Supporting Document A
Questionnaire Testing and Evaluation Methods for Censuses and Surveys

Version 1.2

Issued: 09 Mar 06

Census Bureau Standard
Pretesting Questionnaires and Related Materials for Surveys and Censuses

Authored by:

Theresa J. DeMaio
Principal Researcher
Statistical Research Division

Nancy Bates
Staff, Survey Improvement Coordination Staff
Demographic Surveys Division

Diane Willimack
Chief, Establishment Survey Methods Staff
Economic Statistical Methods and Programming Division

Jane Ingold
Chief, Content and Data Products Branch
Decennial Management Division

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

Document Management & Control

Version	Issue Date	Approval	Description
1.0	25 Jul 03	Associate Directors	Initial Release
1.1	29 Dec 04	Configuration Mgr.	Reformatted to comply with Census Bureau Identity Standard and Quality Program Document Management Plan
1.2	09 Mar 06	Configuration Mgr.	Inserted hyperlink for main standard

The most current version of this document is maintained on the Census Bureau Intranet and may be accessed from the Quality Management Repository.

Questionnaire Testing and Evaluation Methods for Censuses and Surveys[1]

Pretesting is critical for identifying problems for both respondents and interviewers with regard to question content, order/context effects, skip instructions, and formatting. Problems with question content, for example, include confusion with the overall meaning of the question as well as misinterpretation of individual terms or concepts. Problems with skip instructions may result in missing data and frustration by interviewers and/or respondents. Formatting concerns are relevant to self-administered questionnaires and may lead to confusion, as well as loss of information.

Pretesting is a broad term that incorporates many different methods or combinations of methods. This document briefly describes several techniques that can be used to test and evaluate questionnaires. This enumeration and description of potential pretesting and evaluation methods is meant to be exhaustive. These techniques have different strengths and weaknesses that make them valuable for identifying problems with draft questionnaires, and are useful at different stages of questionnaire/instrument development. Typically the use of multiple pretesting methods is more effective than the use of a single method in identifying problem questions and suggesting solutions.

How these methods are used individually or in conjunction with one another determines their suitability to meet the Pretesting Standards. The intent of the standards is that some kind of testing of the questionnaire with respondents is required.[2] Some methods enumerated here are not sufficient by themselves to satisfy this intent, and these will be identified in their descriptions. Other methods may meet the Pretesting Standards by themselves only when a draft questionnaire or instrument is used and when actual survey respondents are the subjects of the test. Nevertheless, the selection of pretesting methods and their application should be determined by their appropriateness to the problem at hand, resources permitting, rather than chosen solely to meet these criteria.

We divide the techniques into two major categories – pre-field and field techniques. Pre-field techniques generally are those used during the preliminary stages of questionnaire development. They include respondent focus groups, exploratory/feasibility company/site visits (for economic surveys), cognitive interviews, usability techniques, and expert reviews. Field techniques are

---

[1]This is based on two sources: 1) Protocol for Pretesting Demographic Surveys at the Census Bureau, prepared by Theresa DeMaio, Nancy Mathiowetz, Jennifer Rothgeb, Mary Ellen Beach, and Sharon Durant, dated June 28, 1993; and 2) Evolution and Adaptation of Questionnaire Development, Evaluation and Testing in Establishment Surveys, by Diane Willimack, Lars Lyberg, Jean Martin, Lilli Japec, and Patricia Whitridge. Monograph Paper for the International Conference on Questionnaire Development, Evaluation and Testing Methods, Charleston, SC, November, 2002.

[2]The only exception is expert review, which is allowed only under extreme time pressure.

those used to evaluate questionnaires tested under field conditions, in conjunction with a field test, or they may be used in conjunction with production data collection, particularly for ongoing or recurring surveys. These include behavior coding of interviewer/respondent interactions, interviewer debriefings, analysts' feedback, respondent debriefings, split sample tests, and analysis of item nonresponse rates, imputation rates, edit failures, or response distributions.

PRE-FIELD TECHNIQUES

Respondent Focus Groups, as noted above, are used early in the questionnaire development cycle and can be used in a variety of ways to assess the question-answering process.

