluation', and 'response selection' problems were collapsed. Table 8 shows the distribution of problems identified according to pretesting Technique. Chi-square testing did not reveal a statistically significant association between category of problem identified and Technique (Chi-sq (4) =4.99, p<.29). However, the most compelling result appears to be that problems related to comprehension and communication are the overwhelming majority of problems identified, which

is consistent with findings from earlier research by Presser and Blair (1994) and Willis, (2001).

TABLE 8. CCS Highest level Coding Category Distribution, by Pretesting Technique

CCS Top Level Categories	**Expert Review**	**QAS**	**Cognitive Interview**	**Total**
Comprehension and Communication	42 79.3%	116 66.3%	46 69.7%	204 69.4%
Retrieval from Memory	5 9.4%	33 18.9%	14 21.2%	52 17.7%
Judgment and Evaluation, and Response Selection	6 11.3%	26 14.9%	6 9.1%	38 12.9%
Total	53 100.0%	175 100.0%	66 100.0%	294 100.0%

Finally, Table 9 displays the distribution of problems identified at the most general cognitive response model level by organization. Overall there was no association between the application of codes and Organization (chi-sq(4) =4.384, p<.357).

TABLE 9. CCS Highest Level Coding Category Distribution, by Organization

CCS Top Level Categories	**Census**	**RTI**	**Westat**	**Total**
Comprehension and Communication	74 69.2%	65 75.6%	65 64.4%	204 69.4%
Retrieval from Memory	21 19.6%	13 15.1%	18 17.8%	52 17.7%
Judgement and Evaluation and Response Selection	12 11.2%	8 9.3%	18 17.8%	38 12.9%
Total	107 100.0%	86 100.0%	101 100.0%	294 100.0%

V. Discussion

A. Assignment of 'problem' status to questions: Quantitative Analysis

1. Comparison of techniques. In the current study, the three pretesting techniques of expert review, question appraisal, and cognitive interviewing revealed (using the dichotomous Problem Assignment indicator) the Question Appraisal System to be the most "productive" in identifying question problems. However, given the extremely high frequency with which this technique detected problems, it is very possible that such an appraisal method, as we applied it, may encourage a low threshold for problem identification, therefore producing a large number of false positives results. Therefore we suspect that the QAS method, as used, has high sensitivity but poor specificity. It is possible that the Question Appraisal method is accurate in its identification of problems, and that the questions used for this research do possess the many problems that system revealed. However, it is also likely that a system that is designed mainly as an aid to the questionnaire designer, rather than a pretesting technique, requires a fair degree of additional expert judgment to be practically effective during the pretest phase.

However, even the finding of vastly greater total problems in the QAS is an ambiguous finding, however, because it is in one sense an artifact of the analysis procedures used. For current purposes, a question was scored as problematic by the QAS if it failed to "pass" each of 26 separate tests, providing an extremely high standard for any survey question. If instead, one were to establish a higher threshold, based on either total number of problems found (e.g., 6 out of 26 failed tests), or an index that was weighted by the anticipated severity of certain types of problems, the results might have looked very different. In fact, the problem is identical to that posed by analysis of behavior coding studies, which provides a continuous distribution of code frequency, and requires the establishment of a threshold value (typically 15-20%) when one is making a dichotomous decision related to whether pretesting has detected a significant problem. In any event, these results do appear to support very emphatically the conclusion of Willis et al. (1999) that any evaluation design depending on the notion that "finding more problems is better" is suspect, because of the exclusive focus on technique sensitivity.

Interestingly, expert review and cognitive interviewing produced very similar results in the current study, in terms of the numbers of problems identified. This is in contrast to the findings of Presser and Blair (1994) and Willis (2001) where expert review was the most productive in identifying question problems. While expert review and cognitive interviewing produced similar numbers of problems, the specific items identified as problematic varied between the two methods, and unlike the results reported by Willis (2001) the correlation between these techniques was rather low. It is not clear what factors led to these discrepancies. However, one difference may relate to the fact that the current study analyzed questionnaires that appear to have contained a multitude of problems, whereas previous ones (Presser and Blair, 1994; Willis 2001) utilized questions that contained flaws, but were generally more useful in their current form. Overall, the current study revealed that approximately half the time an item was evaluated by expert review or cognitive interviewing, and virtually any time it was evaluated via the QAS, it was "flagged" as problematic. Further, this result was obtained independently by three very experienced survey organizations, which suggests a degree of convergent validity. It may be that the tested questions exhibited so many severe problems that each pretesting technique in effect simply selected a different subset of these, and that all may have been "correct" to some extent. Some of the problems with these items may also be because we extracted them from various questionnaires and administered them out of context from their original surveys. Presumably as questions near a more final state in which they contain only one or two serious problems, pretesting techniques might be expected to converge on those problems, producing greater levels of agreement.

