Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 531–535

Discussion

Jennifer H. Madans[1] *and Paul C. Beatty*[2]

Over the last several decades, questionnaire evaluation has become an increasingly prominent component of methodological work aimed at maximizing the quality of survey data. Question evaluation methods are among the most important tools survey methodologists have for describing and improving data quality, but these methods generally require a great deal of investigator interpretation, which makes them difficult to use. The statistical methods available to evaluate sample errors are straightforward when compared to the methods used to evaluate survey questions. Questionnaire evaluation methods also range from the most qualitative to the most quantitative, and require a wide range of expertise.

The questionnaire evaluation methods reviewed in the article by Yan, Kreuter, and Tourangeau, considered together, show the breadth of efforts in this area. As they note, the methods vary widely in their assumptions, implementation, and nature of the data that they produce–and indeed, very different sets of knowledge and skills would be needed to utilize them. For example, expert review presumably requires extensive knowledge of questionnaire design literature, experience crafting questions, and ability to make qualitative judgments; latent class analysis and structural equation models require little of these, but require sophisticated and specific quantitative skills. Given the broad differences in these approaches, and the fact that few survey methodologists are likely to be proficient in all of them, attempts to compare them and evaluate their respective contributions are most welcome. Such comparisons have the potential not only to expand and improve the application of these methods, but also to serve as an impetus for further methodological research. Yan, Kreuter and Tourangeau provide a very useful overview of prominent questionnaire evaluation methods, but the article also illustrates just how difficult it is to compare these methods. The way that specific techniques are used, and the way that results are summarized and combined, can greatly affect any conclusions about the quality of the questions and the data that are produced. Question evaluation methods provide crucial information on data quality, both to improve the quality of data collections and to inform users of existing data. However, they can only be effective if used in a way that provides credible evidence that itself can be evaluated by data users and question developers. This requires that the methods are described and used in a manner that is as transparent as possible.

Comparative methodological evaluation studies such as this one run the risk of oversimplifying the purpose of the methods. In this study and others, there is an implicit

[1] National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20,782, U.S.A. Email: jhm4@cdc.gov
[2] National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20,782, U.S.A. Email: pbb5@cdc.gov

dichotomy between questions that work, and those that have problems that need to be addressed. Presumably, the effectiveness of a method is linked to its ability to identify these problems; methods that identify more problems (or at least, more genuine problems) are considered to be more effective. On the face of things, this does not seem like a particularly controversial set of assumptions. Clearly, questionnaire evaluation often identifies various question flaws, such as ambiguous terms, inappropriate response categories, or overly challenging response tasks, and questions are rewritten to eliminate the problems.

But we suggest that the process of questionnaire evaluation is often more complicated than finding and fixing problems, a paradigm which suggests that there is an "ideal" way to ask a question. The art of question design involves obtaining information about complex concepts through a very limited interaction with a respondent, using questions that the respondent might or might not be paying close attention to. As a result, every question has some degree of imperfection–for example, ambiguity in the way some concept is described. A revision may reduce this ambiguity through adding clarifying details, but these details may add confusion for respondents who would not have had trouble with the original question. Similarly, a term may be problematic for some respondents, and an alternative might be simpler for them, but lack specificity that others need. Questionnaire evaluations should identify the strengths and weaknesses of particular ways of asking a question, and helps conscientious social scientists understand the tradeoffs involved in the various alternatives they could select. Question evaluation should move toward this paradigm to optimize the chances that the appropriate information will be captured. As it is not possible to design questions that mean the same thing to all respondents, or to tap the exact concepts that the researcher desires, evaluation techniques must not only identify problems, but must also provide information to users about what the question means to respondents. Hopefully this will maximize the likelihood that the question will obtain the information desired by its author.

Unfortunately, attempts to quantify this sort of contribution generally fail to capture the nuances of how evaluation methods help to make questionnaire design decisions. Understandably, researchers conducting such evaluations rely on what they can actually measure, such as counts of problems. But quantifying problems is not a very useful metric. For one thing, it requires an operational definition of a problem. Counts also generally assume that problems are of equal weight–but clearly some problems are minor imperfections, while others threaten the usefulness of any data generated by the question. Perhaps more importantly, quantifying problems fails to capture the level of insight produced by various methods. Yan, Kreuter and Tourangeau themselves note that researchers commonly suggest that the main value of cognitive interviewing is that it produces qualitative insight into the fit between the question and the concept it is trying to measure (cf., Beatty and Willis 2007; Miller 2011). Yet many methodological studies, including the current one, evaluate cognitive interviewing in terms of whether it flags the presence of "a problem."

In our view, reducing the output of qualitative methods such as cognitive interviewing in this manner is not only artificial, but undervalues their potential contributions. Similarly, expert reviews may provide rich assessments of which characteristics of questions are likely to lead to particular errors, and here, such insights are only summarized as simple quality ratings. Admittedly, it is hard to quantify the insights gleaned from such methods in a way

that makes them amenable to comparative research. Still, it is problematic to criticize the value of these approaches when the contributions have been reduced to a few variables that do not really represent their contributions to the questionnaire design process.