They can be used to gather information about a topic before questionnaire construction begins. Objectives of focus groups used in this way range from learning how potential respondents structure their thoughts about a topic, to their understanding of general concepts or specific terminology, to their opinions about the sensitivity or difficulty of the questions.

Focus groups also can be used to quickly identify variations in language or terminology or interpretation of questions and response options. Used in this way, they may provide quicker access to a larger number of people than is possible with cognitive interviews.

One of the main advantages of focus groups is the opportunity to observe a large amount of interaction on a topic in a limited period of time. The interaction is of central importance – the focus group provides a means for group interaction to produce information and insights that may be less accessible without the interaction found in a group. However, the focus group, because of the group interaction, does not permit a good test of the "natural" response process, nor does the researcher have as much control over the process as would be true with either cognitive interviews or interviewer-administered questionnaires. One or two people in the group may dominate the discussion and restrict the input from other group members.

Used as described above, the focus group technique does not meet the minimal pretest standard, since it does not expose respondents to a questionnaire. However, another use of focus groups for self-administered questionnaires involves group administration of the instrument followed by a discussion of the experience. This provides information about the appearance and formatting of the questionnaire in addition to knowledge of content problems, and is sufficient to meet the minimal standard.

Exploratory or Feasibility Studies are another common method for specifying survey content relative to concepts. In economic surveys, these studies often take the form of *company or site visits*, which is what economic survey practitioners typically call them.

Because economic surveys rely heavily on business or institutional records, the primary goal of these company or site visits is to determine the availability of the desired data in records and their periodicity, and the definition of the concept as used in company records. Other goals include

assessment of response burden and quality, as well as identification of the appropriate respondent.

There tends to be a great deal of variation in the design of these company or site visits. Since they are exploratory in nature, the activity may continue until the economic survey or program staff understands the concepts sufficiently from the respondents' point of view, resources permitting of course. Purposive or convenience samples are selected that likely target key data providers. Sample sizes are small, perhaps as few as five and rarely more than thirty. Typically, several members of the survey or program staff, who may or may not include questionnaire design experts, conduct meetings with multiple company employees involved in government reporting. Information gained during these visits is beneficial in determining if the survey concepts are measurable, what the specific questions should be, how to organize or structure the questions related to the concept of interest, and identify to whom the form should be sent.

Exploratory or feasibility studies (company or site visits) may be multi-purpose. In addition to exploring data availability for the concept of interest, survey or program staff may also set up reporting arrangements and review operating units to ensure correct coverage. A common by-product of these visits is to solidify relationships between the companies and the survey or program staff. Since these visits are conducted prior to the development of the questionnaire of interest, they do not meet the Pretest Standards.

Cognitive Interviews are used later in the questionnaire development cycle, after a questionnaire has been constructed based on focus groups, site visits, or other sources. They consist of one-on-one interviews using a draft questionnaire in which respondents describe their thoughts while answering the survey questions. These interviews provide an important means of finding out directly from respondents about their problems with the questionnaire. In addition, small numbers of interviews (as few as fifteen) can yield information about major problems as respondents repeatedly identify the same questions and concepts as sources of confusion. Use of this technique is sufficient to meet the minimal standard.

Because sample sizes are small, iterative pretesting of an instrument is often possible. After one round of interviews is complete, researchers can diagnose problems, revise question wording to solve the problems, and conduct additional interviews to see if the new questions are less problematic.

Cognitive interviews may or may not be conducted in a laboratory setting. The advantage of the laboratory is that it offers a controlled environment for conducting the interview, and provides the opportunity for video as well as audio recording. However, laboratory interviews may be impractical or unsuitable. For example, economic surveys rarely, if ever, conduct cognitive interviews in a laboratory setting. Rather, cognitive testing of economic surveys is usually conducted on-site at the offices or location of the business or institutional respondent. One reason for this approach is to enable business or institutional respondents' access to records. Another is business respondents' reluctance to meet outside their workplaces for these interviews.

In many economic surveys, which tend to be relatively lengthy, requiring labor-intensive data retrieval from records, testing may be limited to a subset of questions or sections rather than the entire questionnaire. Thus, researchers must be careful to set the proper context for the target questions.