2. Consistency across organizations. One interesting finding from the current study was that the results among organizations were far more similar than were the results across techniques. Our findings suggest that the different organizations use similar criteria in determining potentially problematic questionnaire items, at least in terms of general proportion of items selected. However, the more significant issue is whether the different organizations selected the same items as having problems; and it was found that selection of problematic items across organizations was only moderate in magnitude, and lower than those previously reported in a

comparison of two organizations by Willis (2001). However, note that these statistical results were based on data having only four potential values (0, 1, 2, 3), and that a value of 0 was used only twice across the 83 items, reducing the effective overall range of the dependent measure to three items. A classical restriction-in-range effect could therefore be responsible for the relative modesty of the obtained relationships, and mask a much greater degree of implicit agreement across organizations.

3. Assignment of type of problem (CCS code) to 15 worst questions: Qualitative analysis. Classification of the types of problems identified through the three pretesting techniques produced interesting results at the 'middle-level' of qualitative coding analysis: Problems associated with question content constituted the single largest category of problems detected for all three techniques. For cognitive interviewing, question content comprised over half of the detected problems, whereas for the other two techniques this category accounted for a little more than a quarter of the identified problems. The appraisal scheme and cognitive interviewing detected problems with information retrieval to a greater extent than did expert review. Expert review and the QAS tended to detect interviewer difficulties, whereas cognitive interviewing tended not to. Overall, the types of problems identified through cognitive interviewing were highly clustered within a few problem types, the categories of problems identified through expert review were somewhat less clustered, and the QAS results were the least clustered in this regard (as it found a multitude of problems).

Examination of the types of problems found at the most general, cognitive processing model level also demonstrated that comprehension and communication problems were identified to the greatest extent by all three techniques, similar to previous findings (Presser and Blair, 1994; Willis, et al., 2000.). Note that in a sense this may not be surprising, simply given the number of total codes devoted to this general category in the CCS system that was developed.

VI. Conclusions and caveats

Based on the results of this research project, each of the three pretesting methods contributes

somewhat differently to the identification of problems in survey questions, in terms of the types of problems identified. However, the differences we observed were largely quantitative, rather than qualitative; with limited variation, these techniques appeared to be the must useful in ferreting out problems related to question comprehension, across three very different questionnaires. The observed consistency of results across organizations is potentially important, because this suggests that there may also be consistency in the ways that the techniques are being used, and the nature of the results produced. The relative lack of consistency across organizations in choosing *which* particular items were problematic is somewhat troubling, although it could also be argued that there was very little disagreement with respect to which of these items were severely flawed.

However, the current study does not address two further vital questions – (a) How do we know that the problems that are identified through pretesting actually exist in the field environment, and (b) Even if the identified problems are "real", what assurance do we have that the modifications that we make to these questions serve to rectify these problems without also introducing new ones? The former issue has been addressed very minimally (Davis and DeMaio, 1993; Willis and Schechter, 1997), and the latter is an almost completely unexplored area. An extension of the current study is now being undertaken to address these research questions. Specifically, we are conducting an experiment in which the original 15 problematic items used in this study, as well as revisions of those items, are administered in a split-sample experiment. Analysis of the field results will then be evaluated, using several independent outcome quality measures (e.g., behavior coding, interviewer rating forms). Comparing the results from each pretesting method with the results of the field study should aid us in determining how well the various pretesting methods identified the types of problems which surfaced during field testing. Further, comparing the outcome quality measures from the field study for the original and revised question wordings will reveal whether revisions in question wording, based on pretesting results, actually improved data quality.

VII. References

Beatty, P. (undated). *Classifying Questionnaire Problems: Five Recent Taxonomies and One Older One*. Unpublished manuscript, Office of Research and Methodology, National Center for Health Statistics.

