Statistical Policy
Working Paper 47

Evaluating Survey Questions: An Inventory of Methods

Statistical and Science Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

January 2016

THE FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY
January 2016

MEMBERS

Jonaki Bose (Chair)
Substance Abuse and Mental Health Services Administration

Nancy Bates
U.S. Census Bureau

Patrick Cantwell
U.S. Census Bureau

Chris Chapman
National Center for Education Statistics

John Eltinge
Bureau of Labor Statistics

Dennis Fixler
Bureau of Economic Analysis

Barry Graubard
National Cancer Institute

Arthur B. Kennickell
Federal Reserve Board

Jennifer Madans
National Center for Health Statistics

Rochelle (Shelly) Martinez
Office of Management and Budget

Jaki McCarthy
National Agriculture Statistic Services

Pam McGovern (Webmaster)
National Agriculture Statistic Services

Grace Medley (Secretary)
Substance Abuse and Mental Health Services Administration

Peter Meyer (CDAC Chair)
National Center for Health Statistics

Peter Miller
U.S. Census Bureau

Renee Miller
Energy Information Administration

Jeri Mulrow
National Science Foundation

Amy O'Hara
U.S. Census Bureau

Polly Phipps
Bureau of Labor Statistics

Nathaniel Schenker
National Center for Health Statistics

Rolf Schmitt
Bureau of Transportation Statistics

Marilyn Seastrom
National Center for Education Statistics

Joy Sharp
Bureau of Transportation Statistics

Katherine K. Wallman (Champion)
Statistical and Science Policy
Office of Management and Budget

G. David Williamson
Agency for Toxic Substances and Disease Registry

Gordon Willis
National Cancer Institute

Statistical Policy
Working Paper 47

Evaluating Survey Questions: An Inventory of Methods

Prepared by
Subcommittee on Questionnaire Evaluation Methods

Statistical and Science Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

January 2016

Evaluating Survey Questions:

An Inventory of Methods

A Report of the Subcommittee on
Questionnaire Evaluation Methods

January 2016

Washington, D.C.

Members of the Subcommittee on Questionnaire Evaluation Methods

Jennifer Madans, Chair
National Center for Health Statistics

Lionel Deang
Social Security Administration

Jennifer Edgar
Bureau of Labor Statistics

Scott Fricker
Bureau of Labor Statistics

Patricia Goerman
U.S. Census Bureau

Kashka Kubzdela
National Center for Education Statistics

Jaki McCarthy
National Agricultural Statistics Service

Kathy Ott
National Agricultural Statistics Service

Kristen Miller
National Center for Health Statistics

Rebecca L. Morrison
National Science Foundation

Cleo Redline
National Center for Education Statistics

Paul Scanlon
National Center for Health Statistics

Kristin Stettler
U.S. Census Bureau

Diane K. Willimack
U.S. Census Bureau

Stephanie Willson
National Center for Health Statistics

,
,

Evaluating Survey Questions: An Inventory of Methods

Contents

Evaluating Survey Questions: An Inventory of Methods

Introduction

This document was developed by the Federal Committee on Statistical Methodology (FCSM) Subcommittee on Question Evaluation Methodology to provide an inventory of methods used by federal statistical agencies to evaluate survey questions. Since question development methods play a crucial role in assuring the quality of survey questions and the data they produce, this document also includes a section on developing survey questions which outlines important steps in creating the survey questions that will subsequently be tested and evaluated. The question development section is followed by a discussion of question evaluation, a description of data collection methods, and a list of the methods federal agencies use to evaluate survey questions, along with a definition, common uses, strengths, and limitations for each evaluation method.

Question Development and Evaluation Methods (QDEM)

The many techniques employed to assure that survey questions evoke the intended response and to evaluate question quality come under the broad umbrella label of "question development and evaluation methods" (QDEM). This includes assuring: (1) that the survey questions ask what the researcher intended, (2) that they are consistently understood as intended, (3) that respondents are willing and able to consistently answer the questions as intended, and (4) that providing the full response is not overly burdensome (Groves et al., 2009). This process involves both development and evaluation of the questions (Figure 1). While conceptually distinct, the development and evaluation processes become intertwined in practice.

Question development involves identifying and defining the concepts to be measured, and drafting the questions to measure these concepts. After questions have been developed, they should be assessed using one or more of the evaluation methods defined in this inventory. The process should be iterative, where questions evaluation (testing) informs revisions (development) to the questions, followed by evaluation of the revised version(s), with further revisions if necessary. Through this process, revisions constitute a later stage of question development.

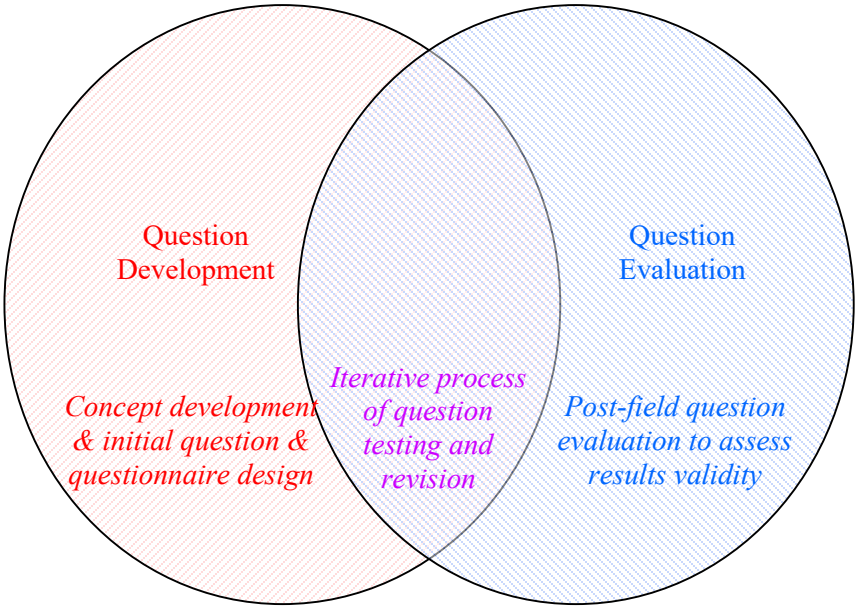

Fig. 1. Graphical representation of the relationship between survey question development and evaluation methods. The focus of this document is depicted by the blue circle (survey question evaluation methods), which includes methods that are used both in question development and evaluation. The overlap depicted is intended to be illustrative and not to convey the actual size of the overlap.

Question Development

The purpose of a survey question is to collect data in such a way that respondents (surveyed members of the target population) interpret the questions consistently and as the question developer intended. The resulting information will then reflect one or more underlying *concepts* of interest to the researcher and will satisfy the needs of data users. This document focuses on methods that may be used to evaluate the effectiveness of survey questions in achieving this purpose, after the survey questions have already been drafted. However, before survey questions can be effectively evaluated, the underlying concepts of interest must be clearly defined, with a deliberate and comprehensive approach implemented for drafting the survey questions themselves. This is essential to ensure that the questions capture the intended concepts and collect appropriate data. Thus question development activities are critical to assuring the quality of survey questions and the data they produce. This section provides a brief overview of conducting question development activities.

Concepts are made up of *attributes*, which may be thought of as bits of measureable information. When developing questions, the first step is to define the concepts of interest and then to identify the attributes associated with the underlying concept(s). The identified attributes become the raw material for drafting survey questions.

The degree to which survey questions capture data that truly reflect the desired underlying concepts is referred to as *construct validity* (e.g., Groves, 1989; Groves et al., 2009). Construct validity is one of the quality factors assessed by question evaluation. However, if poor conceptual specification is not recognized until question evaluation research is being conducted, then researchers need to backtrack into earlier stages of question(naire) development using methods that define underlying concepts, identify their attributes, determine measurements associated with these attributes, and aid in the creation of draft questions that better reflect construct validity.

Many seemingly straightforward concepts are actually quite complex when it comes to identifying their attributes, specifying practical measurements, and formulating survey questions. People can differ in how they articulate, define, and operationalize these concepts. Consider, for example, the concept of employment. While it may seem obvious to ask a respondent "Do you have a job?" there are nuances in the concept that need to be clarified to ensure respondents interpret the question as intended and answer accurately. Volunteer work, unpaid work and temporary positions are all examples of situations which might cause respondent confusion or uncertainty when answering questions about their employment.

