Examining expert reviews as a pretest method

DeMaio, Terry; Landreth, Ashley

EXAMINING EXPERT REVIEWS AS A PRETEST METHOD[1]

TERRY DEMAIO &ASHLEY LANDRETH

Introduction

Expert reviews are frequently used as a method of evaluating draft questionnaires. Either alone or in combination with other methods, people who have theoretical questionnaire knowledge or practical experience are asked to review draft questionnaires with an eye to identifying questionnaire problems. This can be done either by having individuals review the questionnaire alone or convening a group, also known as an "expert panel."

Presser/Blair (1994) included expert panels in their research on the effectiveness and reliability of different methods of pretesting questionnaires. But to our knowledge no work has been done to evaluate the consistency of the results produced by individual expert reviewers. We believe it is important to look at the results of individual expert reviews, because we suspect that time and resource constraints cause individual reviewers to conduct the bulk of expert reviews in the early stages of pretesting.

As part of an experiment on alternative cognitive interviewing methods, we used the results from individual expert reviews to gauge the breadth of results produced by three different teams of cognitive interviewers. Rather than having the perspective of one expert represent the potential for problems contained in the questionnaire, we chose to spread the responsibility by recruiting three experts, who worked separately to review the same questionnaire[2] using the same set of instructions.[3]

---

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 are not necessarily those of the U.S. Census Bureau.

2 The questionnaire was a pre-existing CATI general population survey on recycling containing 48 items. Its objectives included determining trash removal practices among households, determining the level of participation in recycling among households, eliciting attitudes about

We expected to find a reasonable level of agreement among the experts in their evaluations of the questionnaire. But we were somewhat surprised at the outcome. In this paper, we lay out our findings and discuss what they might mean for questionnaire development and pretesting.

Methods

Three survey methodologists, from three different Federal government agencies, were enlisted to conduct the expert reviews. Each reviewer had 10 or more years of experience in questionnaire design, cognitive interview methods, and/or survey interview process research and were selected for their ability to complete the task in the time required. Each expert was asked to enumerate problems question by question in a 48-question survey on recycling. They were also asked to identify the five worst questions in the questionnaire, the five worst (i.e. most major) problems with the questionnaire, and the question numbers that reflected those problems. The experts reported their review on paper forms that were provided to them (see Attachment A for a sample of the report forms). The forms were then coded by applying a questionnaire appraisal coding scheme containing 28 problem types (see Attachment B).[4] Problem types and their locations were recorded in a database and compared across the experts.

Results

Table 1 presents the degree to which the experts agreed among themselves in identifying the number and types of problems in the questionnaire. As the top row shows, there are vast differences among the experts regarding the total number of problems each identified, with Evaluator B having identified less than one quarter (17 percent), and Evaluator C having identified less than half (41 percent), of the 158 problems identified by Evaluator A.

---

recycling, and eliciting opinions on alternative recycling strategies designed to increase the level of this behavior.

3 For information about the larger research project, see DeMaio/Landreth (in press).

4 The coding scheme is a close adaptation of that used in recent experimental research on alternative pretesting methods (Rothgeb/Willis/Forsyth, 2001), and was first created by Lessler/Forsyth (1996). The questionnaire problems documented by the expert review forms were coded by both authors, with a good level of inter-coder agreement (76.3 percent).

We examined the level of agreement among experts in identifying specific problems in particular questions, and it was extremely low (21 percent).[5] However, the possibility exists that the experts found similar types of problems (e.g. vague terms such as "recycling" and "household trash") but elected to document them at different points in the questionnaire. The data in Table 1 are consistent with this hypothesis. The percentages of problem types identified by experts at the highest level of aggregation (i.e. the categories labeled *interviewer difficulties, comprehension, retrieval, judgment,* and *response)* rank similarly across experts. The *comprehension* category ranks highest in terms of the percentage of these problem types found by each expert, between 40.8 and 60.0 percent. The *response* category ranks second highest, between 29.2 and 33.6 percent, and the *interviewer difficulties* category ranks third for at least two experts, between 13.9 and 22.2 percent. These three categories contain the majority of the problems identified by the interviewers. Agreement among experts also seems consistent for the lowest ranked categories, *retrieval* and *judgment,* which captured the fewest problem types, ranging from 0 to 7.4 percent.