Evaluating methods that produce qualitative insights is difficult for other reasons as well. Cognitive interviewing, in particular, is practiced in a variety of forms and with varying degrees of expertise. For example, some variants place strong emphasis on "thinking aloud" with minimal interviewer intervention, while others rely heavily upon probing – sometimes prescripted, sometimes determined based on interviewer content. But inevitably, methodological studies must define the practice of cognitive interviewing in a particular way, and the evaluation can only really address the way that it is conducted at that time. In other words, results of an evaluation don't generalize to "the method" – only the way the method was carried out in the particular study. More generally, it is difficult to perform evaluation of qualitative methods because comparisons require that the methods be standardized to some degree – otherwise, it is impossible to specify what exactly is being evaluated. However, this is problematic if one believes that a key strength of the method lies in its ability to adapt to issues that emerge in an interview in ways that would be difficult to predict in advance – in other words, its non-standardization. By standardizing the method, the researcher has compromised its strength and introduced a high degree of artificiality. While the authors are transparent about the approach and assumptions taken in the cognitive interviews within their study, it is very difficult to say anything conclusive about the overall value of "cognitive interviewing" as a method because the method, researchers and particular questions can all be confounded in challenging ways. Hopefully, the development and adoption of best practices for conducting cognitive interviews and for analyzing and reporting results will greatly facilitate question evaluation.

For any evaluation method to be effective, it should produce *measurably better questions*. Insights that do not actually contribute to that goal may be interesting, but are ultimately irrelevant. Evidence regarding the quality of these insights should be generated through carefully designed studies that use appropriate techniques. Question validity is often used as the gold standard for comparing the results of various evaluation methods. Theoretically, methods that produce more *valid* questions would be demonstrably better than alternatives.

The problem is that in practice, true validity is unknown, and attempts to quantify it have numerous problems of their own. In many cases there really is no "gold standard" for comparison – and even if there is, obtaining it is often either difficult or expensive. Furthermore, while latent class analysis is useful for some things, it does not truly measure validity. An alternative is to use the more limited concept of *construct validity*, in which researchers examine correlations with items that should theoretically be related to the question. Unfortunately, such validity assessments are only as good as the external comparators used, which might not be tapping the intended concept. More importantly, being correlated with another item is not the same as actually measuring what is intended. Statements that questions have been "validated" are powerful, but must be used with great caution. Measures of validity need to be improved, and evaluations of validity should report findings in a way that the criteria used to measure validity are clearly defined.

For all of these reasons, we do not find it particularly surprising that the methods evaluated in this study did not produce the same results. The differences are partially attributable to the fact that methods naturally create different sorts of insights, which

cannot be easily compared. They are also partially attributable to the fact that questions are not easy to rank in terms of quality, nor readily categorized as "good" or "flawed" – realistically, most questions are imperfect and multifaceted, better for some purposes than others. Furthermore, they are partially attributable to the fact that standardizing methods, and abstracting results into scoring measures, alters both the methods and the results that they produce. Although the authors stop short of concluding that any of the methods are "better" in an absolute sense, they do suggest that qualitative methods have more to prove than quantitative measures of reliability and validity. We suggest that such conclusions are in part based on assumptions about the comparability of measures that are difficult to support. In fact, we wonder whether the question "which approach is best?" is really the right one to ask. It is an advantage that the methods provide different types of information, as this provides richer evaluation.

Instead, it is very important to ask "what does each method contribute?" and "under which circumstances is each method likely to be useful?" Yan, Kreuter and Tourangeau's article offers a helpful response to the first question through a solid review of the variety of evaluation methods currently available. Their analysis did not really address the second question, but could have through a different approach: rather than attempting to determine overall measures of the quality of each method, and thereby suggesting varying degrees of methodological value, they could have started with the assumption that each method was *likely* to produce different results. From there, it would be possible to examine the nature of evidence from each method, how each are used to draw conclusions, and what sorts of decisions are actually made as a result of each. Such an analysis would not need to assume that all of the methods produced results of equal quality – in fact, it could still conclude that methods produced results of limited worth, at least within the current study. But it would probably not lend itself to conclusions about the relative value of each method. Then again, we find such conclusions to be limited, for the various reasons described above. There will always be interest in combining findings using different methods, and in learning about questionnaire design in general from all methods. As findings will be very dependent on the methods used, those who undertake question evaluations need to be explicit about how tests are done and how evidence from the tests is summarized and evaluated. A lack of information on question behavior is major threat to data quality, but acting on information that does not accurately convey what is known is even more dangerous.

Whether subsequent researchers build from the approach taken by Yan, Kreuter and Tourangeau, or instead decide to pursue different strategies, we think it is useful for their analysis to be part of a larger discussion. Hopefully our reservations with their approach and findings serve a similar purpose and will be seen in that light. Question evaluation remains a vital component of survey quality, and it is clear that we do not yet know enough about the contributions of the various approaches that are available. It is certainly undesirable for methodologists to work without more knowledge about what these approaches do and do not accomplish, although it is also undesirable to draw unwarranted conclusions about their merits. As we have seen, comparative research is difficult, and we have much more work to do before definitive statements can be made about what each method produces and when it should be used.

References

Beatty, P. and Willis, G.B. (2007). The Practice of Cognitive Interviewing. Public Opinion Quarterly, 71, 287–311.

Miller, K. (2011). Cognitive Interviewing. In Question Evaluation Methods, J. Madans, K. Miller, A. Maitland, and G. Willis (eds). Hoboken, NJ: John Wiley & Sons.

Received October 2012