Beatty, P., Willis, G. B., and Schechter, S. (1997). Evaluating the generalizability of cognitive interview findings. In *Office of Management and Budget Seminar on Statistical Methodology in the Public Service, Statistical Policy Working Paper 26*, pp. 353-362. Washington, DC: Statistical Policy Office.

Campanelli, P. (1997). Testing survey questions: New directions in cognitive interviewing. *Bulletin de Methodologie Sociologique, 55*, 5-17.

Conrad, F., and Blair, J. (1997). From impressions to data: Increasing the objectivity of cognitive interviews. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.

Davis, W. L., and DeMaio, T. J. (1993). Comparing the think-aloud interviewing technique with standard interviewing in the redesign of a dietary recall questionnaire. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 565-570.

DeMaio, T., Mathiowetz, N., Rothgeb, J., Beach, M. E., and Durant, S. (1993). *Protocol for pretesting demographic surveys at the Census Bureau*. Unpublished manuscript, Center for Survey Methods Research, U.S. Bureau of the Census.

Esposito, J. L., and Rothgeb, J. M. (1997). Evaluating survey data: Making the transition from Pretesting to Quality Assessment. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (Eds.), *Survey Measurement and Process Quality*, pp 541-571. New York: Wiley.

Foddy, W. (1996). The In-Depth Testing of Survey Questions: A Critical Appraisal of Methods. Quality and Quantity, 30, 361-370.

Forsyth, B., and Lessler, J. (1991). Cognitive Laboratory Methods. In P. Biemer *et al.* (eds.), *Measurement Errors in Surveys*, New York: Wiley.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly*, *56*: 218-231.

Fowler, F. J., and Cosenza, C. (1999). Evaluating the results of cognitive interviews. *Proceedings of the Workshop on Quality Issues in Question Testing*, Office for National Statistics, London, 35-41.

Gerber, E. R., and Wellens, T. R. (1997). Perspectives on pretesting: "Cognition" in the cognitive interview? *Bulletin de Methodologie Sociologique, 55*, 18-39.

Groves, R. M. (1996). How do we know that what we think they think is really what they think? In N. Schwarz and S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 389-402). San Francisco: Jossey-Bass.

Lessler, J. T., and Forsyth, B. H. (1996). A coding system for appraising questionnaires. In N. Schwarz and S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 259-291). San Francisco: Jossey-Bass.

Lessler, J. T., and Rothgeb, J. (1999). Integrating cognitive research into household survey design. In *A New Agenda for Interdisciplinary Survey Research Methods: Proceedings of the CASM II Seminar*. National Center for Health Statistics, pp. 67-69.

Lessler, J.T., Tourangeau, R. and Salter, W. (1989). Questionnaire design research in the cognitive research laboratory. Vital and Health Statistics (Series 6, No. 1; DHHS Publication No. PHS-89-1076). Washington, DC: U.S. Government Printing Office.

Oksenberg, L.,Cannell, C., and Kalton, G. (1991), New Strategies for Pretesting Survey Questions. *Journal of Official Statistics*, 7, 3, pp. 349-365.

Presser, J., and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results?, in P.V. Marsden (ed.), *Sociological Methodology*, Vol. 24, Washington, DC: American Sociological Association, pp. 73-104.

Rothgeb, J., and Willis, G. (1999). Evaluating pretesting techniques for finding and fixing questionnaire problems. *Proceedings of the Workshop on Quality Issues in Question Testing*, Office for National Statistics, London, 100-102.

Tucker, C. (1997). Measurement issues surrounding the use of cognitive methods in survey research. *Bulletin de Methodologie Sociologique, 55*, 67-92.

Tourangeau, R. (1984). Cognitive Science and Survey Methods. In T. Jabine *et al.* (eds.), *Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines*, Washington: National Academy Press, pp. 73-100.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). The Psychology of Survey Response. Cambridge: Cambridge University Press.

Willis, G. B. (2001). *A Comparison of Survey Pretesting Methods: What do Cognitive Interviewing, Expert Review, and Behavior Coding Tell Us?* Paper submitted for publication.

