

“A Comparison of Appraisal and Cognitive Interview Methods for Testing Organizational Survey Questionnaires.”

Barbara H. Forsyth (Westat), Elisa S. Weiss (The New York Academy of Medicine) and Rebecca Miller Anderson (Mount Sinai School of Medicine)

Introduction

Today we'll be talking about a study we conducted to compare results from two methods for pretesting organizational survey questionnaires: cognitive interviewing and a questionnaire appraisal tool, the Questionnaire Review Coding System or QRCS (Forsyth et al, 1999). Both pretest methods produce qualitative results and so, we are wealthy with data. Today, we'll give you a general accounting. There's a lot more we can share with anyone who's interested.

We used cognitive interviews and the QRCS appraisal tool to pretest two questionnaires developed for the National Study of Partnership Functioning (NSPF) – a study of factors affecting the success of collaborative processes in partnerships developed to promote health and well-being in local communities.

As a psychologist, one of us has spent way too much time thinking about primacy and recency effects. We'd like to get them both working for us today, so we'll start with the punch line. Our results suggest that the two methods are complementary. They produce similar results, but with different emphases. It seems to us that the differences make the two pretest methods useful for slightly different purposes and particularly effective when used in combination.

Now, we'll turn to the part of this talk that you're less likely to remember. We'll start with a few words about why the methods comparison is interesting to us. Then, we'll give a little background about the survey questionnaires we pretested. We'll spend more time talking about our pretesting methods and results and finish – in time for recency to kick in -- with a few conclusions and observations.

So why is it interesting to compare pretest results based on cognitive interviews and questionnaire appraisals – especially in this organizational survey context? Cognitive interviewing is a pretesting staple. It's a pretty standard method for pretesting household survey questionnaires and, based on other papers presented at this conference, it seems to be on the rise for pretesting establishment and organizational surveys. Questionnaire appraisal is in essence a structured expert review. Appraisal methods are less common for testing household survey questionnaires. Furthermore,

appraisal tools appropriate for organizational surveys are relatively new developments.

Early developmental testing with the QRCS tool indicated that questionnaire revisions made based on cognitive interviewing seemed to address at least some potential problems identified by an independent evaluation using the QRCS (Forsyth et al., 1999). Our aim is to extend that preliminary finding by making more explicit comparisons between the two sets of pretest results.

The NSPF is an intriguing context for this methods comparison because the partnerships studied in the NSPF are inter-organizational collaborations. As a consequence, the organizational factors and social dynamics that affect any organizational survey are particularly rich in the NSPF. We'll show you what we mean by giving you a little more information about the NSPF.

The National Study of Partnership Functioning (NSPF)

The National Study of Partnership Functioning was an exploratory study designed to learn more about the characteristics of partnerships that are related to their ability to create a successful collaborative process. The study was also developed to test the validity and reliability of new measures. The sample of partnerships included 63 partnerships – for example, consortia, coalitions, and alliances – in 28 states across all regions of the U.S. All partnerships had been in existence for two years or more and had at least ten partners working together to improve health and well-being in their communities.

Because the focus of the study was on characteristics of the partnership as a whole, the study design called for collecting data from multiple informants within each partnership. Informants from each partnership included the person who was responsible for coordinating the partnership's activities as well as partner representatives who represented organizations participating in the partnership and who were identified as knowledgeable about the partnership.

The study was designed as a self-administered mail survey. We developed two slightly different questionnaires for the partnership coordinators and partner representatives. Several sets of items were included in both questionnaires, but both questionnaires contained additional items that reflected the different roles and responsibilities for partnership coordinators and partner representatives.

Pretest Methods

We used cognitive interview methods to pretest the two draft questionnaires and we used the results to revise the questionnaires in preparation for the

main study. We conducted the QRCS review roughly two years later. This methods study we're reporting on today is something we talked about off and on during cognitive interviewing and questionnaire revision phases of the NSPF. We took this conference as an opportunity to follow through on those early ideas.

We'll say more about the pretest staffing and time lag in a moment. First, we'll tell you about the pretest methods we used.