"Think aloud" interviews, as this technique has come to be called, can be conducted either concurrently or retrospectively – that is, the respondents' verbalizations of their thought processes can occur either during or after the completion of the questionnaire. As the Census Bureau conducts them, cognitive interviews incorporate follow-up questions by the researcher in addition to the respondent's statement of his/her thoughts. *Probing questions* are used when the researcher wants to have the respondent focus on particular aspects of the question-response task. For example, the interviewer may ask how respondents chose among response choices, how they interpreted reference periods, or what a particular term meant. *Paraphrasing* (that is, asking the respondents to repeat the question in their own words) permits the researcher to learn whether the respondent understands the question and interprets it in the manner intended, and may reveal better wordings for questions.

In surveys of businesses or institutions, in which data retrieval often involves business records, probing and paraphrasing techniques are often augmented by questions asking respondents to describe those records and their contents or to show the records to the researcher. Since data retrieval tends to be a labor-intensive process for business respondents frequently requiring the use of multiple sources or consultation with colleagues, it is often unrealistic for researchers to observe the process during a cognitive interview. Instead, *hypothetical probes* are often used to identify the sources of data, discover respondents' knowledge of and access to records, recreate likely steps taken to retrieve data from records or to request information from colleagues, and suggest possible estimation strategies.

Usability Techniques are used to aid development of automated questionnaires. Objectives are to discover and eliminate barriers that keep respondents from completing the automated questionnaire accurately and efficiently, with minimal burden. Aspects that deserve attention during testing include the language, fonts, icons, layout, organization, and interaction features, such as data entry, error recovery, and navigation. Typically, the focus is on instrument performance in addition to how respondents interpret survey questions. Problems identified during testing can then be eliminated before the instrument is finalized.

As with paper questionnaires, there are different usability techniques available depending upon the stage of development. One common technique is called the *usability test*. These tests are similar to cognitive interviews – that is, one-on-one interviews that elicit information about respondents' thought process. Respondents are given a *task*, such as "Complete the questionnaire," or smaller subtasks, such as "Send your data to the Census Bureau." The *think aloud, probing* and *paraphrasing* techniques are all used as respondents complete their assigned tasks. Early in the design phase, usability testing with respondents can be done using *low fidelity questionnaire prototypes* (i.e., mocked-up paper screens). As the design progresses, versions of the automated questionnaire can be tested to choose or evaluate basic navigation features, error

correction strategies, etc. Usability tests meet the recommended standard for electronic self-administered questionnaires, but need to be combined with more questionnaire-centered cognitive interviews to meet the minimal standard.

Disability accommodation testing is a form of usability testing which evaluates the ability of a disabled user to access the questionnaire through different assisted technologies such as a screen reader. *Expert reviews* (see below) are also part of the repertoire of usability techniques.

Research has shown that as few as three participants can uncover half of the major usability problems; four to five participants can uncover 80 percent of the problems; and ten participants can uncover 90 percent of the problems (Dumas and Redish, 1999).[3]

Methodological Expert Reviews, which are conducted by survey methodologists or questionnaire design experts rather than subject-matter experts, have as their objective the evaluation of the questionnaire for potential interviewer and respondent task difficulty. The means for achieving that objective is to draw on the expertise of seasoned survey researchers who have extensive exposure to either the theoretical or practical aspects of questionnaire design. Usually these reviews are conducted early in the questionnaire development process and in concert with other pretest methods.

The cognitive appraisal coding system (Forsyth and Lessler, 1991)[4] is a tool providing a more systematic approach to the methodological expert review process. This tool may be used effectively by questionnaire design experts who understand the link between the cognitive response process and potential measurement. Or it may also be used by novice staff or subject-area staff, as a guide for what to look for in their reviews of questionnaires. Like methodological expert reviews, results are used to identify questions that have potential for reporting errors.

Methodological expert reviews also can be conducted as part of a usability evaluation. Typically this review is done with an automated version of the questionnaire, although it need not be fully functional. Experts evaluate the questionnaire for consistency and application of user-centered principles of user-control, error prevention and recovery, and ease of navigation, training, and recall.

Both the methodological expert review and the cognitive appraisal coding system are efficient methods for identifying potential questionnaire problems. However, because there is no input directly from respondents when using these techniques, these are considered last resort methods

---

[3]Dumas, J. and Redish, J. 1999. *A Practical Guide to Usability Testing*, Portland, OR: Intellect.