Willis, G. B. (1994). *Cognitive interviewing and Questionnaire Design: A Training Manual*. National Center for Health Statistics: Cognitive Methods Staff (Working Paper No. 7).

Willis, G.B., DeMaio T.J, and Harris-Kojetin B. (1999). *Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive interviewing Techniques.* In M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (Eds.). Cognition and Survey Research. New York: Wiley, 133-153.

Willis, G.B., and Schechter, S. (1997). Evaluation of Cognitive interviewing Techniques: Do the Results Generalize to the Field? *Bulletin de Methodologie Sociologique*, 55, pp. 40-66.

Willis, G. B., S. Schechter, and K. Whitaker (2000). "A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What do They Tell Us?" *American*

Statistical Association, Proceedings of the Section on Survey Research Methods.

ATTACHMENT A

Census Bureau/RTI/Westat Pretesting Research Project

QUESTION APPRAISAL SYSTEM (QAS):

CODING FORM

INSTRUCTIONS. Use one form for EACH question to be reviewed. In reviewing each question:

1) WRITE OR TYPE IN QUESTION NUMBER AND INCLUDE THE FULL QUESTION TEXT (INCLUDING RESPONSE CATEGORIES) HERE:

Question number or question here:

2) Proceed through the form - Circle or highlight YES or NO for each Problem Type (1a... 8).

3) Whenever a YES is circled, write detailed notes on this form that describe the problem.

STEP 1 - READING: Determine if it is difficult for the interviewers to read the question uniformly to all respondents.	
1a. **WHAT TO READ**: Interviewer may have difficulty determining what *parts* of the question should be read.	YES NO
1b. **MISSING INFORMATION**: Information the interviewer needs to administer the question is *not* contained in the question.	YES NO
1c. **HOW TO READ**: Question is *not* fully scripted and therefore difficult to read.	YES NO
STEP 2 - INSTRUCTIONS: Look for problems with any introductions, instructions, or explanations from the *respondent's* point of view.	
2a. **CONFLICTING OR INACCURATE INSTRUCTIONS**, introductions, or explanations.	YES NO

2b. **COMPLICATED INSTRUCTIONS**, introductions, or explanations.	YES NO
STEP 3 - CLARITY: Identify problems related to communicating the *intent or meaning* of the question to the respondent.	
3a. **WORDING**: Question is lengthy, awkward, ungrammatical, or contains complicated syntax**.**	YES NO
3b. **TECHNICAL TERM(S)** are undefined, unclear, or complex.	YES NO
3c. **VAGUE:** There are multiple ways to interpret the question or to decide what is to be included or excluded.	YES NO
3d. **REFERENCE PERIODS** are missing, not well specified, or in conflict.	YES NO
STEP 4 - ASSUMPTIONS: Determine if there are problems with assumptions made or the underlying logic.	
4a. **INAPPROPRIATE ASSUMPTIONS** are made about the respondent or about his/her living situation**.**	YES NO
4b. **ASSUMES CONSTANT BEHAVIOR** or experience for situations that vary.	YES NO
4c. **DOUBLE-BARRELED**: Contains more than one implicit question.	YES NO

STEP 5 - KNOWLEDGE/MEMORY: Check whether respondents are likely to *not know* or have trouble *remembering* information.	
5a. **KNOWLEDGE** may not exist: Respondent is unlikely to *know* the answer to a factual question.	YES NO
5b. **ATTITUDE** may not exist: Respondent is unlikely to have formed the attitude being asked about.	YES NO
5c. **RECALL** failure: Respondent may not *remember* the information asked for.	YES NO
5d. **COMPUTATION** problem: The question requires a difficult mental calculation.	YES NO
STEP 6 - SENSITIVITY/BIAS: Assess questions for sensitive nature or wording, and for bias.	
6a. **SENSITIVE CONTENT** (general): The question asks about a topic that is embarrassing, very private, or that involves illegal behavior.	YES NO
6b. **SENSITIVE WORDING** (specific): Given that the general topic is sensitive, the wording should be improved to minimize sensitivity.	YES NO
6c. **SOCIALLY ACCEPTABLE** response is implied by the question.	YES NO