Table 1 Percent of Problem Types Identified by Experts

Problem Types	Experts A (N = 158)	B (N = 27)	C (N = 65)
Interviewer Difficulties	13.9	22.2	4.6
Comprehension	50.6	40.7	60.0
Retrieval	1.3	0.0	4.6
Judgment	0.6	7.4	1.5
Response	33.6	29.6	29.2
Total %	100	99.9	99.9

However, the coder agreement among problem types does not necessarily support the notion that experts reported the same problems at different points in the questionnaire because it could also be the case that the problems they identified were of the same general type (e.g. comprehension) but focused on different terms in the questions (e.g. "recyclables," "trash," etc.). Our coding was not detailed enough to capture these differences.

5 Agreement statistic was generated by dividing the total occurrences of cases where two or more experts agreed on problem type and location (i.e. question number) by the total number of mutually exclusive problem types across all experts (N = 204).

We looked at these results in another way, focusing on the number of *questions* the experts identified as having at least one problem, rather than the number of problems themselves. The second row of Table 2 shows that there is quite a bit more similarity here, at least between two of the experts. Of course, there are external limits imposed here by the number of questions they had to evaluate. But the lowest number of problem questions identified is 41.3 percent of the highest number (19/46); in contrast, the lowest number of problems identified is only 16.4 percent of the highest number (27/165). In other words, there seemed to be a great deal more disparity across experts when comparing the number of problems each found, while the differences seem far less dramatic when comparing the number of flawed questions they identified.

Table 2 Number of Questionnaire Problems and Flawed Question by Expert

	Experts		
Problems & Flawed Questions	**A**	**B**	**C**
Number of problems found	158	27	65
Number of questions w/problems	46	19	40
Number of questions affected by major problems	38	14	10

Experts were asked to identify the five worst questions in the questionnaire. The variation in identifying problem questions was just as great as identifying individual problems. Only one question was named as the worst question by all three experts. One other question was mentioned by two experts, and all the other "worst questions" were named by only one expert. That is, there were 10 "worst questions" that none of the experts agreed on.

Experts were also asked to identify the five "most important problems"[6] they found with the questionnaire, starting with the most broad ones (i.e. those broad problems that could potentially affect more than one question or aspect of the questionnaire). For each problem, they were instructed to list the question numbers of the items that were most likely to be affected by it. Again, we found great variability in the evaluations of the experts. The magnitude of the problems varied from large things like "the survey wants household level information but the questions ask for person level estimates" to fairly

6 Experts were not provided any criteria for identifying the most important problems. They relied on their own interpretations of this concept.

small things like “the ordering of the scale items.” In terms of agreement across the experts, there was no major problem that was mentioned by all three experts. Three of the major problems had agreement by two of the experts, and six problems were only mentioned by one of the experts. In one case, an expert listed two major problems that were related and both fit under one problem listed by another expert. So, experts agreed less often on these general overarching problems than we would have expected.

Even when the same problems were identified, however, there were large differences in the number of questions that were reported as being affected by the problem. Overall, as the bottom line of Table 2 shows, many more questions were identified by Expert A than Experts B or C as being affected by the major problems they reported. Specifically Expert A reported an average of 7 questions affected by each major problem, while Experts B and C each reported an average of 2.2 problems.

Discussion

An analysis of the output of the experts suggests different review styles are operating. Expert A clearly has a very detailed focus, finding more than twice as many problems as the nearest other expert. This included problems with almost all of the questions (46 out of 48), with each question having an average of 3.4 problems. In addition, the five worst problems enumerated by this expert were reported to affect almost 80 percent of the questions (38 out of 48). A very careful and critical review was necessary to elicit the level of evaluative information contained in this report.