Development work must be done to identify the relevant attributes of the construct of interest to ensure the question can measure it accurately and reliably. Thus, question development (Willeboordse, 1998) begins with:

1. identifying how data produced by the survey are to be used,
2. specifying the concepts that address the identified use,
3. dismantling the concepts into their component attributes,
4. developing specifications for the attributes, and
5. defining practical measurements.

Based on the information gathered in these steps, questions may then be written, and placed in the appropriate context given the mode of delivery of the survey questionnaire. Specifically, mode of collection may affect concept measurement so that question wording and structure may not be exactly the same across modes (Martin et al, 2007). For instance, a visual, written self-administered collection may display all response categories at once, whereas a telephone interviewer-administered collection may be better suited for a series of shorter questions to measure the same concept. Question development and concept measurement should capitalize on the strengths of each mode of collection while being mindful of how these differences might affect the comparability of the information generated.

There are a number of methods useful in aiding the initial development of survey questions to appropriately reflect underlying concepts of interest and achieve construct validity at the question development stage (Boivin & O'Rear, 2012; Snijkers & Willimack, 2011). A common feature of many question(naire) development methods is that they lend themselves to exploration and discovery, using open, free-form, in-depth discussions or interviews with

stakeholders, subject matter experts, and members of the target population. Often researchers take an iterative approach, adjusting protocols and preliminary draft survey questions as the concept gains further specification.

Question Development Methods

Pre-questionnaire methods, including one-on-one interviews, questionnaire mining, and focus groups, among other methods, constitute the questionnaire and question development phase. While there is no strict methodology for translating the findings from these various methods into usable survey questions, these tools are useful in providing the question developer with enough information to begin the process. This section is not intended to provide a guide to question design, but rather to point out some of the more common methods used in this process. For more information, there is a vast literature on survey question and questionnaire development and design that one may consult.

We identify two primary types of methods used during the question and questionnaire development phase—those in which the potential survey respondents play an active role, and those in which they play a passive role. Both of these categories, and the methods within them, have various strengths and weakness, which are briefly described below.

Development using Respondents Actively

We describe five major types of pre-questionnaire interviews and interactions in which targeted respondents or data users play an active role: ethnography, early stage scoping, focus groups, concept mapping, and data user needs assessments. These interactions are generally unstructured and designed to understand how potential respondents or data users think about the survey topic and what information they have access to. Ethnographic interviews are most commonly used in (though not limited to) the design of household surveys, whereas similar techniques are referred to as early stage scoping in establishment surveys. Focus groups are used commonly in the design of both household and establishment surveys. Concept mapping and data user needs assessments can be used for either household or establishment surveys. All of these methods use purposive samples, produce mainly qualitative data, and can give survey designers insight into how to best word and frame questions and response categories to both reduce cognitive burden and to increase construct validity.

Ethnography: Ethnography is a holistic qualitative methodology whose overall purpose is to place and describe a culture or cultural phenomenon within its natural context. In contrast to others in this inventory, ethnography is not a singular method, but rather an approach that modifies other methods. For instance, semi-structured interviews can be considered "ethnographic interviews" when their focus is specifically understanding how participants view and understand an issue or concept from their own cultural point-of-view. The findings from ethnographic methods are normally qualitative, and particular to whatever culture or subculture the participants belong to. They typically provide detailed descriptions about specific cultures or phenomena (Geertz 1973, Bernard 2011). Ethnography and

ethnographic interviews are incredibly flexible, and may incorporate other methods, even ones that produce quantitative data (Berlin and Kay, 1969; Weller and Romney, 1988).

Early-Stage Scoping: The primary goal of early stage scoping (Stettler and Featherston, 2010) is to determine how closely desired data matches available data and gain knowledge of the survey concepts from the respondent's point of view (Willimack, et al., 2004). To do this, researchers meet with potential respondents within a survey population and ask open-ended questions about survey concepts, data availability, record keeping practices, timing, and data sensitivity (Freedman and Rutchik, 2002). Ideally, these studies are conducted in the early phases of questionnaire development, as a means of assessing what questions it may be possible to ask in the context of a survey (Stettler and Featherston, 2012). The method also allows researchers and survey managers to become familiar with the terminology used by respondents, which can then be used in writing questions and developing survey protocols. Depending on the study needs and available resources, the interviews can be conducted in-person or over the telephone.

Focus Groups : Focus groups are group discussions that are designed to listen to and gather information from participants that can help the researcher identify trends and patterns (Krueger & Casey, 2000). This technique can also be used to elicit consensuses, or "group think." These discussions are normally led by a single researcher, or moderator, and tend to include individuals who are homogenous across some socio-cultural characteristic of interest (i.e. education, race, sexual orientation). In the context of question and questionnaire development, focus groups represent a relatively inexpensive method for determining how a particular social group thinks and speaks about a particular issue or construct.

Concept Mapping (Hox, 1997; Haraldsen, 2003): Concept mapping offers a structured way of studying concepts, to allow researchers to move from a concept to an actual question. Participants selected from the target population first generate statements that describe relevant aspects, or attributes, of the concepts under study. Next, participants individually group these statements according to their own views. The statement groupings are combined into a similarity matrix and analyzed using a multi-dimensional scaling technique. The result is a concept map, showing clusters of the statements, which may be interpreted as the attributes related to the concepts of interest. Using the concept map, participants discuss possible meanings and acceptable labels for each statement cluster. The concept map guides translation of the statements into survey questions, as well as structuring the questionnaire into blocks of related questions.

Data User Needs Assessments (Willimack, et al., 2004; Mulrow et al., 2007): Data users, researchers, and stakeholders work in concert with survey methodologists to develop survey content by specifying and clarifying data needs. In a workshop setting they identify how and why the data of interest are to be used; determine gaps in existing data; gain insight into data needs for specific issues; and create a preliminary set of data priorities. Lists of requested data items may be ranked iteratively by different groups of stakeholders and data users, while survey personnel distinguish items previously collected with high quality as opposed to items of questionable quality.

Panels of industry experts aid researchers in drilling down from concepts to attributes to common definitions and metrics. Early drafts of questions and questionnaires are developed based on these exercises.

Strengths: The first four methods, ethnography, early stage scoping, focus groups, and concept mapping, provide questionnaire developers information about how members of a cultural group or organization understand and perceive their world, and can elicit the cultural underpinnings needed to understand how respondents think about, process, and respond to questions. They also allow developers to better understand how respondents' cultural domains (classes or groups of related phenomena) relate to the constructs they wish to explore. Data user needs assessments can provide question developers with a clearer understanding of the information that is needed by the data users. These methods can shorten the time required to write and subsequently test and evaluate questions, help refine data measures to match records, and limit pretesting to data measures that are feasible for respondents (Stettler & Featherston, 2010).

Limitations: Because the focus of four of the pre-questionnaire methods (ethnography, early stage scoping, focus groups, concept mapping) is on the level of a culture or sub-culture, these methods may not be considered ideal for *post hoc* questionnaire evaluation, and may be better suited for survey and question design and development. That is, they focus on refining the concepts under study, rather than the questions used to measure them. In addition, like other methods that rely on non-random samples, their findings are limited in their generalizability. Questionnaire developers need be sure to take this limitation into account both when constructing a sample for the pre-questionnaire interactions and when analyzing their findings. Both of these focus on the refining concept under study, rather than on actual questions used to measure it. Moreover, data user assessments based on selected respondents may not cover the full range of attributes or user needs, potentially leading to gaps in measurement.

Development using Respondents Passively

We describe two development methods that utilize data or other information that is already available rather than collecting new information from respondents. These analyses leverage data collection efforts that have already been conducted, and so have the advantage of not requiring as many resources as the previously discussed methods. All of these methods use existing information or expertise to arrive at a refined understanding of the concept. This understanding is critical to ensure that the survey designers can develop questions to measure the construct as intended.

Dimension or Attribute Analysis (Hox, 1997): This method involves researchers specifying a concept using existing theory and logical reasoning, though they may also incorporate previous empirical research. The goal of this work is to identify a network of concepts that are logically tied together. Using that network of concepts, one or more appropriate empirical attributes or indicators are defined for each concept. These attributes are used

to form the basis for survey questions. For example, a researcher trying to measure the concept of health might identify the attributes of weight, height, blood pressure and eating habits.

