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

Jennifer Rothgeb, Gordon Willis, and Barbara Forsyth,
Jennifer Rothgeb, U.S. Census Bureau, SRD, Washington, D.C. 20233-9100

Key Words: Pretesting Techniques, Pretesting Methods,

I. Introduction

During the past 15 years, in an effort to improve survey data quality, survey practitioners have significantly increased their use of an evolving set of questionnaire pretesting methods. 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.

The overall objective of this study was to rectify this limitation. 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 pretesting techniques, when applied by each of three survey research organizations. For this research, multiple researchers within each of the organizations used three pretesting methods: Informal expert review, Formal cognitive appraisal, and Cognitive Interviewing. A classification scheme was developed to code problems identified through any of the methods, and by each organization[1].

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

II. Design

The experimental design was developed to balance each experimental factor, to render the analysis as unambiguous as possible. Further, we 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 be meaningful. We decided that each of the three researchers would evaluate all three questionnaires, and each would use all three techniques (each on a different questionnaire module.) Each researcher applied an invariant ordering of techniques, (expert review, forms appraisal, cognitive interviewing) rather than varying this ordering.

III. Method

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

A. Survey Instruments

We evaluated 83 items distributed among three questionnaire modules on different topics, deliberately choosing subject matter with which none of the researchers had substantial experience. We selected topics (consumer expenditures, transportation, and environmental issues) and items which could be administered to the general population by telephone and which contained few skip patterns to maximize the number of 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 frequently applied in Federal cognitive laboratories, and we decided to also include the forms appraisal method, as it is more systematic than an expert review, but less labor intensive than cognitive interviewing.

1. Expert Review

The first method used was individually-based expert review. Researchers independently conducted an expert review on an assigned questionnaire determining whether each item was problematic. The questionnaire review form was designed so that each item had by a 'problem indicator box' which the researcher marked if he/she perceived a potential problem. Space was provided at each question for notes about suspected problems. 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. The QAS provides a checklist-based means of identifying potential flaws in survey questions. For each survey question, the researcher completes a QAS form that leads the user to consider specific characteristics and the

researcher decides whether the item may be problematic with respect to that characteristic. The eight dimensions on which each item is evaluated are: 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. When a "yes" is marked, the researcher also enters notes about the nature of the potential problem. 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 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 of the three researchers within each organization conducted three cognitive interviews with their assigned modules. 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 nine cognitive interviews at each organization were completed, the head researcher from each organization reviewed the results, making a determination of whether, for each tested item, significant problems had been detected.

IV. Results

A. Item Summary Score computation

The initial analysis involved only the number of problems identified as problematic, not the qualitative nature of problems. 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. Then, for each of the 83 items, 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; and b) the number of Techniques that identified that item as problematic, within each Organization. Each of these scores could range between 0 and 3.

B. Analysis of Summary Scores

The foundation of our analysis was the Summary Scores of each pretesting Technique, and of each Organization, and examined differences between mean item scores, and correlations between item scores. Results of each are described below.

1. Analysis of differences between Pretesting Techniques

The mean item scores 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. 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. The means of the items scores for the expert review and cognitive interviews indicate that they both identified potential problems about half the time.

To determine whether pretesting techniques found significantly different numbers of problems, and whether they found different numbers of problems in each of the three questionnaire modules, analysis of variance (ANOVA) was conducted. The unit of analysis was the questionnaire item; the independent variables were questionnaire module and pretesting technique. 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 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 numbers 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 $p < .001$.

Overall, the pattern of results portrayed above showed that different Organizations behaved fairly consistently with respect to how often they selected questions as problematic. 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.

C. Qualitative coding of problems

Although it is useful to determine whether different techniques and organizations produce different number of problems, 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 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.

1. Classification Coding Scheme

The attached Classification Coding Scheme (CCS) was developed to reflect several 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: comprehension and communication, retrieval, judgement and evaluation, and 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 codes be independent of one another and that rules be established on the use of any codes which may be ambiguous. The CCS was developed in order to attempt to maximize inter-rater agreement, with respect to assignment of individual codes[2].