Cognitive Interview Pretest Methods. We recruited 11 respondents to participate in the cognitive interview pretest. These respondents included 3 partnership coordinators and 8 partner representatives. Geographic diversity was an important factor in the pretest design. This meant that the cognitive interviews had to be conducted by telephone. We chose to mail survey materials to pretest volunteers and have them complete the questionnaire before their cognitive interview appointments. The cognitive interview protocol used a combination of retrospective think-aloud, paraphrasing and detailed probing methodologies to explore questionnaire design issues, including:

- instructions that might make the survey task easier,
- question wordings that might make items difficult to understand or answer,
- vocabulary that might be specialized or unfamiliar,
- response scales that might be difficult to interpret,
- response sets that might be incomplete,
- questions that might seem overlapping or redundant, and
- important partnership experiences that were not covered by the draft questionnaire items.

Based on detailed notes, we drafted comprehensive summaries for each interview. We used qualitative analytic methods to review the summaries and identify the key findings from the cognitive interview pretest.

Questionnaire Appraisal Pretest Method. We used the QRCS organizational survey appraisal tool to review the two draft questionnaires. There are a variety of questionnaire appraisal methods available. Most of them are designed for pretesting household survey questionnaires (e.g., Forsyth et al., 1992; Lessler & Forsyth, 1996; Rothgeb et al., 2001; Willis & Lessler, 1999). Fewer of the tools are designed for organizational survey questionnaires (e.g., Forsyth et al., 1999; O'Brien, 2000).

All of these appraisal systems are intended as methods for conducting a structured expert review. Under the general appraisal method, experts

review a questionnaire item-by item. For each item, reviewers identify question characteristics that could cause problems for respondents or interviewers. Reviewers document problematic question characteristics using a set of question codes. For example, most appraisal systems include a code for “undefined terminology” that can be assigned to items that use technical terms without defining them. As another example, most appraisal systems include a code for “social desirability” that can be assigned to items that are likely to elicit responses that are biased in the direction of presenting the respondent or the respondent’s organization in a favorable light.¹

We don’t have the time needed to fully review the appraisal method we used. Instead, we are making available copies of a 1999 paper describing the QRCS for anyone interested in more detail. (Copies are at the back of the room.) Also, your handouts contain a copy of the QRCS codes that we used to review the draft coordinator and partner representative questionnaires. (See Exhibit 1 attached.)

Pretest Timing and Staffing. As I said, we used the QRCS to evaluate the two draft questionnaires roughly two years after we completed the cognitive interview pretesting and questionnaire revision phases of the NSPF. Obviously the two pretests were not independent of each other. I was the lead cognitive interviewer and I was also the lead QRCS reviewer. I played a minor role in revising the original draft questionnaires for the main study, and I was not involved at all in the main study data collection. I had plenty of distractions between the fall of 1999 and the spring of 2002. While the two sets of pretests are not independent, we believe it is also true that there is less carry-over across the methods than there would be under an iterative pretest design that used cognitive interview results to inform the QRCS review. We’re somewhere in-between independence and staged implementation.

Results

Let’s move on to discuss results from our comparisons of the two pretest methods. We’ll begin by sharing findings from the QRCS and then move to the cognitive interview findings. We’ll discuss the cognitive interview findings in relation to the QRCS. We’ll follow with our conclusions.

QRCS Results. We computed the frequencies with which individual QRCS codes were assigned in order to get an overview of the appraisal results. A few findings are evident based on the code frequencies.

¹ The set of QRCS codes was developed to identify questionnaire problems. Since problems result from interactions between respondents and questionnaires (and sometimes interviewers too), “problems” identified by the QRCS can turn out to be merely question “characteristics” under actual survey implementation.

Finding #1: The two questionnaires are very similar in terms of the number of potential problems identified by the QRCS.

Exhibit 2 shows frequency distributions for the numbers of codes assigned to individual items. The average number of codes per item was 4.6 for the coordinator questionnaire and 5.4 for the representative questionnaire, and the standard deviations for the number of codes assigned were also very similar: 2.48 for the coordinator questionnaire and 2.37 for the representative questionnaire. In other words, the two distributions in Exhibit 2 are very similar.