Mining of Questionnaires (Snijkers & Willimack, 2011; Giesen & Hak, 2005): Researchers "mine" existing questionnaires for questions that seem to be relevant in the context of the study. Subject matter experts help to appraise the selected questions as to how they are related to the central concept(s). Reports of pre-test studies for these questions, analysis of item non-response, examination of the types and frequency of edit failures, or analysis of the response data may be used to study the validity and measurement problems of questions. Questions identified during this exercise should be used judiciously, and may require some modification, in subsequent survey questionnaires, to fill measurement gaps and avoid new error sources.

Strengths: These techniques are relatively easy to use since they do not require collection of new information from active participants. In addition, their financial costs could be significantly lower. Finally, in some cases, questions that have already been pretested and/or used in production can be used, which means the questionnaire developer may draw upon prior expertise and experience with the questions and simply use or modify questions already in existence, rather than creating new questions.

Weaknesses: These methods make a priori assumptions about how to measure concepts and tend to replicate the status quo. These techniques rely on historical information and concepts that may not be accurate in the contemporary context. Finally, the question(naire) developer should exercise caution when mining questionnaires. An existing question's context plays an important role in the effectiveness of the question, and a question may be successful in one context but not another. The universe, purpose for which the data are being collected, number of questions or attributes measuring a concept, question flow, and mode of collection all provide important context affecting any single question or concept measure. Questions should be used or adapted after careful consideration of these contextual factors that affect concept measurement.

These are just a few techniques to guide survey researchers, stakeholders, and sponsors in developing survey questions that measure the chosen underlying concepts. In addition, some of the question evaluation methods described in the following sections may also be adapted and used for refining draft or previously used questions as well as for early stages of question development.

Question Evaluation

Question evaluation generally aims to assess whether questions consistently evoke the intended response, and is typically done using a set of methods, with each method providing different insight about the question's performance. The selection of appropriate evaluation methods is driven by the type and mode of data collection, the

target population, and previous use and performance of the survey questions. Using a set of complementary question evaluation methods is typically more effective in identifying problems and suggesting solutions than using a single method (U.S. Census Bureau Statistical Quality Standards, Reissued Jun 2012[1]). The evaluation methods used in a given survey need to be carefully selected and applied in order to successfully develop questions that provide the intended information and evaluate their performance.

Data Collection Vehicles

Language describing different data collection vehicles for question evaluation is not used consistently across federal agencies, and may even be idiosyncratic to individual agencies. For example, what one agency calls *dress rehearsal* another calls *field test* while another agency uses terms like *pretest* or *pilot test*. The goal of this section is to introduce the key features of different data collection vehicles, and provide the considerations typically used when selecting one vehicle over another. This document does not attempt to define the mechanisms or to delineate what constitutes a pretest, versus a pilot test, field test, dress rehearsal, etc.

The question evaluation methods described in this inventory can be carried out in a variety of data collection vehicles, which vary along several important characteristics: sample size (i.e., small scale to large scale); sample type (i.e., convenience versus seeded versus probability sample); setting (i.e., the laboratory versus the field); timing (i.e., before, during, or after survey production); and the extent to which the vehicle mimics the conditions of the full production survey (i.e., very little similarity to complete replication of field conditions).

It is critical that the data collection vehicle in which the question evaluation method (or sets of methods) is carried out be chosen based on the aims of the research. The characteristics of the vehicle will ultimately determine the data that can be obtained from an evaluation method and what can be learned about the survey questions. In addition, as is evident in the descriptions of the methods in the following section, some question evaluation methods lend themselves better to a given data collection settings than others.

To illustrate more specifically, consider choosing to conduct a small-scale study in the field after the survey is conducted. This type of vehicle is conducive to including respondent and interviewer debriefings, where respondents are questioned soon after they complete the survey to identify problematic questions (e.g., confusing, burdensome, or requiring data that are not available or reliable), and interviewers may be asked to suggest improvements to the survey instruments and procedures based on their observations and interactions with respondents. As such, small field tests allow for examining specific aspects of the survey instrument, procedures, and/or respondent behavior under field conditions, in a relatively short time, and at less expense than larger tests. At the same time, because conditions are not exactly the same as the production survey's, including the use of a small

[1] https://www.census.gov/quality/standards/index.html

sample size, these tests may not reveal issues with regard to the survey instrument, respondent recruitment, or data collection procedures that are specific to subpopulations insufficiently represented in the pretest's sample.

In contrast, choosing to conduct a field test with a large sample size allows for a more representative sample of the target population and more extensive evaluation of the survey instrument and procedures. Large-scale field tests, typically done before a survey goes into production, can support a wider array of quantitative question evaluation methods, such as item response theory, latent class analysis, validation and record-check studies, or split-sample/randomized experiments. They also permit greater quantitative analyses of response quality and are more likely to reveal issues specific to different subpopulations. At the same time, the large sample sizes do not lend themselves to in-depth qualitative analysis and thus they are generally not able to reveal issues such as the reasons behind question wording problems. Also, given the time and expense larger field tests involve, they are not well suited to quick iterative cycles, where the questions are evaluated, revised, and then re-evaluated.

To give another example, choosing to conduct a small-scale test in the laboratory permits the use of methods that yield different strengths and weaknesses from those of the larger-scale field studies. The strength of laboratory-based vehicles is in their ability to provide in-depth information regarding respondents' thought processes, respondents' reactions to questions, and the underlying rationales for revising problematic questions. Although it is inappropriate to make statistical inferences using results from qualitative research methods, or those based generally on non-representative samples, gaining a rich and complex understanding of respondent interaction with the survey questions is invaluable in evaluating and improving question quality, even if one is unsure of the extent to which the results can be generalized to the full target populations. Laboratory-based vehicles also lend themselves to quick turnaround evaluations, and can be done before, during, or after a survey is fielded, depending on the research goals. Evaluation of new questions typically occurs before a survey is put into production, while unexpected results can be explored using laboratory-based vehicles after a survey has been conducted.

In summary, just as the question evaluation methods described in the next section vary in the benefits and limitations that characterize them, so do decisions about the characteristics of the data collection vehicle that will support the methods used. Determining the data collection vehicle is critical, because these decisions will affect which evaluation methods can be used and thus what can be learned about the survey questions. When reporting on use of these vehicles, agencies should clearly describe all features of the study, such as sample size and replication of field conditions. The absence of consensus among agencies on the terminology and definitions of data collection methods renders any reference to or comparison of such mechanisms across agencies misleading, unless a reference or complete description is accompanied. Finally, as noted with multiple evaluation methods below, there are advantages to using multiple data collection vehicles. Small and large scale studies, or those done in a laboratory setting and those in a field setting, can be done in conjunction with each other to provide complementary information that might otherwise be missed.

Combining Multiple Methods

This document contains an inventory of question evaluation methods. A number of the methods described in the next section may be applied to different research purposes in a variety of contexts. Thus, we must reiterate that our focus here is on the use and utility of the methods in evaluating survey questions. Although we describe single methods in this inventory, there are advantages to using them together. Using multiple methods together permits researchers to leverage the strength of each and to compensate for the weaknesses of each. The combination of methods should be done in a thoughtful way, to ensure that each method works alone to meet a specific research objective, and the group of methods put together also allows the researcher to meet the overall research goals. Since there is no prescription for which methods to combine or what order to do them in, Appendix A provides three examples of research using multiple methods to address a research question.

References

Berlin, Brent, and Kay, Paul. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press.

Bernard, Russell. (2011). *Research Methods in Anthropology*: *Qualitative and Quantitative Approaches.* Lanham, MD: Alta Mira Press.

Boivin, S., and O'Rear, I. (2012). "Measurement Strategies for Identifying Holders of Certificates and Certifications," Proceedings of the *FCSM Research Conference, Federal Committee on Statistical Methodology*, Washington, DC.

Freedman, S.R., and Rutchik, R.H. (2002). "Information Collection Challenges in Electric Power and Natural Gas." Presentation at the Joint Statistical Meetings. http://www.amstat.org/sections/srms/Proceedings/y2002/files/JSM2002-000226.pdf.

Geertz, Clifford. (1973). *The Interpretation of Cultures*. New York, NY: Basic Books.

Giesen, D., and Hak, T. (2005). "Revising the Structural Business Survey: From a Multi-Method Evaluation to Design," *Proceedings of the FCSM Research Conference, Federal Committee on Statistical Methodology,* Washington, DC.