STEP 7 - RESPONSE CATEGORIES: Assess the adequacy of the range of responses to be recorded.	
7a. **OPEN-ENDED QUESTION** that is inappropriate or difficult.	YES NO
7b. **MISMATCH** between question and response categories.	YES NO
7c. **TECHNICAL TERM(S)** are undefined, unclear, or complex.	YES NO
7d. **VAGUE** response categories are subject to multiple interpretations.	YES NO
7e. **OVERLAPPING** response categories.	YES NO
7f. **MISSING** eligible responses in response categories.	YES NO
7g. **ILLOGICAL ORDER** of response categories.	YES NO
STEP 8 - OTHER PROBLEMS: Look for problems not identified in Steps 1 - 7.	
8. Other problems not previously identified.	YES NO

ATTACHMENT B

Question Wording of 15 "Worst" Items selected for CCS (qualitative) coding

Consumer Expenditure Questions

- What property(ies) was (*were*) the telephone bill for?

  Mobile (car) phone ☐
  Rented sample unit ☐
  Other rented unit ☐
  Property not owned or rented by CU ☐

- What is the name of the company which provides telephone services for (*property description*)?

- What was the total amount of bills (*bill numbers*)? Exclude any unpaid bills from a previous billing period?
  $ _____

- In what month was the bill received?

Transportation Questions

- How many cylinders does it have?

- Is it used for business?
  Yes, used for business ☐
  No, personal use only ☐

- Is local bus service available in your town or city?
  Yes ☐
  No ☐

  (Include only services that are available for use by the general public for local or commuter travel, including dial-a-bus and senior citizen bus service. Do not include long distance buses or those chartered for specific trips.)

- Is subway, commuter train, or streetcar service available in your town or city?
  Yes ☐
  No ☐
  (Include only services that are available for use by the general public for local or commuter travel, including elevated trains. Do not include long distance services or those chartered for specific trips.)

Environmental Questions

- First, I am going to read you a list of different issues that may or may not occur in the *[PLACE NAME OF MSA HERE*] area. Some issues are about the urban environment and others are about topics, such as schools and roads. Understanding how important all of these issues are to you, will help EPA and other agencies better serve you. I am going to read the list of issues and I want you to tell me how high or low a priority each is in the [*PLACE NAME OF MSA HERE*] area. Use a scale of one to ten, with one meaning "very low priority" and ten meaning "very high priority."

- Greater protection for ground water and wells 1 2 3 4 5 6 7 8 9 10 Dk

- Depletion of the water table 1 2 3 4 5 6 7 8 9 10 Dk

- Now I would like you to rate the following groups and organizations on how well they provide you with information about environmental conditions in the *[PLACE NAME OF MSA HERE*] area. Please rate these groups using a scale from 1 to 10, with 10 being EXCELLENT and 1 being VERY POOR.

- Let's start with...

  Television 1 2 3 4 5 6 7 8 9 10 Dk

- The next few questions are about your household and the environment. When we use the word "environment" we mean the air you breathe, the water you drink, the place where you live, work and play, and the food you eat. It also means the climate, wild animals, recycling and more. When you think about the environment this way, have you or anyone else in your household age 18 and older:

- Requested environmental information in person, in writing, or by phone?

ATTACHMENT C: CCS Coding Scheme

Comprehension and Communication		Retrieve from Memory	Judgement and Evaluation	Response Selection
		17 Shortage of memory cues 18 High detail required or information unavailable 19 Long recall period or long reference period	20 Complex estimation, difficult mental arithmetic required; (Guessing or heuristic estimation may be likely) 21 Potentially sensitive or desirability bias	Response Terminology 22 Undefined term(s) 23 Vague term(s)
Interviewer Difficulties 1. Inaccurate Instructions (move to wrong place; skip error 2. Complicated instruction 3. Difficult for interviewer to administer	Question Structure 8. Transition needed 9. Unclear respondent instruction 10 Question too long 11 Complex or awkward syntax 12 Erroneous assumption 13 Several questions			Response Units 24 Responses use wrong or mismatching units 25 Unclear to R what Resp. option are
Question Content 4 Vague topic/unclear Q 5 Complex topic 6 Topic carried over from earlier question 7 Undefined term(s)) vague term	Reference 14 Reference period carried over from earlier question 15 Undefined reference period 16 Unanchored or rolling reference period			Response Structure 26 Overlapping categories 27 Missing response categories

28 Something else _____