The actual numbers of codes assigned is difficult to interpret. What's the range that indicates "small problems" with the questionnaire? How many codes are needed to signal "big problems?" For our purposes, the numbers of codes assigned are most helpful for understanding how the two questionnaires compare. These two questionnaires seem very similar in terms of the overall magnitude of potential problems identified by the QRCS review.

Finding #2: The two questionnaires are also very similar in terms of the general types of potential problems identified by the QRCS.

As you can see from the QRCS in your handout (Exhibit 1), the codes fall into five general problem areas:

- item instructions
- question comprehension
- information retrieval
- synthesis, evaluation and judgment, and
- response selection.

Exhibit 3 shows the number of items having one or more potential problems within each of the five general problem areas. The first set of frequencies in Exhibit 3 shows results for the coordinator questionnaire and the second set of frequencies in Exhibit 3 shows results for the representative questionnaire.

Comprehension-related problems were the most frequent in both questionnaires. Most of these potential problems were linked to the use of potentially vague terminology and undefined reference periods.

Retrieval- and judgment-related problems were also relatively frequent in both questionnaires. Most of these potential problems were due to

- potential difficulties accessing relevant organization or partnership records
- likely use of guessing strategies, and
- items involving potentially complex estimation.

Potential problems with response selection were somewhat less frequent in both questionnaires. Again, the use of vague terminology is particularly common – in this case to identify response options.

Finding #3: The QRCS revealed some small differences between the two questionnaires in the areas of judgment and response selection. In these problem areas, the QRCS review identified more potential problems for the representative questionnaire than for the coordinator questionnaire. The higher number of judgment-related codes in the representative questionnaire was due to three sets of items that were coded for potential "social desirability" bias. These items asked representatives to report on experiences that could reflect more or less favorably on their organizations and on the partnerships they were involved in.

The higher number of response-related codes in the representative questionnaire was due to two sets of items that were coded as having both "overlapping response categories" and "missing response categories." These codes were linked to the particular formatting selected for them.

Cognitive Interview Results. Let's turn to the cognitive interview results. Are they similar or different? It turns out that the answer to this question is "Yes." Yes they're similar and yes, they're different. We'll try to highlight some of the similarities and differences with a few more findings.

Finding #4: There were several points of agreement between the QRCS and the cognitive interview results. For example, the QRCS review identified very few instruction-related problems. Cognitive interview respondents agreed, reporting that the questionnaire instructions were clear and easy to follow.

Cognitive interview respondents identified questionnaire terminology that confused them, and most of these terms were also identified as potentially problematic by the QRCS review. For example, cognitive interview participants reported that the questionnaire seemed to use the term "partner" in different ways – sometimes referring to partner organizations participating in a partnership and sometimes referring to the individual partner representatives who work on behalf of those organizations. The QRCS review also identified the "partner" terminology as potentially problematic.

In results related to question structure, cognitive interview respondents reported that they were confused about the reporting unit for several sets of questionnaire items. They wondered whether they should be reporting their own experiences and perspectives or the experiences and perspectives of the organizations they represent. In many cases, these concerns raised by cognitive interview participants were also identified by the QRCS review, typically through the QRCS code for "unclear question goal."

There were other points of agreement between the two sets of results that we can talk about more informally if anyone is interested. But, the main point is that the two methods had many similar conclusions.

Finding #5. In those areas of agreement between the QRCS and cognitive interview results, there was value added from talking with respondents in the cognitive interviews. Often respondents articulated the problems they had understanding or answering an item. Most importantly, respondents were able to suggest solutions that made sense to them; this is something that QRCS tool does not offer. For example, cognitive interview participants generally thought we could clarify the ambiguous “partner” terminology by reserving the term “partner” to refer to partner organizations, and explicitly referring to individuals as “partner representatives.” We followed this advice in revising the questionnaire and also added instructional passages to make the distinction explicit. This consistent and clearly defined vocabulary also helped to address confusion that cognitive interview participants reported about the intended reporting unit. Revised items focusing on individuals used different vocabulary from revised items focusing on organizations.

Finding #6: Comments from the cognitive interview participants addressed a few important topics that were beyond the reach of the QRCS review. For example, several participants commented on their high interest in the topics covered by the questionnaires – something an expert reviewer can only dream of. They also mentioned that the questionnaire was easy to complete – low burden. Expert reviewers can make informed guesses about burden, but verification from cognitive interview participants is a valuable asset. (Response rates are even more valuable for assessing burden. Among other things.)