Expert B, in contrast, can be thought of as having a minimalist focus. Relatively few problems were identified by this expert compared to either of the other two. Less than half the questions (19) were seen as having problems, with each question having, on average 1.4 problems. The five worst questionnaire problems identified by this evaluator were fairly narrow in the number of questions they applied to (14). On the basis of this information, one would have a hard time believing that experts A and B were reviewing the same questionnaire.

Expert C has a more middle-of-the-road focus. The number of questionnaire problems identified was in the middle of the other two experts. The number of questions identified as problematic was similar to Expert A, while the number of questions affected by the five worst problems was similar to Expert B. While more questions were found to be problematic, the number of problems identified per question is fairly low (1.6 on average). In fact, there is a lower ratio of problem questions that are affected by a broad

problem (10/40) for Expert C than for either of the others – 38/46 for Expert A and 14/19 for Expert B.

There are several explanations for the disparity of these results. One is the amount of time the experts were able to devote to the task. A priori, one could argue that the more time is allotted, the more comprehensive the review. (However, this information was not collected for this project.) Second is differential expectations about the level of detail required in the assignment. Once an expert identified a problem in question 3, for example, he/she may not have felt it necessary to report it again in the item-by-item portion of the task for every other question that suffered from the same problem. Third is a difference in the perceptions of the experts as to what constitutes a good or bad question. Fourth, although the experts all had 10 or more years of experience in questionnaire design, cognitive interview methods, and/or survey interview process research, their particular experience and expertise may have left them better or worse at evaluating questionnaires. Finally, some experts may be used to working in review panels rather than individually, and feel hampered by the non-collaborative style here.

The low level of agreement among the experts in our research is enough to cause concern about the generalizability of expert review results. While expert reviews are typically considered as a quick, low-cost method of obtaining input about questionnaire problems, some thought should be given to specific aspects of the review procedures. Who does the expert reviews and how they are done may have important implications for the quality of the review. We thought we were providing specific guidelines to our reviewers. But although the reporting format was standardized, the process of problem discovery was not. Some experts may have used a question appraisal scheme to guide their review, while others may have taken a less structured approach. Without more controlled research on this topic, we would suggest that the results of a single expert may not be sufficient, either by itself as a pretest method or as a preliminary step for cognitive interviews. It seems to us that expert review panels (even small ones) would, by their collaborative nature, yield more consistent results. And in addition, some structured procedures such as using a question appraisal scheme to guide their review should be presented to experts. Perhaps further research in this area could determine what the best method of approaching the expert review task would be.

References

DeMaio, T.J./Landreth, A.D., Cognitive Interviews: Do Different Methodologies Produce Different Results?, in S. Presser/J. Rothgeb/M. Couper/J. Lessler/E. Martin/J. Martin/E. Singer (eds.), Questionnaire Development Evaluation and Testing Methods, New York: Wiley Interscience (forthcoming).

Lessler, J./Forsyth, B., 1996: A Coding System for Appraising Questionnaires, in N. Schwarz/S. Sudman (eds.), Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco: Jossey-Bass.

Presser, S./Blair. J., 1994: Survey Pretesting: Do Different Methods Produce Different Results? In P.V. Marsden (ed.), Sociological Methodology: Volume 24, Beverly Hills, CA: Sage, pp. 73-104.

Rothgeb, J./Willis, G./Forsyth, B., 2001: Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results? Paper prepared for presentation at the Annual Meetings of the American Association for Public Opinion Research, May 2001.

Contact

Terry DeMaio
U.S. Census Bureau
SRD/Center for Survey Methods Research
Washington, DC 20233-9100
U.S.A.
email: Theresa.J.DeMaio@census.gov

Ashley Landreth
U.S. Census Bureau
SRD/Center for Survey Methods Research
Washington, DC 20233-9100
U.S.A.
email: ashley.denele.landreth@census.gov

Attachment A:
Selected Pages from Expert Review Report Forms

INSTRUCTIONS

Independent Evaluator Record Sheets

Feel free to use the attached forms to record your analyses. If you prefer to submit typed feedback, please adhere to the general format outlined in the following pages.