[4]Forsyth, B. H., and Lessler, J. T. (1991), "Cognitive Laboratory Methods: A Taxonomy," in Measurement Errors in Surveys, Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S.(eds.), New York: John Wiley and Sons, Inc., pp. 393-418.

for meeting the minimal standard. Their use by themselves is sufficient to meet the minimal standard ONLY when extreme time constraints prevent the use of other pretesting methods. In such instances, they must be conducted by survey methodology experts and the results must be documented in a written report. This decision is made by subject-matter areas in consultation with methodological research areas in SRD and ESMPD.

FIELD TECHNIQUES

Field techniques may be used with pretests or pilot tests of questionnaires or instruments and survey processes. They may also be associated with ongoing periodic (or recurring) surveys. The value of testing draft questionnaires with potential survey respondents cannot be overstated, even if it simply involves observation and evaluation by questionnaire developers. However, use of some of the following pretest methods maximizes the benefit that can be gained from field testing.

Behavior Coding of Respondent/Interviewer Interactions involves systematic coding of the interaction between interviewers and respondents from live or taped field or telephone interviews to collect quantitative information. The focus here is on specific aspects of how the interviewer asks the question and how the respondent reacts. When used for questionnaire assessment, the behaviors that are coded focus on behaviors indicative of a problem with either the question, the response categories, or the respondent's ability to form an adequate response. For example, if a respondent asks for clarification after hearing the question, it is likely that some aspect of the question caused confusion. Likewise, if a respondent interrupts the question before the interviewer finishes reading it, then the respondent misses information that might be important to giving a correct answer. For interviewer-administered economic surveys, the coding scheme may need to be modified from traditional household applications, because interviewers for establishment surveys tend to be allowed greater flexibility. Use of this method is sufficient to meet the minimal standard.

In contrast to the pre-field techniques described earlier, the use of behavior coding requires a sample size sufficient to address analytic requirements. For example, if the questionnaire contains many skip patterns, it is necessary to select a large enough sample to permit observation of various paths through the questionnaire. In addition, the determination of sample sizes for behavior coding should take into account the relevant population groups for which separate analysis is desirable.

The value of behavior coding is that, since it evaluates all questions on the questionnaire, it allows systematic detection of questions that have large numbers of behaviors that reflect problems. However, it is not usually designed to provide answers about the source of the problems. It also may not distinguish which of several similar versions of a question is better. And finally, behavior coding does not always provide an accurate diagnosis of problems. It can only detect problems that are manifest in interviewer or respondent behavior. Some important problems, such as respondent misinterpretations, are largely hidden because both respondents

and interviewers tend to be unaware of them. Behavior coding is not well-suited for identifying such problems.

Respondent Debriefing involves using a structured questionnaire following data collection to elicit information about respondents' interpretations of survey questions. This may be done by incorporating structured follow-up questions at the end of a field test interview, or it may require re-contacting the respondent following return of a completed self-administered questionnaire. In economic surveys, respondent debriefings sometimes are called "response analysis surveys (RAS)" or "content evaluations." Although usually interviewer-administered, the mode for respondent debriefings may be self-administered. Some Census Bureau economic surveys have conducted respondent debriefings by formulating them as self-administered questionnaires and enclosing them with survey forms during pilot tests or production data collections. Use of this method is sufficient to meet the minimal standard.

Sample sizes and designs for respondent debriefings vary. Sample sizes may be as small as 20 or as large as several hundred. Designs may be either random or purposive, such as conducting debriefings with respondents that exhibited higher error rates or errors on critical items. Since the debriefing instrument is structured, empirical summaries of results may be generated.

When used for testing purposes, the primary objective of respondent debriefing is to determine whether concepts and questions are understood by respondents in the same way that the survey designers intend. Sufficient information is obtained to evaluate the extent to which reported data are consistent with survey definitions. For instance, respondents may be asked whether they included or excluded particular items in their answers, per definitions. In economic surveys, use of records and/or estimation strategies may also be queried. In addition, respondent debriefings can be useful in determining the reason for respondent misunderstandings. Sometimes results of respondent debriefing show that a question is superfluous and can be eliminated from the final questionnaire. Conversely, it may be discovered that additional questions need to be included in the final questionnaire to better operationalize the concept of interest. Finally, the data may show that concepts or questions cause confusion or misunderstanding as far as the intended meaning is concerned.