Groves, R.M. (1989). *Survey Errors and Survey Costs.* New York: John Wiley and Sons.

Groves, R.M., Fowler, F.J., Jr., Couper M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*, 2nd edition. Hoboken, NJ: John Wiley and Sons.

Haraldsen, G. (2003). "Searching for Response Burden in Focus Groups with Business Respondents," in Prüfer, P., Rexroth, M., and Fowler, F. (eds.), *Proceedings of the 4th Workshop on Questionnaire Evaluation Standards* (pp. 113–123), QUEST 2003, October 21–23, 2003, ZUMA Nachrichten, Spezial Band 9, Mannheim, Germany.

Hox, J. (1997). "From Theoretical Concepts to Survey Questions," in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.), *Survey Measurement and Process Quality,* New York: Wiley.

Krueger, R.A., and Casey, M.A. (2000). *Focus Groups: A Practical Guide for Applied Research,* 3rd edition. Thousand Oaks, CA. Sage Publications.

Martin, E., Hunter Childs, J., DeMaio, T., Hill, J., Reiser, C., Gerber, E., Styles, K., and Dillman, D. (2007). *Guidelines for Designing Questionnaires for Administration in Different Modes.* Washington DC: U.S. Census Bureau. http://www.census.gov/srd/mode-guidelines.pdf.

Mulrow, J.M., Carlson, L., Jankowski, J., Shackelford, B., and B. Wolfe, R. (2007). "A Face-Lift or Reconstructive Surgery? What Does It Take to Renew a 53-Year Old Survey?" *Proceedings of the 3rd International Conference on Establishment Surveys* (pp. 208–213), June 18–21, 2007, Montreal, Canada. American Statistical Association, Alexandria, VA.

Office of Management and Budget (OMB). (2006.) Guidance on Agency Survey and Statistical Information Collections, January. http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/pmc_survey_guidance_2006.pdf.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Rothgeb, J., and Singer, E. (2004). "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly 68(1):* 109–130.

Snijkers, G., and Willimack, D.K. (2011). "The Missing Link: From Concepts to Questions in Economic Surveys," paper presented at the 2nd European Establishment Statistics Workshop (EESWII), Neuchatel, Switzerland, Sept. 1214.

Stettler, K., and Featherston, F. (2010). "Early Stage Scoping: Building High Quality Survey Instruments without High Costs." Presentation at the International Field Directors and Technologies Conference. http://ifdtc.org/PC2010/presentation_2010_files/12B-Kristin%20Stettler.pdf.

Stettler, K., and Featherston, F. (2012). "Early Stage Scoping: Bridging the Gap between Survey Concepts and Survey Questions." Presentation at the Fourth International Conference on Establishment Surveys. www.amstat.org/meetings/ices/2012/papers/301938-A.pdf.

U.S. Census Bureau. (2012) *Statistical Quality Standards*, Reissued June, https://www.census.gov/quality/standards/index.html.

Weller, Susan, and Romney, A. Kimball. (1988). *Systematic Data Collection*. Newbury Park, CA: Sage Publications.

Willeboordse, A. (ed.) (1998). *Handbook on the Design and Implementation of Business Surveys*. Luxembourg: Eurostat.

Willimack, D.K., Lyberg, L., Martin, J., Japec, L., and Whitridge, P. (2004). "Evolution and Adaption of Questionnaire Development, Evaluation, and Testing Methods for Establishment Surveys," in Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, Hoboken, NJ: Wiley.

Question Evaluation Methods

Expert Reviews (Methodological)

Methodological expert reviews are done in order to assess potential respondent and interviewer difficulty with administering and responding to survey items (U.S. Census Bureau, 2003). The reviewers should be experienced survey methodologists with knowledge of the theoretical or practical aspects of questionnaire design. Expert review teams can also include subject matter experts. The expert review process can include a broad range of tasks from a cursory read-through to a more systematic checklist exercise. At its core, an expert review attempts to catch and resolve typical survey design errors.

Cognitive appraisals are coding systems or checklists which allow for a more systematic approach to the expert review of a survey instrument. One example is the Question Appraisal System (QAS) by Willis and Lessler (1999).

Heuristic evaluations involve the expert review of electronic instruments based on human-computer interaction principles or heuristics (Willimack, et al., 2004). An example of a heuristic principle is: "make it easy to navigate." Evaluators might conduct usability testing to see if this principle is met throughout the instrument (Hansen and Couper, 2004).

Expert reviews are typically used to check a questionnaire for major errors before the survey is either pretested or fielded. While expert reviews are usually conducted early in the questionnaire development process, they can be done at different times when resources or circumstances call for them. An expert review is often done in combination with other pretesting methods (U.S. Census Bureau, 2003). One benefit to conducting expert reviews early in the questionnaire development process is that the results of the review sometimes send researchers "back to the drawing board" with problematic questions (Willis, 2005). Expert reviewers typically describe the questionnaire's characteristics and the type of tasks that will be required of respondents. Reviewers then make a list of problems that they predict respondents will have (Forsyth and Lessler, 1991).

Cognitive appraisal is a more formal process through which codes are assigned to describe the response process and to identify problems. These codes often focus on "difficulties that respondents have in understanding questions, recalling information, and formulating responses, as well as the sensitivity of questions…" (Forsyth and Lessler, 1991; p. 398).

Strengths: Expert reviews and heuristic evaluations can be relatively inexpensive and quick to accomplish (Tourangeau, 2004; Hansen and Couper, 2004). They can eliminate major design errors that may create problems or add undue respondent burden. Conducting expert reviews prior to other evaluation methods can improve

efficiencies by identifying and fixing obvious problems and allowing the empirical evaluations to be done on the best possible version of the questions.

Limitations: Expert review, however systematic, does not provide transparent, empirical, or analyzable data, and cannot be considered a scientific method. A review can eliminate obvious errors or pitfalls on a questionnaire before a rigorous questionnaire evaluation method is employed. The method often solicits just the "opinions" of researchers with a lack of consistency across reviewers and no use of uniform evaluation criteria (Tourangeau, 2004). In addition, when experts have opposing views about a question it can be difficult to reconcile the disagreement. Another important drawback of expert review is that it does not include respondent input (U.S. Census Bureau, 2003). This can be problematic because respondents' backgrounds and life circumstances vary a great deal and it can be difficult for experts to anticipate and understand all of these variables without respondent input. Because of this limitation, some organizations view expert review as meeting only the "minimal standard" of their pretesting requirements. In some cases, expert review may be used independently of other pretesting methods only in situations where time and resource problems leave this as the only option.

References:

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

Groves, Robert M., Fowler, Floyd J., Couper, Mick P., Lepkowski, James M., Singer, Eleanor, and Tourangeau, Roger. (2004). *Survey Methodology.* Hoboken, NJ: Wiley and Sons.

Hansen, S.E., and Couper, M.P. (2004). "Usability Testing to Evaluate Computer-Assisted Instruments," in Presser, S., Rothgeb, J.M., Couper, M., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. (pp. 337–360). Hoboken, NJ: John Wiley & Sons.

Lessler, Judith T., and Forsyth, Barbara H. (1996). "A coding system for appraising questionnaires," in Schwarz, N., and Sudman, S. (eds.), *Answering Questions*. San Francisco, CA: Jossey-Bass.

Ramirez, C. (2002). "Strategies for subject matter expert review in questionnaire design." International Conference on Questionnaire Development, Evaluation and Testing Methods (QDET), Charleston, SC.

Tourangeau, R. (2004). "Experimental Design Considerations for Testing and Evaluating Questionnaires," in Presser, S., Rothgeb, J.M., Couper, M., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (eds.). *Methods for Testing and Evaluating Survey Questionnaires* (pp. 209–224). Hoboken, NJ: John Wiley & Sons.

U.S. Census Bureau. (2003). *Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses*. http://www.census.gov/srd/pretest-standards.pdf.

Willimack, D.K., Lyberg, L., Martin, J., Japec, L., and Whitridge, P. (2004). "Evolution and Adaptation of Questionnaire Development Evaluation, and Testing Methods for Establishment Surveys," in Presser, S., Rothgeb,

J.M., Couper, M., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (eds.). *Methods for Testing and Evaluating Survey Questionnaires* (pp. 385–407). Hoboken, NJ: John Wiley & Sons.