Although it’s difficult to prove without a careful review of problems identified during data collection, we believe there is also value added from the QRCS results. We believe the QRCS effectively identified more subtle problems in questionnaire design that cognitive interview respondents may not be attentive to. I’ll return to this point when I summarize our conclusions.

Finding #7: In a few instances, cognitive interview participants and the QRCS review came to different conclusions about particular questions. Three examples seem particularly important to us.

First, a few cognitive interview participants reported that some items in the questionnaire led them to think about their partnership experiences in ways they hadn’t before. Cognitive interview participants saw this as a positive feature that made the questionnaire items appealing to them. However, several of these items were identified as potentially problematic by the QRCS review because the items made “implicit assumptions” about respondents’ partnership experiences.

Implicit assumptions can interfere with comprehension when the assumptions built into a question don’t match a respondent’s experience. Well-motivated respondents might notice and be interested by the disparity. Less motivated respondents could be less aware and consequently confused

or frustrated by it. Additional evaluation is required to resolve this apparent divergence.

As a second example, most cognitive interview respondents made favorable comments about the general questionnaire format. Many of the items were formatted as items in series -- with a single question stem and multiple items, all set up using a common response scale. Other authors have commented on the benefits and risks of the item-in-series format (e.g., Dillman, 2000).

Our cognitive interview respondents commented on the benefits of the item-in-series format. It helped them move through the questionnaire quickly. The QRCS review identified some of the potential risks. Notably, the question is distant from items placed late in the series, making it easy for respondents to lose track of the intended question as they move through the list of items in the series. Again, further evaluation is required to determine whether the series were a suitable length to take advantage of the format's benefits without introducing comprehension errors.

Our third example focuses on potential social desirability biases. Recall that the QRCS results suggested risks of desirability reporting in both questionnaires. The QRCS suggested a stronger risk of desirability reporting in the representative questionnaire due to sets of items that asked representatives to report on experiences that could reflect more or less favorably on their organizations and on the partnerships they were involved in.

In contrast, only one cognitive interview participant reported feeling uncomfortable answering any of the questionnaire items and that one respondent was reacting to the coordinator survey questionnaire. While the QRCS results suggested relatively strong risks of desirability reporting, the cognitive interview results suggested little risk. As before, further evaluation is required to determine the actual prevalence of response bias in the direction of social desirability.

We've highlighted some similarities between and differences across the QRCS and cognitive interview results. We'll close by summarizing a few things we think we've learned.

Conclusions

Well then, what have we learned?

First, there were similarities across the two sets of pretest results -- particularly in the general areas of instruction and question comprehension. These similarities provide some convergent support for both pretest methods. This convergence is particularly important to us because the QRCS appraisal tool is relatively new and largely untested. Our results suggest that

the QRCS tool works reasonably well as a pretest method, compared with more traditional cognitive interviewing.

Of course, there were also differences across the two sets of pretest results. The cognitive interviews provided information about levels of interest and perceived burden that we wouldn't trust expert reviewers to be able to assess. In addition, the cognitive interview results suggested solutions for problems identified.

We also believe that the QRCS appraisal method provides insights that are not likely to be available from cognitive interviewing alone. For example, insights into factors like social desirability that influence judgment and response selection. Social desirability may operate in ways respondents are not aware of, making cognitive interviews an ineffective approach for identifying desirability effects.

Following similar logic, we expect there are other areas, besides social desirability, where the QRCS tool provides insights we would not have access to based on cognitive interviewing alone. For example, if it doesn't occur to respondents that there may be organizational records or other persons in their organization who they should talk with in order to answer a question, then difficulties accessing records or other organizational resources may not be evident from cognitive interview responses. Similarly, if cognitive interview respondents believe they understand a question, then there's a risk that misunderstood terms or misinterpreted reference periods will not be evident in cognitive interview responses.

We anticipate that the role of awareness differs depending on the particular cognitive interviewing methods used. Our design and budget required that we adopt a retrospective approach that may be particularly insensitive to more subtle factors affecting response processes. However, the general concern holds for all pretest methods that rely on self-report (e.g., Forsyth & Lessler, 1991).