PART I: Question-by-Question Problem Identification

1. Question wordings, instructions, and response categories are considered in-scope for this evaluation. For each survey question, identify and briefly explain each specific problem you find. Each problem should be recorded separately and given a number. The attached form only allows for seven (7) problems, but if you find this is insufficient, feel free to continue the numbering scheme to add to the problem list. If additional problems are identified, remember to record all the relevant data associated with these problems – as outlined in items 2 and 3 below.
2. For each problem you identify, mark one box – labeled "H" or "L" – to classify it as a high or low priority problem according to the following definitions:

   High priority: A problem that <u>should</u> be addressed before the instrument is fielded, because it <u>will likely</u> adversely affect the response process in unacceptable ways.

   Low priority: A problem that <u>could</u> be addressed before instrument is fielded, but <u>may not</u> adversely affect the response process in unacceptable ways.
3. For each specific problem identified, mark one box – labeled "A" or "R" or "B" – to record whether it will be a problem with administering the question (A = administration), a response problem (R = response), or both (B = both).

PART II: Five (5) Most Important Problems

1. Briefly state the five (5) most important problems you found with this questionnaire. Please list any broad/general problems first (i.e. those that apply to more than one question or aspect of the questionnaire).
2. For each of the problems you identify, please list the question numbers that are likely to be affected.

PART III: Five (5) Worst Questions

1. Identify the five (5) worst questions. For each, please include a short (i.e. 1-2 sentences) explanation for its selection.

RECORD SHEET EXAMPLES

PART I: Question-by-Question Problem Identification

Q10

	Priority:	Problem for:
Problem 1:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**

This question is double-barreled; it asks respondents to enumerate the number of years since they bought the horse AND moved to Montana.

PART II: Five (5) Most Important Problems

Briefly state the problem and list affected question numbers:

Problem 1:

Awkwardly worded questions will be difficult for respondents to comprehend the first time the question is read.

Affected question numbers: Q9, Q42, Q75, and Q99

PART III: Five (5) Worst Questions

Identify question number and provide brief explanation for why it was selected (1-2 sentences):

Problem 1. Q #: Q76 Explanation:

This question's response set is not mutually exclusive, and it will be impossible for respondents to select only one option – as the question's instruction suggests.

PART I: Question-by-Question Problem Identification

Q1

	Priority:	Problem for:
Problem 1:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 2:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 3:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 4:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 5:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 6:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 7:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**

PART II: Five (5) Most Important Problems

Briefly state the problem and list affected question numbers:

Problem 1:

Affected question numbers: _____

Problem 2:

Affected question numbers: _____

Problem 3:

Affected question numbers: _____

Problem 4:

Affected question numbers: _____

Problem 5:

Affected question numbers: _____

PART III: Five (5) Worst Questions

Identify question number and provide brief explanation for why it was selected (1-2 sentences):

Problem 1. Q #: _____ Explanation:

Problem 2. Q #: _____ Explanation:

Problem 3. Q #: _____ Explanation:

Problem 4. Q #: _____ Explanation:

Problem 5. Q #: _____ Explanation:

Attachment B:
Questionnaire Appraisal Coding Scheme

Interviewer Difficulties	**Comprehension**	**Retrieval**	**Judgment**	**Response Selection**
IR Difficulties	Question Content	Retrieval from Memory	Judgment & Evaluation	Response Terminology
1 Inaccurate instruction 2 Complicated instruction 3 Difficult for interviewer to administer	4 Vague/unclear Q 5 Complex topic 6 Topic carried over from earlier Q 7 Undefined/vague term	17 Shortage of memory cues 18 High detail required or info unavailable 19 Long recall or reference period	20 Complex estimation, difficult mental calculation required 21 Potentially sensitive or desirability bias	22 Undefined term 23 Vague term
	Question Structure			Response Units
	8 Transition needed 9 Unclear respondent instruction 10 Question too long 11 Complex/awkward syntax 12 Erroneous assumption 13 Several questions			24 Responses use wrong or mismatching units 27 Unclear to respondent what response options are 28 Multi-dimensional response set
	Reference Period			Response Structure
	14 Period carried over from earlier Q 15 Undefined period 16 Unanchored/rolling period			25 Overlapping categories 26 Missing response categories