Willis, G., and Lessler, J.T. (1999). *The BRFSS-QAS: A Guide for Systematically Evaluating Survey Question Wording*. Rockville, MD: Research Triangle Institute.

Willis, G. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications, Inc.

Cognitive interviews

Definition: Cognitive interviewing studies investigate how survey questions perform when asked of respondents, that is, if respondents understand the questions according to their intended design and if respondents can provide accurate answers based on that intent. Cognitive interviewing studies determine respondent interpretations and detail the phenomena considered by respondents in forming their answer. Findings from cognitive interviewing studies can indicate whether a survey question captures the intended construct as well as difficulties that respondents experience when formulating a response.

Cognitive interviewing studies are qualitative studies. Interviews generate textual data that include explanations and examples of respondent circumstances and how those circumstances inform the question response process. Cognitive interviews consist of one-on-one, open-ended semi-structured interviews. The typical interview structure consists of respondents[2] first answering the evaluated survey question and then a series of follow-up questions that reveal what respondents were thinking and their rationale for that specific response. Common follow-up questions include: "what were you thinking when you answered the question?" and "what does the question mean to you?" Through this semi-structured design, cognitive interviewing provides rich, contextual insight into the ways in which respondents 1) interpret a question, 2) consider and weigh out relevant aspects of their experiences and, finally, 3) formulate a response based on that consideration. As such, cognitive interviewing provides in-depth understanding of the ways in which a question operates, the kind of phenomena that it captures, and if and how it ultimately serves the scientific goal.

Sample selection for a cognitive interviewing study is purposive rather than random. The intent is to select respondents who can address the objectives of the study rather than serve as a representative of the population. For example, when studying questions designed to identify persons with disabilities, the sample would likely consist of respondents with a previously known disability and, to discover potential causes of false positive or false negative reporting, some respondents with no known disability. Analysis of cognitive interviews does not produce

[2] Cognitive interviewees are also sometimes referred to as participants. The term respondent is used here to emphasize the fact that cognitive interviewees are not only responding directly to survey questions, but that they are contributing data at the same individual level as survey respondents.

generalizable findings in a statistical sense, but rather, provides insight into patterns of interpretation and the potential for measurement error.

Raw data of a cognitive interview study consists of either a video or audio recording or a written transcript of the interview. As is the case for all analyses of qualitative data, the general process involves data synthesis and reduction—beginning with a large amount of textual data and resulting in conclusions that are meaningful to the ultimate purpose of the study. Additionally, as previously described, cognitive interviewing studies can serve different purposes that pertain to question evaluation. Those purposes include:

- Identifying difficulties that respondents may experience when attempting to answer a survey question. These difficulties may occur within one of the four stages of the question response process: comprehension, retrieval, judgment, and response. Once identified, these findings can provide clues as to how a question might be improved so the recognized difficulties can be reduced.

- Identifying experiences or events that respondents consider and ultimately include or exclude in their answer to a particular question. This type of study is an examination of construct validity since it identifies the actual phenomenon captured by a survey question.

- Examining issues of comparability, for example, the accuracy of translations or equivalence across socio-cultural groups. This type of study is an examination of bias since it investigates how different groups of respondents may interpret or process questions differently.

Findings from a cognitive interviewing project typically lead to recommendations for improving a survey question. Results are also beneficial to post-survey analysis by informing data interpretation.

Strengths: The method can determine the ways in which respondents interpret questions and apply those questions to their own sets of experiences and perceptions. As such, the method identifies difficulties that survey respondents may experience when answering the question, as well as the phenomena or sets of phenomena that a variable would measure once the survey data is collected.

Limitations: While cognitive interviewing studies indicate that a particular interpretive pattern does exist, because the method is purposive and not random, it cannot determine the extent to which that pattern would occur in a survey sample. Nor can cognitive interviewing studies reveal the extent to which interpretive patterns differ would occur across groups. Additionally, the studies cannot fully determine the extent to which respondents experience difficulty when attempting to answer a question.

References:

Boeije, H., and Willis, G. (2013). The cognitive interviewing reporting framework (CIRF): towards the harmonization of cognitive testing reports. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 9(3):87–95

Miller, K., Willson, S., Chepp, V., and Padilla, J. (2014). *Cognitive Interviewing Methodology*. Hoboken, NJ: John Wiley & Sons.

Willis, G. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, California: Sage Publications, Inc.

Embedded Probing/Web Probing

Definition: Similar to respondent debriefing, *embedded probing* is a "field-based" activity in which information about the performance of items is collected within a field environment. However, whereas the probes administered in respondent debriefing are fully retrospective (in that they tend to be administered as a separate set following the completed survey interview), those employed through embedded probing are usually concurrent—that is, administered during the survey interview, immediately following the response to the item under evaluation. Embedded probing has a fairly long history in survey methods. In *random probing* (Schuman, 1966), several items are randomly selected for follow-up probing within the field interview, to gather additional information helpful for understanding the respondent's answer. However, embedded probes are more typically applied strategically to only those specific questions that the designers are particularly interested in evaluating (Converse and Presser 1986, Baena and Padilla 2014).

Embedded probes can be read by a field interviewer, in the case of an interviewer-administered survey, or can be included within a self-administered survey instrument. Increasingly, the embedding of probes has taken the form of *web probing*, in which the respondent is asked to elaborate on his or her response to a prior survey item by typing an explanation within a dedicated text box (Behr, Braun, Kaczmirek, and Bandilla, 2014; Murphy, Edgar, and Keating, 2014). For example, Murphy et al. (2014) evaluated the item "Since the first of May, have you or any other member of your household purchased any swimsuits or warm-up or ski suits?" by then administering the open-ended probe: "What types of items did you think of when you read this question?" Embedded probing can either be incorporated into a final version of a questionnaire, or it can be included within a pilot or pretest version of the questionnaire.

Strengths: As in respondent debriefing, embedded probing allows researchers to maintain the ecological validity of their evaluation studies by collecting information about the respondents' survey response processes *in situ*. Further, this method allows for a large number of responses, which provides two substantial benefits. First, unlike purely qualitative methods, embedded probing can be incorporated into statistical samples, thus allowing the evaluative

results to be extrapolated to the survey population. Secondly, a large number of responses—sometimes collected at a very low cost—allows for a full range of coding and analysis activities, in which response tendencies can be ascertained for each of several demographic or other defined sub-groups (e.g., cultural, racial, ethnic, etc.).

Limitations: It is feasible to conduct embedded probing for only a small subset of survey items – generally no more than a half-dozen – without unacceptably increasing respondent burden and questionnaire administration time. Therefore, investigators can target only a few, carefully selected items for which to embed probes. Further, probes must be scripted in order to be embedded in a questionnaire, resulting in a rigid and inflexible approach in which the probes cannot be worded in a flexible manner appropriate to the individual interview, and no follow-up probing can be done, as in a more traditional form of cognitive testing (Willis, 2015). Finally, more research is necessary on the validity of the embedded probes themselves, and how respondents answer them (especially self-administered web probes that require open-ended typing).

References:

Baena, I.B., & Padilla, J.L. (2014). Cognitive Interviewing in Mixed Research. In Miller, K., Willson, S., Chepp, V. and Padilla, J.L. (editors), *Cognitive Interviewing Methodology*. Hoboken, NJ: Wiley.

Behr, D., Braun, M., Kaczmirek, & Bandilla, W. (2014). Item Comparability in Cross-National Surveys: Results From Asking Probing Questions in Cross-National Surveys About Attitudes Towards Civil Disobedience. Quality and Quantity, 48(1), 127-148.

Converse, J. M., & Presser, S. (1986). Survey Questions: Handcrafting the Standardized Survey Questionnaire. Newbury Park, CA: Sage.

Murphy, J., Edgar, J., & Keating, M. (2014, May). Crowdsourcing in the Cognitive Interviewing Process. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Anaheim, CA.

Schuman, H. (1966). The Random Probe: A Technique for Evaluating the Validity of Closed Questions. American Sociological Review, 31(2), 218–222.

Willis, G. (2015). Analysis of the Cognitive Interview in Questionnaire Design. New York: Oxford.