2. Application of CCS Scheme to Questions

The three lead researchers worked together to assign CCS codes to the 15 items receiving the highest total Item Summary Scores, by reviewing the open-ended researcher notes concerning the problems that had been identified through each of the three pretest methods by each of the three organizations. Each item received as many codes as the researchers agreed were found to apply to that item, based on the written comments only[3].

3. Results of coding scheme application

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. A small number of codes accounted for a large proportion of problems identified. Six codes *(Difficult for interviewer to administer, Vague topic/unclear question, Undefined/vague term, Undefined reference period, High detail required/information unavailable, and Erroneous assumption)* accounted for 69.9 percent of all identified problems

Note that all of these codes were classified 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 unclarity dominated their qualitative coding.

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

The data were collapsed according to each of the stages in a four-stage cognitive response model. Due to small cell sizes, the 'judgment and evaluation', and 'response selection' problems were collapsed. 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).

Overall there was no association between the application of codes and Organization (chi-sq(4) =4.384, p<.357).

V. Discussion

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

1. Comparison of techniques. In this study, the Question Appraisal System was the most "productive" in identifying question problems. Given the high frequency with which this technique detected problems, it is very possible that the method, as we applied it, encouraged a low threshold for problem identification, producing a large number of false positives results. We suspect that the QAS method, as used, has high sensitivity but poor specificity. The finding of vastly greater total problems in the QAS is an ambiguous one, however. It could be an artifact of the analysis procedures

[2] Note that the lowest-level CCS codes are 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.

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

used. For current purposes, a question was scored as problematic by the QAS if it failed to "pass" any of 26 separate tests. If one were to establish a higher threshold, based on either total number of problems found or an index weighted by the anticipated severity of certain types of problems, the results might look very different. In any case, these results do seem to support 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 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 studies (Presser and Blair, 1994; Willis 2001) used questions with less severe flaws. 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 replicated 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 be responsible for the modesty of the obtained relationships, and mask a greater degree of agreement across organizations.

B. Assignment of type of problem: Qualitative Analysis

Examination of the types of problems found at the most general, cognitive processing model level 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 methods contributes somewhat differently to the identification of problems in 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 most 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 be consistency in the ways that the techniques are being used. The relative lack of consistency across organizations in choosing *which* items were problematic is somewhat troubling, although it could be argued that there was little disagreement about which 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? An extension of the current study is now being undertaken to address these research questions.

For those persons interested in the details of this research, please request the full-length version of the paper from the contact author.

NOTE: This paper reports the results of research and analysis undertaken by Census Bureau staff and colleagues. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

II. 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*. 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.

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.*

CLASSIFICATION CODING SCHEME

COMPREHENSION AND COMMUNICATION

Interviewer Difficulties

1. Inaccurate instructions (move to wrong place; kip error)
2. Complicated instructions
3. Difficult for interviewer to administer

Question Content

4. Vague topic/unclear Q
5. Complex topic
6. Topic carried over from earlier question
7. Undefined term(s)/vague term

Question Structure

8. Transition needed
9. Unclear respondent instruction
10. Question too long
11. Complex or awkward syntax
12. Erroneous assumption
13. Several questions

Reference Period

14. Reference period carried over from earlier question
15. Undefined reference period
16. Unanchored or rolling reference period

RETRIEVE FROM MEMORY

17. Shortage of memory cues
18. High detail required or information unavailable
19. Long recall period or long reference period

JUDGEMENT AND EVALUATION

20. Complex estimation, difficult mental arithmetic required; (Guessing or heuristic estimation may be likely)
21. Potentially sensitive or desirability bias

RESPONSE SELECTION

Response Terminology

22. Undefined term(s)
23. Vague term(s)

Response Units

24. Responses use wrong or mismatching units
25. Unclear to R what response options are

Response Structure

26. Overlapping categories
27. Missing response categories