So what do we think after using both pretest methods? We believe that expert reviewers and pretest respondents have complementary types of expertise. We hypothesize that each makes useful contributions to pretest findings. Also, we hypothesize that using a mix of pretest methods that take advantage of both is likely to be more effective than relying on either one alone. In other words, we recommend using the QRCS and cognitive interviewing together for best results.

Using the NSPF as an example, cognitive interviewing found areas in the questionnaires that needed improvement. Based on results from the main study, we believe that the QRCS identified additional problem areas that were not detected based on the cognitive interview results. If we had conducted the QRCS review first and used the results to develop the cognitive interview protocol, we could have focused more closely on some of

the additional areas flagged by the QRCS. Perhaps some problems we found in the main study could have been reduced if we'd used both pretest methods before launching the main study. In general, we believe that using the two methods together will produce a better questionnaire.

Exhibit 1. Questionnaire Review Coding System (QRCS) for the National Study of Partnership Functioning (NSPF)

Instructions	Comprehension	Information Retrieval	Synthesis and Judgment	Response Selection
Instruction Content	Question Terminology	Organization Characteristics	Match: Record and Item	Response Terminology
Conflicting instructions Inaccurate instructions Hidden instructions Complicated content Complex syntax Separate from item Nearby but not embedded in item Instructions provided too late Unclear examples Unclear layout Transition needed Provide info on finding details	Critical definition(s) missing Add or add to examples Ambiguous or vague term(s) Multiple definitions Mismatch to technical language Industry-specific terminology	Distributed knowledge likely (or multiple sources) Seasonal or periodic trends	Incompatible with regulatory requirements Incompatible with organizational or unit objectives Survey-specific system unlikely (e.g., panel versus one-time survey) Variability in recorded units Incompatible time frames	Critical definition(s) missing Vague term(s) Mismatch to technical language Industry-specific terminology
	Question Structure	Source Identification		Response Units
	Hidden question Complex syntax Implicit assumption Several questions Unclear goal Q/A mismatch No question	Provide assistance to help identify source(s) Sources may not be accessible (for survey purposes)		Response unit mismatch to organizational units Responses use wrong units
	Reference Period	Memory Retrieval	Judgment Processes	Response Structure
	Carry-over reference period Undefined reference period Embedded reference period Abrupt change Problematic length	Non-routine summary or breakdown required Shortage of (memory) cues Detail problem/item specificity Unanchored reference period Rolling reference period	Coordination or collaboration Guessing or estimation likely	Overlapping categories Missing response categories
Navigational Instructions		Record Retrieval	Task Characteristics	
Inaccurate instructions (move to wrong place) Confusing convention, flow or typographic Complex information Not salient		Records unavailable or don't support estimation Record access issues Authority issues	Non-routine time frame Complex estimation Potentially sensitive Social desirability Proprietary information Strategic factors Timing issues	
Question Content				
Complex topic Under specified topic Topic carried over Assumes consistent behavior Provide assistance to identify source(s)				

Exhibit 2. Frequency distribution for number of QRCS codes assigned to individual items, by questionnaire.

Number of Codes Assigned	Coordinator Questionnaire		Representative Questionnaire	
	Number of items	Percent of items	Number of items	Percent of items
0	4	4.0%	5	3.9%
1	9	9.1%	8	6.2%
2	9	9.1%	4	3.1%
3	12	12.1%	9	7.0%
4	15	15.1%	12	9.3%
5	15	15.1%	19	14.7%
6	11	11.1%	29	22.5%
7	9	9.1%	20	15.5%
8	9	9.1%	13	10.1%
9	5	5.0%	8	6.2%
10	1	1.0%	2	1.6%
TOTAL	99		129	

Exhibit 3. Number of Items coded as having one or more potential problems within each of five general problem areas.

		Coordinator Questionnaire		Representative Questionnaire	
		Frequency	Percent of items	Frequency	Percent of items
General Problem Area					
	Instructions	12	12%	10	8%
	Comprehension	86	87%	117	91%
	Retrieval	69	70%	96	74%
	Judgment	55	56%	93	72%
	Response	42	42%	82	64%