Vignette Studies and Fictional Scenarios

Definition: Vignettes and fictional scenarios are increasingly being used in survey research to evaluate questions or responses to questions. Vignettes describe hypothetical situations or objects, which respondents are asked to judge (Martin, 2004). For example, the following vignette describes a living situation to explore respondents' definition of residency:

> Mary asked her friend Helen if she could stay with her for a few days while she looked for a place of her own. It has been five months since then. Mary's suitcases are still packed, and are at the front door. Should Helen count Mary as usually living there? (Gerber et al., 1996)

The situations or objects in vignettes can take on different dimensions, the values of which are often varied, so that their impact on respondents' judgments can be assessed. To describe residency to respondents, Tourangeau et al. 2006 varied whether a person in their vignette contributed toward the rent; whether he or she slept at the residence; and whether he or she was related to household members. Ultimately, the purpose is to determine whether respondents' understanding of survey concepts are consistent with survey definitions.

Hypothetical situations that include additional materials, such as floor plans, purchase receipts, business records, or video clips and that ask respondents to role-play, are often called fictional scenarios (see, for example, Schober and Conrad, 1997). Since the researchers know the "correct" answer in these cases, the accuracy of respondents' answers can be assessed.

Vignettes are also used by researchers to elicit respondents' opinions under hypothetical situations that vary along a dimension. For example, to determine how willing respondents would be to allow their private information to be shared between governmental agencies, Smirnova and Scanlon (2013) altered the description of benefits that would be realized from the sharing of information.

Finally, anchoring vignettes are used when measuring concepts, like political efficacy, for which there is no single commonly held metric. In these instances, a self-assessment question might be asked: "How much say do you have in getting your local government to consider issues that interest you?" followed by a vignette and a second assessment question regarding the character from the vignette (Hopkins and King, 2010). Responses to the second assessment question are then used to provide a common reference point for rescaling the self-assessment questions.

Vignette studies are typically attached to fielded surveys in the pre-testing phase and are usually analyzed quantitatively. However, they can also be embedded into cognitive interview or focus groups and be analyzed qualitatively.

Strengths: By asking all respondents to assess the same description—regardless of their personal experience with a construct—these studies can provide insight into respondents' conceptualizations of survey concepts, opinions, or self-assessments with efficiency and power. Vignettes enable researchers to control and present situations, the likes of which would be difficult, if not impossible, to recruit for. In addition, the accuracy of respondents' answers can be assessed when researchers provide respondents with fictional scenarios from which to answer questions. Furthermore, since attitudes and opinions can change in response to contextual variables, vignettes allow researchers to isolate attitudes and opinions by controlling for extraneous variables. Finally, vignettes allow for the calibration of respondents' self-assessments, which potentially makes comparisons across different subgroups (and cultures) more valid.

Limitations: The results of administering vignettes to respondents are not direct measures of their behavior (Martin, 2004). To the extent respondents' behaviors differ from their hypothetical assessments, the hypothetical

assessments will be biased. Also, vignettes have been shown to suffer from the same order and wording problems as the questions they are designed to study (Auspurg and Jäckle, 2012; Buckly, 2008; Hopkins and King, 2010). Furthermore, it cannot be assumed that respondents will understand vignettes the same way across cultures and language groups (Goerman and Clifton, 2011). Thus vignettes too need to be evaluated in order to assure that they are adequate measuring devices in a given context.

References:

Auspurg, K., and Jäckle, A. (2012). First Equals "Most Important? Order Effects in Vignette-Based Measurement," ISER Working Paper Series, No. 2012-01, http://hdl.handle.net/10419/65949.

Beck, J. (2010). *On the Usefulness of Pretesting Vignettes in Exploratory Research*. Research Report Series (Survey Methodology #2010-02).

Buckley, J. (2008). "Survey Context Effects in Anchoring Vignettes." Working paper. http://polmeth.wustl.edu/workingpapers.php.

Gerber, E., Wellens, T.R., and Keeley, C. (1996). "Who Lives Here? The Use of Vignettes in Household Roster Research." *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 962–967.

Goerman, P., and Clifton, M. (2011). "The Use of Vignettes in Cross-Cultural Cognitive Testing of Survey Instruments." *Field Methods* 23(4):362–378.

Hopkins, D. J., and King, G. (2010). "Improving Anchoring Vignette: Designing Surveys to Correct Interpersonal Incomparability." *Public Opinion Quarterly* 74:201–222.

Martin, E. (2004). "Vignettes and Respondent Debriefing for Questionnaire Design and Evaluation," in Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J. and Singer, E., *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken, NJ: John Wiley & Sons.

Schober, M.F., and Conrad, F.G. (1997) "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly* 61:576–602.

Smirnova, M., and Scanlon, P. (2013). "To Share or Not to Share? Cognitive Interviewing and Focus Groups in Eliciting Privacy and Confidentiality Concerns." Unpublished Manuscript.

Tourangeau, R., Conrad, F., Arens, Z., Fricker, S., Lee, S., and Smith, E. (2006). "Every day Concepts and Classification Errors: Judgments of Disability and Residence." *Journal of Official Statistics*, 22:385–418.

Usability Testing

Definition: Usability testing is a methodology for evaluating data collection instruments, including computerized self-administered questionnaires and web survey instruments. It employs techniques to collect empirical data while observing representative end users (such as survey respondents or field interviewers) using the product (in this case, the electronic instrument) performing representative tasks. The goals of usability testing are to: 1) collect

quantitative and qualitative data as users complete tasks, 2) determine users' satisfaction with the product, 3) identify any usability problems, and 4) identify changes to improve user performance.

Researchers ask participants to complete tasks typically encountered in using the instrument, and then observe them as they perform actions such as entering data, navigating skip patterns, correcting errors, and printing or submitting completed forms. During this time, the researchers will collect quantitative data such as task time and number of errors, along with qualitative data, such as their own observations and participants' comments. At the end of the test session, researchers usually ask participants to respond to some debriefing questions about their experience. Researchers analyze all these data to identify usability problems and potential solutions. Behaviors of respondents interacting with the instrument may also be coded, supporting empirical analyses of usability problems.

Eye-tracking research uses special equipment to detect exactly where people's eyes are focused when they look at a computer screen. This form of usability research makes it easier to understand what part of an electronic instrument users are attracted to, and which parts they tend to overlook.

Testing may begin with paper prototypes or nonfunctioning screen shots, called low-fidelity prototypes. Results from this testing may then be used to assist in the design of the system. Generally, most usability testing is conducted with partially or fully functioning, or high-fidelity, electronic prototypes.

Usability testing can take place either in a usability laboratory or in the field at the participant's home or work site. Lab studies are easier for the researcher to manage and control, but field studies allow the researcher to observe participants in their natural environment, to study the atmosphere of the home or workplace, and to see how the environment affects the reporting task.

Cognitive methods, such as thinking aloud, concurrent probes, retrospective debriefings, user or respondent ratings, and vignettes are often incorporated in usability tests.

Strengths: The method can be performed in a relatively short period of time, allowing for quick turnarounds. The method allows researchers to observe actual users and get feedback from them. Done early in the development process, usability testing can provide results that will avoid costly design mistakes.

Limitations: Usability testing is sometimes conducted in a simulated environment, which may affect the participant's behavior or reactions, possibly limiting the generalizability of the results.

References:

Dumas, J.S., and Redish, J.C. (1999). *A Practical Guide to Usability Testing*. Norwood, NJ: Intellect.

Nielsen, J., and Pernice, K. (2010). *Eye Tracking Web Usability*. Berkeley, CA: New Riders.

Rubin, J. and Chisnell, D. (2008). *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests.* Hoboken, NJ: John Wiley & Sons, Inc.

Snijker, G. and Haraldsen, G. (2013). *Designing and Conducting Business Surveys*. John Wiley & Sons, Hoboken, NJ.

Feedback from Survey Personnel

Once a survey has been fielded, survey methodologists can seek feedback from a variety of sources. Aside from respondent debriefings, which are discussed at greater length in a different section of this report, debriefings are also often done with field interviewers in the case of household surveys and with survey analysts and help desk personnel or other staff involved in the data collection process in the case of establishment surveys (U.S. Census Bureau, 2003).

Both interviewer and analyst debriefings can be carried out via focus groups, rating forms or structured questionnaires. Interviewers or analysts are asked to describe their assessment of problems they encountered while administering survey items or problems they noticed respondents having with particular questions (U.S. Census Bureau, 2003). Interviewer debriefings are often done in the context of interviewer-administered household surveys. They can also be done to evaluate field or pilot tests or in order to evaluate periodic surveys prior to an update or redesign.

