1 Introduction

JENNIFER MADANS, KRISTEN MILLER,
and AARON MAITLAND
National Center for Health Statistics

GORDON WILLIS
National Cancer Institute

If data are to be used to inform the development and evaluation of policies and programs, they must be viewed as credible, unbiased, and reliable. Legislative frameworks that protect the independence of the federal statistical system and codes of conduct that address the ethical aspects of data collection are crucial for maintaining confidence in the resulting information. Equally important, however, is the ability to demonstrate the quality of the data, and this requires that standards and evaluation criteria be accessible to and endorsed by data producers and users. It is also necessary that the results of quality evaluations based on these standards and criteria be made public. Evaluation results not only provide the user with the critical information needed to determine whether a data source is appropriate for a given objective but can also be used to improve collection methods in general and in specific areas. This will only happen if there is agreement in how information on data quality is obtained and presented. In November 2009, a workshop on Question Evaluation Methods (QEM) was held at the National Center for Health Statistics in Hyattsville, Maryland. The objective of the workshop was to advance the development and use of methods to evaluate questions used on surveys and censuses. This book contains the papers presented at that workshop.

To evaluate data quality it is necessary to address the design of the sample, including how that design was carried out, as well as the measurement characteristics of the estimates derived from the data. Quality indicators related to the sample are well developed and accepted. There are also best practices

Question Evaluation Methods: *Contributing to the Science of Data Quality,* First Edition.
Edited by Jennifer Madans, Kristen Miller, Aaron Maitland, Gordon Willis.

for reporting these indicators. In the case of surveys based on probability samples, the response rate is the most accepted and reported quality indicator. While recent research has questioned the overreliance on the response rate as an indicator of sample bias, the science base for evaluating sample quality is well developed and, for the most part, information on response rates is routinely provided according to agreed-upon methods. The same cannot be said for the quality of the survey content.

Content is generally evaluated according to the reliability and validity of the measures derived from the data. Quality standards for reliability, while generally available, are not often implemented due to the cost of conducting the necessary data collection. While there has been considerable conceptual work regarding the measurement of validity, translating the concepts into measurable standards has been challenging. There is a need for a critical and creative approach to evaluating the quality of the questions used on surveys and censuses. The survey research community has been developing new methodologies to address this need for question evaluation, and the QEM Workshop showcased this work. Since each evaluation method addresses a different aspect of quality, the methods should be used together. Some methods are good at determining that a problem exists while others are better at determining what the problem actually is, and others contribute by addressing what the impact of the problem will be on survey estimates and the interpretation of those estimates. Important synergies can be obtained if evaluations are planned to include more than one method and if each method builds on the strength of the others. To fully evaluate question quality, it will be necessary to incorporate as many of these methods as possible into evaluation plans. Quality standards addressing how the method should be conducted and how the results are to be reported will need to be developed for each method. This will require careful planning, and commitments must be made at the onset of data collection projects with appropriate funding made available. Evaluations cannot be an afterthought but must be an integral part of data collections.

The most direct use of the results of question evaluations is to improve a targeted data collection. The results can and should be included in the documentation for that data collection so that users will have a better understanding of the magnitude and type of measurement error characterizing the resulting data. This information is needed to determine if a data set is fit for an analytic purpose and to inform the interpretation of results of analysis based on the data. A less common but equally if not more important use is to contribute to the body of knowledge about the specific topic that the question deals with as well as more general guidelines for question development. The results of question evaluations are not only the end product of the questionnaire design stage but should also be considered as data which can be analyzed to address generic issues of question design. For this to be the case, the results need to be made available for analysis to the wider research community, and this requires that there be a place where the results can be easily accessed.

A mechanism is being developed to make question test results available to the wider research community. Q-Bank is an online database that houses science-based reports that evaluate survey questions. Question evaluation reports can be accessed by searching for specific questions that have been evaluated. They can also be accessed by searching question topic, key word, or survey title. (For more information, see http://www.cdc.gov/qbank.) Q-Bank was first developed to provide a mechanism for sharing cognitive test results. Historically, cognitive test findings have not been accessible outside of the organization sponsoring the test and sometimes not even shared within the organization. This resulted in lost knowledge and wasted resources as the same questions were tested repeatedly as if no tests had been done. Lack of access to test results also contributed to a lack of transparency and accountability in data quality evaluations. Q-Bank is not a database of good questions but is a database of test results that empowers data users to be able to evaluate the quality of the information for their own uses. Having the results of evaluations in a central repository can also improve the quality of the evaluations themselves, resulting in the development of a true science of question evaluation. The plan is for Q-Bank to expand beyond cognitive test results to include the results of all question evaluation methods addressed in the workshop.

The QEM workshop provided a forum for comparing question evaluation methods, including behavior coding, cognitive interviewing, field-based data studies, item response theory modeling, latent class analysis, and split-sample experiments. The organizers wanted to engage in an interdisciplinary and cross-method discussion of each method, focusing specifically on each method's strengths, weaknesses, and underlying assumptions. A primary paper followed by two response papers outlined key aspects of a method. This was followed by an in-depth discussion among workgroup participants. Because the primary focus for the workgroup was to actively compare methods, each primary author was asked to address the following topics:

- Description of the method
- How it is generally used and in what circumstances it is selected
- The types of data it produces and how these are analyzed
- How findings are documented
- The theoretical or epistemological assumptions underlying use of the method
- The type of knowledge or insight that the method can give regarding questionnaire functioning
- How problems in questions or sources of response error are characterized
- Ways in which the method might be misused or incorrectly conducted
- The capacity of the method for use in comparative studies, such as multicultural or cross-national evaluations

- How other methods best work in tandem with this method or within a mixed-method design
- Recommendations: Standards that should set as criteria for inclusion of results of this method within Q-Bank

Finally, closing remarks, which were presented by Norman Bradburn, Jennifer Madans, and Robert Groves, reflected on common themes across the papers and the ensuing discussions, and the relevance to federal statistics.

One of the goals for the workshop was to support and acknowledge those doing question evaluation and developing evaluation methodology. Encouragement for this work needs to come not only from the survey community but also from data users. Funders, sponsors, and data users should require that information on question quality (or lack thereof) be made public and that question evaluation be incorporated into the design of any data collection. Data producers need to institutionalize question evaluation and adopt and endorse agreed-upon standards. Data producers need to hold themselves and their peers to these standards as is done with standards for sample design and quality evaluation. Workshops like the QEM provide important venues for sharing information and supporting the importance of question evaluation. More opportunities like this are needed. This volume allows the work presented at the Workshop to be shared with a much wider audience—a key requirement if the field is to grow. Other avenues for publishing results of evaluations and of the development of evaluation methods need to be developed and supported.