A critical aspect of a successful respondent debriefing is that question designers and researchers must have a clear idea of potential problems so that good debriefing questions can be developed. Ideas about potential problems can come from pre-field techniques (e.g., cognitive interviews) conducted prior to the field test, from analysis of data from a previous survey, from careful review of questionnaires, or from observation of earlier interviews.

Respondent debriefings have the potential to supplement the information obtained from behavior coding. As noted above, behavior coding demonstrates the existence of problems but does not always indicate the source of the problem. When designed properly, the results of respondent debriefing can provide information about the sources. In addition, respondent debriefing may reveal problems not evident from the response behavior.

Interviewer Debriefing has traditionally been the primary method used to evaluate field or pilot tests of interviewer-administered surveys. It also may be used following production data collection prior to redesigning an ongoing periodic or recurring survey. Interviewer debriefing consists of group discussions and/or structured questionnaires with the interviewers who conducted the test to obtain their views of questionnaire problems. The objective is to use the interviewers' direct contact with respondents to enrich the questionnaire designer's understanding of questionnaire problems. While it is a useful evaluation component, it is not sufficient as an evaluation method and does not meet the minimal pretest standards. Interviewers may not always be accurate reporters of certain types of questionnaire problems for several reasons. When interviewers report a problem, we do not know whether it was troublesome for one respondent or for many. Interviewers' reports of problem questions may reflect their own preference for a question rather than respondent confusion. Also, experienced interviewers sometimes change the wording of problem questions as a matter of course to make them work, and may not even realize they have done so.

Interviewer debriefings can be conducted in several different ways: in a group setting, through rating forms, or through standardized questionnaires. *Group setting debriefings* are the most common method. They essentially involve conducting a focus group with the field test interviewers to learn about their experiences in administering the questionnaire. *Rating forms* obtain more quantitative information by asking interviewers to rate each question in the pretest questionnaire on selected characteristics of interest to the researchers (e.g., whether the interviewer had trouble reading the question as written, whether the respondent understood the words or ideas in the question). *Standardized interviewer debriefing questionnaires* collect information about the interviewers' perceptions of the problem, prevalence of a problem, reasons for the problem, and proposed solutions to a problem. They can also be used to ask about the magnitude of specific kinds of problems to test the interviewers' knowledge of subject-matter concepts.

Analysts' Feedback is a method of learning about problems with a questionnaire specific to the economic area. Since at the Census Bureau most economic surveys are self-administered, survey or program staff analysts in the individual subject areas, rather than interviewers, often have contact with respondents. Feedback from analysts about their interactions with respondents may serve as an informal or semi-formal evaluation of the questionnaire and the data collected. These interactions include "Help Desk" phone inquiries from respondents, as well as follow-up phone calls to respondents by analysts investigating suspicious data flagged by edit failures. This information is more useful when analysts systematically record feedback from respondents in a log. This enables qualitative evaluation of the relative severity of questionnaire problems, because strictly anecdotal feedback may sometimes be overstated.

Another way of gathering feedback is for questionnaire design experts to conduct focus groups with analysts who review data and resolve edit failures to identify potential candidate questions for redesign and/or evaluation by other methods. Regardless of how respondent feedback is captured, it is typically conveyed by analysts early in the questionnaire development cycle for recurring surveys for the purpose of indicating problematic questions. However, while collecting

feedback from analysts is a useful evaluation component, it does not meet the minimal pretesting standard.

Split Panel Tests refer to controlled experimental testing of questionnaire variants or data collection modes to determine which one is "better" or to measure differences between them. Split panel experiments may be conducted within a field or pilot test, or they may be embedded within production data collection for an ongoing periodic or recurring survey. For pretesting draft versions of a questionnaire, the search for the "better" questionnaire requires that an a priori standard be determined by which the different versions can be judged.

Split panel tests can incorporate in a single question a set of questions or an entire questionnaire. It is important to provide for adequate sample sizes in designing a split sample test so that differences of substantive interest can be measured. In addition, it is imperative that these tests involve the use of randomized assignment within replicate sample designs so that differences can be attributed to the question or questionnaire and not to the effects of incomparable samples. Split panel testing is sufficient to meet both the minimal standard and the recommended standard for data with important policy implications.

