

Response to Esposito's “Some Thoughts on the Use of Field Tests to Evaluate Survey Questionnaires”

Brian A. Harris-Kojetin, Ph.D.
Statistical and Science Policy
U.S. Office of Management and Budget

October 21, 2009

The views expressed in this presentation are those of the author and do not represent changes in OMB policy.



Overview

- Role of Field Testing in Questionnaire Evaluation
- When Should Agencies Conduct Field Tests?
- How do Federal Agencies Use Field Tests?
- Implications for Q-Bank



Role of Field Testing in Questionnaire Evaluation

- Key part of the process of developing and evaluating questions
- Involve more realistic production conditions and personnel
- Introduce more variables than laboratory methods, which generally precede field testing
- Field tests often include experiments, and may include behavior coding, and may be the source of data for IRT modeling.



When Should Agencies Conduct Field Tests?

- *Questions and Answers When Designing Surveys for Information Collections*
 - When developing new questionnaires
 - To see how respondents actually answer questions, and
 - Identify potential data quality problems, such as high item nonresponse rates.
 - To gather data to refine questionnaire items and scales and assess reliability or validity.
 - When planning a change in methodology or questions in an ongoing survey
 - To obtain quantifiable data to decide among the different questions.
 - To measure the effect of the change in methods or questions on resulting estimates.



When Should Agencies Conduct Field Tests?

- *Standards and Guidelines for Statistical Surveys*
 - **Standard 1.4:** Agencies must ensure that all components of a survey function as intended when implemented in the full-scale survey and that measurement error is controlled by conducting a pretest of the survey components or by having successfully fielded the survey components on a previous occasion.



When Should Agencies Conduct Field Tests?

- **Guideline 1.4.2:** Use field tests prior to *implementation of the full-scale survey* when some or all components of a survey system cannot be successfully demonstrated through previous work. The design of a field test should *reflect realistic conditions*, including those likely to pose difficulties for the survey. Elements to be tested include, for example, frame development, sample selection, *questionnaire design*, data collection, item feasibility, electronic data collection capabilities, edit specifications, data processing, estimation, file creation, and tabulations. A complete test of all components (sometimes referred to as a dress rehearsal) may be desirable for highly influential surveys.



How Do Federal Agencies Use Field Tests?

- New survey questionnaires or a major redesign of an ongoing survey.
 - Methodological as well as questionnaire changes
- Some periodic surveys (as opposed to ongoing collections), build in some kind of field test prior to each new collection period.
 - to assess new content or ordering or other methodological changes being considered



How Do Federal Agencies Use Field Tests?

- The first 50 or 100 fielded cases may be treated as a pretest
 - To make sure the items and the programmed instrument are functioning as planned.
 - May involve interviewer and/or respondent debriefings
 - Also some quantitative analysis of item responses
 - The goal is often to use the field test as a “trial run” and make only minor adjustments as necessary before going forward with the full-scale collection.



How Do Federal Agencies Use Field Tests?

- The field test is a separate and distinct precursor to the full-scale collection
 - Focus on obtaining quantitative data for further analysis to refine the survey instrument.
 - Assess timing of the administration
 - Assess how items are functioning
 - Distribution of responses and item missing rates
 - Psychometric analyses of item characteristics



How Do Federal Agencies Use Field Tests?

- To compare different questions through split panel tests.
 - May utilize a separate sample, such as a “methods panel” for the survey or may be used on “live” production sample depending on the survey.
 - May involve one or more modes depending on the modes used in the survey and the available resources.
 - For a major redesign effort, an agency may launch a large-scale separate survey in parallel to production survey to assess the impact of the new survey and/or methods on the key survey estimates.



The American Community Survey (ACS) Contest Test

- ACS is the replacement for the Decennial “long form” that is sent to 2.9 million households per year.
- Process for adding/changing content
 - Agency justification to Census and OMB
 - Cognitive testing
 - Split-panel field test
- Field test conducted on a large sample separate from the ongoing ACS
 - Testing must be conducted in all three modes of data collection: mail, CATI, and CAPI.



The American Community Survey (ACS) Contest Test

- The primary results are quantitative comparisons of different questions in terms of panel response rates, item response distributions, and item response rates.
 - Additional qualitative and quantitative methods, such as interviewer debriefings and behavior coding are often included.
 - Also typically include a reinterview to assess reliability over time.



The American Community Survey (ACS) Contest Test

- In 2008 tested two different versions of a Field of Degree question
 - For National Science Foundation's Division of Science Resources Statistics.
- The results from the content reinterview were key to assessing which of the two versions better captured the field of degree.
 - An open-ended question asking for the respondent's major performed much better than a set of response categories
 - Respondents had a difficult time figuring out which category best fit their major degree and so they checked multiple categories



Hospital Consumer Assessment of Health Plans Survey (HCAHPS)

- AHRQ and its contractors developed and cognitively tested items on hospital patients perceptions of their care.
- 66 items administered to medical, surgical, and obstetric patients who had an overnight hospital stay at one of 109 hospitals in Maryland, Arizona, or New York
- The focus of the field test was to obtain sufficient data for psychometric analysis



Hospital Consumer Assessment of Health Plans Survey (HCAHPS)

- Data Analysis consisted of
 - Exploratory factor analyses were conducted to guide refinements to the initially hypothesized structure,
 - The revised structure was evaluated using item-scale correlations, internal consistency reliability, and correlations with global ratings.
 - The revised instrument included 32 questions tapping seven domains of care as well as several global items.
 - Additional analyses were also conducted to identify variables for case-mix adjustment.



National Assessment of Education Progress (NAEP)

- NAEP is the source of the Nation's report card, providing an indication evaluation of how well students and schools are performing.
- NCES employs several strategies to develop NAEP items:
 - Small and large-scale pilot testing to test potential assessment items and select those that will be included on the final instruments.
 - Pilot tests are conducted with nationally representative samples of students to gather information about performance across the whole spectrum of student achievement.
 - In general, two items are tested for each one that will appear on the final instrument.



National Assessment of Education Progress (NAEP)

- NCES conducts “pre-calibration tests” of the final instruments (after pilot testing) to obtain IRT parameters in advance of the main assessment.
- Thus, the data from the main assessment can then be analyzed using the defined item parameters and produce scale scores much more quickly.



National Survey of Drug Use and Health (NSDUH)

- SAMHSA collects data on the use of alcohol, tobacco, and illicit substances to track the prevalence of substance use in the U. S. for all persons 12+ years of age.
- SAMHSA conducted separate validity and reliability studies using field tests
- Validity study used a separate sample (from the main) and obtained hair and urine specimens to compare to laboratory results to self-reports.
- Reliability Study used a subsample of the main sample and reinterviewed respondents between 5 and 15 days later .



Implications for Q-Bank

- Field tests provide vital information to assist in the development, refinement, and selection of survey questions
- Wide range of field test designs and purposes
 - Corresponding array of quantitative and qualitative data
- Hopefully, creative and useful ways will be found to incorporate this information