Feedback from survey analysts and help desk personnel is often sought in the context of establishment surveys. This is because these surveys are typically self-administered and some survey analysts come into direct contact with respondents during data collection. Survey analysts and help desk personnel can both receive incoming respondent requests for help and contact respondents to follow up on missing data or unclear responses. Feedback is sought from those analysts who have direct contact with respondents.

Strengths: Interviewer debriefing can be useful in shedding light on respondent experience and difficulty with survey items. Feedback from survey analysts and help desk personnel can be examined through recorded phone logs, which reflect the relative frequency of particular respondent problems and can make qualitative analysis richer.

Limitations: Interviewer debriefing alone does not meet the "minimal pretesting standard" at some agencies, since it is difficult to quantify the relative frequency of a given problem that an interviewer may report via this method. Interviewers also may sometimes focus on their own preferences with regards to question wording, rather than any issues they've observed respondents to be having. Experienced interviewers also sometimes tend to "correct" for problematic question wording and may not even realize that they are doing so in order to be able to report this behavior during the feedback (U.S. Census Bureau, 2003).

Focusing only on interviewer perception of problems has drawbacks, since problems can exist that respondents are either unaware of or that they do not signal while answering the questions. In addition, interviewers' signaling of problems does not suggest ways to resolve the problems (Presser, et al., 2004).

In terms of establishment surveys, feedback from survey analysts and help desk personnel can be anecdotal when gathered from analysts' memory alone.

References:

Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J. and Singer, E., *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken, NJ: John Wiley & Sons.

U.S. Census Bureau (2003). *Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses.* http://www.census.gov/srd/pretest-standards.pdf.

Respondent Debriefing

Definition: Respondent debriefing is a retrospective field probing technique, where survey respondents are re-contacted following field data collection and asked about the response strategies and data sources actually used to provide answers to specific questions. Thus, the degree to which the collected data meet the intent of the survey question(s) can be assessed against actual respondent behavior. A typical purpose for conducting respondent debriefings is to evaluate the quality of collected data.

One common goal of respondent debriefing is to ascertain question meaning in the context of actual administration of the survey. Following survey completion by the respondent, survey researchers conduct interviews using probe questions to investigate cognitive processes actually undertaken by respondents–interpretation, retrieval, judgment, and reporting–in answering the survey questions. Not only can survey researchers assess discrepancies relative to the question's intent, but they may also identify sources of error to be examined and addressed when revising questions. Another goal may be to determine respondents' subjective experience and reactions to taking the survey, for example, whether the content was too sensitive or the instrument too lengthy and burdensome.

Respondent debriefings are typically applied in conjunction with other methods where data are collected from respondents under field conditions. Sample designs vary and may provide either qualitative or quantitative results depending on research goals. Qualitative respondent debriefing studies typically rely on small purposive samples, as few as 20 cases. Respondent debriefing studies that support quantitative results are usually designed as formal response analysis surveys (described in the next section). In-person interviews are the preferred mode for conducting qualitative respondent debriefing studies, where researchers follow a protocol using retrospective focused interviewing techniques to identify reporting errors.

Strengths: Since respondent debriefings are conducted following field data collection, the respondent's actual response strategy can be uncovered. That is, the respondent will have completed a self-administered questionnaire or survey interview under typical field conditions, rather than the somewhat artificial conditions of pre-field cognitive or exploratory interviews.

Limitations: Since respondent debriefings rely on respondents' recall of their actual response processes, debriefing interviews must be conducted temporally as near response as possible. This may make it difficult to plan and manage the study. Also, the response process may not be sufficiently salient for respondents to recall with accuracy. In addition, small sample sizes or purposive sample designs preclude generalizability of results.

References:

Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses. http://www.census.gov/srd/pretest-standards.html.

DeMaio, T. (ed). (1983). "Approaches to Developing Questionnaires," Statistical Policy Working Paper 10, Federal Committee on Statistical Methodology, Washington, DC.

DeMaio, T., and Rothgeb, J. (1996). "Cognitive Interviewing Techniques: In the Lab and in the Field," in Schwarz, N., and Sudman, S. (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research.* San Francisco, CA: Jossey-Bass.

Martin, E. (2004). "Vignettes and Respondent Debriefing for Questionnaire Design and Evaluation," in Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J. and Singer, E., *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken, NJ: John Wiley & Sons.

Willis, G. (2005). *Cognitive Interviewing: A Tool for Improving Question Design.* Thousand Oaks, CA: Sage Publications, Inc.

Response Analysis Surveys

Definition: Respondent debriefings that are formalized through use of a structured questionnaire are known as response analysis surveys (RAS) (Utter, 1983). They are conducted following data collection. Quantitative summaries are typically generated, evaluating the extent to which reported data meet the definitions, along with respondents' use of records and estimation strategies.

Although commonly associated with pilots of new or redesigned surveys, RASes may also be used to evaluate data quality in recurring surveys. (Goldenberg, 1994; Goldenberg and Phillips, 2000; Goldenberg, Butani, and Phipps, 1993; Palmisano, 1988; Phipps, 1990; Phipps, Butani, and Chun, 1995). Depending on research goals, random samples may be used, or sample selection may be purposive to focus on respondents who made reporting errors (Phipps, 1990) or to compare early and late respondents (Ware-Martin, Adler, and Leach, 2000).

Since they utilize a structured questionnaire, RASes are typically based on a statistical sample selected from a

defined target population, thus supporting statistical summaries, analyses and inferences from collected data. RAS structured interviews are commonly conducted via telephone.

Strengths: In addition to the strengths associated with respondent debriefings, RASes permit researchers to quantify behaviors and strategies associated with the response process. If they are based on a random sample, statistical inferences are also supported.

Limitations: In addition to the limitations associated with respondent debriefings, RASes increase data collection costs and respondent burden when conducted on a broad enough scale or with large sample sizes.

References:

Goldenberg, K.L. (1994). "Answering Questions, Questioning Answers: Evaluating Data Quality in an Establishment Survey," *Proceedings of the Section on Survey Research Methods,* (pp. 1357–1362), American Statistical Association, Alexandria, VA.

Goldenberg, K.L., and Phillips, M.A. (2000). "Now That the Study is Over, What Did You Really Tell Us? Identifying and Correcting Measurement Error in the Job Openings and Labor Turnover Survey Pilot Test," *Proceedings of the Second International Conference on Establishment Surveys,* contributed papers only available on CD, American Statistical Association, Alexandria, VA.

Goldenberg, K.L, Butani, S., and Phipps, P.A. (1993). "Response Analysis Surveys for Assessing Response Errors in Establishment Surveys," *Proceedings of the International Conference on Establishment Surveys* (pp. 290–299), American Statistical Association, Alexandria, VA.

Kydoniefs, L. (1993). "The Occupational Safety and Health Survey," *Proceedings of the International Conference on Establishment Surveys* (pp. 99–106), American Statistical Association, Alexandria, VA.

Palmisano, M. (1988). "The Application of Cognitive Survey Methodology to an Establishment Survey Field Test," *Proceedings of the Survey Research Methods Section* (pp. 179–184), American Statistical Association, Alexandria, VA.

Phipps, P.A. (1990). "Applying Cognitive Theory to an Establishment Mail Survey," *Proceedings of the Section on Survey Research Methods* (pp. 608–612), American Statistical Association, Alexandria, VA.

Phipps, P.A., Butani, S.J., and Chun, Y.I. (1995). "Research on Establishment-Survey Questionnaire Design," *Journal of Business and Economic Statistics* 13:337–346.

Utter, C.M. (1983). "Response Analysis Surveys," Chapter 11 in DeMaio, T. (ed.), Approaches to Developing Questionnaires, Statistical Policy Working Paper 10, Federal Committee on Statistical Methodology, Washington, DC.

Ware-Martin, A., Adler, R.K., and Leach, N.L. (2000). "Assessing the Impact of the Redesign of the Manufacturing Energy Consumption Survey," *Proceedings of the Second International Conference on Establishment Surveys* (pp. 1488–1492), American Statistical Association, Alexandria, VA.

Willimack, D.K., Lyberg, L., Martin, J., Japec, L., and Whitridge, P. (2004). "Evolution and Adaption of Questionnaire Development, Evaluation, and Testing Methods for Establishment Surveys," in Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J. and Singer, E., *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken, NJ: John Wiley & Sons.