Another use of the split panel test is to calibrate the effect of changing questions. Although split panel tests are expensive, they are extremely valuable in the redesign and testing of surveys for which the comparability of the data collected over time is an issue. They provide an all-important measure of the extent to which changes in major survey redesigns are due to changes in the survey instrument, interview mode, etc., as opposed to changes over time in the subject-matter of interest.

Comparing response distributions in split panel tests produces measures of differences but does not necessarily reveal whether one version of a question produces a better understanding of what is being asked than another. Other question evaluation methods, such as respondent debriefings, interviewer debriefings, and behavior coding, are useful to evaluate and interpret the differences observed in split panel tests.

Analysis of Item Nonresponse Rates, Imputation Rates, Edit Failures, or Response Distributions from the collected data can provide useful information about how well the questionnaire works. In household surveys, examination of item nonresponse rates can be informative in two ways. First, "don't know" rates can determine the extent to which a task is too difficult for respondents to do. Second, refusal rates can determine the extent to which respondents find certain questions or versions of a question to be more sensitive than others.

In economic surveys item nonresponse may be interpreted to have various meanings, depending on the context of the survey. In some institutional surveys (e.g., hospitals, prisons, schools) where data are abstracted from individual person-level records, high item nonresponse is considered to indicate data not routinely available in those records. Item nonresponse may be more difficult to detect in other economic surveys where questions may be left blank because

they are not applicable to the responding business, or the response value may be zero. In these cases, the data may not be considered missing at all.

Response distributions are the frequencies with which answers were given by respondents during data collection. Evaluation of the response distributions for survey items can determine if there is variation among the responses given by respondents or if different question wordings or question sequencings produce different response patterns. This type of analysis is most useful when pretesting either more than one version of a questionnaire or a single questionnaire for which some known distribution of characteristics exists for comparative purposes. Use of this method in combination with a field pretest is sufficient to meet the minimal standard.

The quality of collected data also may be evaluated by making comparisons, reconciling or benchmarking to data from other sources. This is especially true for economic data, but benchmarking data are also available for some household surveys.

CONCLUSION

The Census Bureau advocates that both pre-field and field techniques be undertaken, as time and funds permit. Pre-field methods alone may not be sufficient to test a questionnaire. In terms of pre-field techniques, the choices include focus groups, exploratory/feasibility studies and cognitive interviews, and usability techniques. For continuing surveys that have a pre-existing questionnaire, cognitive interviews should be used to provide detailed insights into problems with the questionnaire whenever time permits or when a redesign is undertaken. Cognitive interviews may be more useful than focus groups with a pre-existing questionnaire because they mimic the question-response process. For one-time or new surveys, focus groups are useful tools for learning about the way respondents think about the concepts, terminology, and sequence of topics prior to drafting the questionnaire. In economic surveys, exploratory/feasibility studies, conducted as company or site visits, also provide information about structuring and wording the questionnaire relative to data available in business/institutional records. Usability techniques are an increasingly important technique as surveys move to automated means of data collection.

Some type of testing in the field is encouraged, even if it is only evaluated based on observation by questionnaire developers. More helpful is small-to-medium-scale field or pilot testing with more systematic evaluation techniques. Pretesting typically is more effective when multiple methods are utilized. The objective and the subjective methods, and the respondent-centered and the interviewer or analyst-centered methods, complement each other with respect to identifying problems, source of the problems, and potential solutions. The relative effectiveness of the various techniques for evaluating survey questions depends on the pretest objectives, sample size, questionnaire design, and mode of data collection.

In terms of meeting the pretesting standards, cognitive testing, focus groups involving administration of questionnaires, or field testing including behavior coding or respondent debriefing is necessary to meet the minimal standard; split panel testing is necessary to meet the recommended standard for data with important policy implications; cognitive testing is necessary

to meet the recommended standards for supplemental instruments and materials; and usability testing is necessary to meet the recommended standard for electronic self-administered questionnaires. However, pretesting should not be limited to these techniques. One or more other pretesting methods should be carefully considered and used as time permits to provide a thorough evaluation of questionnaire problems and documentation that a question or questionnaire "works," according to the pretesting standards.