Randomized Experiments

Definition: Randomized experiments use rigorous procedures to explore the effect of manipulating variables (Shadish, Cook, and Campbell 2002). These procedures usually include having well-defined objectives (hypotheses for predicting the effect of manipulating the variables), adequate precision (adequate sample size for measuring the effect), the ability to estimate sampling error, and the ability to distinguish the various effects of manipulating the variables.

Experiments used to evaluate questions are often called split-ballot (or split-sample) experiments. Split-ballot experiments produce quantitative data that can be analyzed using statistical methods. Ideally, split-ballot experiments are used to complement or to follow-up qualitative methods (Fowler, 2004). Split-ballot experiments compare the original version of the question (or questionnaire) to alternative versions of a question (or questionnaire). These alternative versions typically incorporate proposed solutions based on problems identified from the qualitative methods. Such experiments can aid in determining if issues uncovered by qualitative methods actually impact survey estimates and whether the alternative versions of the question (or questionnaire) improve data quality.

Randomized experiments are also used to bridge or calibrate the differences between an original version of a question (or questionnaire) and a redesigned version in repeated or longitudinal surveys. For example, in the Current Population Survey (CPS), the old version of the CPS was compared to the final redesigned version to determine the effect of the overall redesigned version on labor force estimates (Polivka, 1994). Using experiments to calibrate across surveys maintains the ability of researchers to conduct trend analyses.

Randomized experiments can be embedded into the production environment, that is, they can be embedded into a field pretest, pilot test, field test, or actual production survey (refer to Moore et al. (2004) for examples in the Survey of Income and Program Participation). The primary goal of these surveys is to produce estimates for data uses, and the secondary goal is to evaluate the questions through the use of experiments. Randomized experiments can also be conducted outside the confines of a production survey. For example, Tourangeau et al. (2011) designed numerous omnibus surveys for the express purpose of conducting experiments on survey questions. The primary goal of the latter approach is to evaluate the questions; they do not have the goal of producing estimates for data users as their objective.

Strengths: The act of randomization controls for bias between treatment (or comparison) groups. Thus, randomized experiments allow for a causal description of effects; that is, they describe the consequences attributable to deliberately varying a treatment. When single items vary in one way, the effect can be attributable to that variable. Such designs have the potential to further the science of questionnaire design.

Limitations: Randomized experiments are often expensive and time-consuming, especially those conducted under production survey conditions. Thus, researchers often package variables together between treatments, that is, two or more individual variables are allowed to vary at once. The problem with this is that the contribution of the individual variables and their interaction to the overall effect cannot be parsed out. The individual variables are confounded. In addition, differences may exist between treatments, but it may be difficult to identify which of the treatments actually leads to a better understanding and more valid data without the use of triangulating data. Finally, experiments may suffer from the same limitation as all methods for evaluating survey data, that is, inference to a target population is possible only if a statistical sample with sufficient statistical power was drawn from the target population for conducting the experiment.

References:

Fowler, F. J. (2004). "The Case for More Split-Sample Experiments in Developing Survey Instruments," in Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T, Martin, E., Martin, J., and Singer, E. (eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 173–188). New York: John Wiley and Sons.

Moore, J., Pascale, J., Doyle, P., Chan, A. and Klein Griffiths, J. (2004). "Using Field Experiments to Improve Instrument Design: The SIPP Methods Panel Project," in Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T, Martin, E., Martin, J., and Singer, E. (eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 189–208). New York: John Wiley and Sons.

Polivka, A.E. (1994). "Comparisons of Labor Force Estimates from the Parallel Survey and the CSP During 1993: Major Labor Force Estimates," in *CPS Overlap Analysis Team Technical Report 1*. Washington, DC: Bureau of Labor Statistics. http://www.bls.gov/osmr/abstract/cp/cp940100.htm.

Redline, C. (2013). "Clarifying Survey Questions in a Web Survey," *Public Opinion Quarterly* 77, Special Issue: 89–105.

Shadish, W.R., Cook, T.D., and Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Snedecor, G.W., and Cochran, W.G. (1989). *Statistical Methods*, 8th ed. Ames, Iowa: Iowa State University Press.

Tourangeau, R. (2004). "Experimental Design Considerations for Testing and Evaluating Questionnaires," in Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J. and Singer, E., *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken, NJ: John Wiley & Sons.

Tourangeau, R., Couper, M.P., and Conrad, F.G. (2013). "'Up Means Good': The Effect of Screen Position on Evaluative Rating in Web Surveys," *Public Opinion Quarterly* 77, Special Issue: 69–88.

Validation Studies

Definition: Validation studies are studies in which data external to the survey of interest are used to evaluate the validity and accuracy of survey data. These external data may be from official registers, other surveys, administrative records, or other available sources. In these studies, the external validation data are generally considered of good quality and thus a reasonable proxy for the true values. These data are then compared to survey data and discrepancies are identified and evaluated.

Validation studies can be done at a macro-analysis level or a micro-analysis level. At the macro-analysis level, a survey estimate is compared to a "gold standard" or benchmark for that estimate to determine if there is measurement error in the survey estimate. If large errors are found, a micro-analysis validation study can be done to try to determine potential reasons or sources for the errors or how the errors may vary by subgroup. In the micro-analysis validation, record linkage is used to compare individual record level data from the survey to the individual record level data from the benchmark. Examination of the type of differences, the magnitude of differences, or the distribution of differences may suggest possible reasons for errors. For example, administrative records from government programs can be matched to survey reports of receipt of benefits. Underreporting or overreporting of program benefits in survey reports can be identified if survey respondents' data do not match the administrative records.

Measurement errors identified in these studies can be the result of poorly designed questionnaires, but may also be the result of poorly designed data collection procedures, or respondent errors. Thus, validation studies are not exclusively focused on questionnaire evaluation.

Strengths: Both macro-analysis and micro-analysis validation studies can be used to provide an empirical indication of survey measurement error, which is often impossible to do with survey data alone. If data are significantly under- or overreported on the survey, this can suggest that potential improvements to question wording may produce more accurate answers (for example, the inclusion of more salient examples of potential misreporting, additional reporting instructions, or definitions).

Limitations: The validation data for macro-analysis or micro-analysis validation studies are usually considered a "gold standard" and assumed to be error free and matched correctly to individual survey records. In addition, the definition for the item(s) of interest must be the same for both the survey and the external validation data to allow for valid comparisons between the two data sources. The extent to which these assumptions are true determines how useful the method is for evaluating the survey data. When using macro-analysis, the comparison of survey data

to validation data alone cannot pinpoint the cause of the discrepancies between them. In fact, the differences in estimates may be due to multiple sources of error, such as nonresponse, coverage, processing errors, etc.

Micro-analysis may neither be possible nor a useful evaluation if record-level data for the survey items of interest are not available for all (or nearly all) of the survey respondents. Micro-analysis may also not provide clear evidence of underlying causes of errors in survey estimates, but may simply provide more specific information and suggest hypotheses that require more in-depth study. In addition, since surveys are often conducted to provide timely estimates, these external data used for validation may not be available until well after the survey period.

References:

Blumberg, S. J., & Cynamon, M. (1999). "Misreporting Medicaid Enrollment: Results of Three Studies Linking Telephone Surveys to State Administrative Records," Proceedings of the Seventh Conference on Health Survey Research Methods, DHHS Publication No. (PHS) 01-1013. Hyattsville, MD: Department of Health and Human Services.

Davern, M. E., Call, K. T., Beebe, T. J., Bland, P., Ziegenfuss, J., and Blewett, L. A. (2008). "Validating Health Insurance Coverage Survey Estimates: A Comparison between Self-Reported Coverage and Administrative Data Records," *Public Opinion Quarterly* 72:241–259.

Fulton, J., and Kreuter, F. (2010). "Evaluation of Administrative Record Quality in the U.S. Federal Statistical System," *Proceedings of the Section on Survey Research Methods, American Statistical Association*,

Goldstein, K.P., Kviz, F.J., and Daum, R.S. (1993, Nov. 10). "Accuracy of immunization histories provided by adults accompanying preschool children to a pediatric emergency department," *Journal of the American Medical Association* 270(18):2190–2194.

Kalsbeek, W., Weigle, K., Allred, N., and Liu, P. (1991). "A comparison of survey designs for estimating childhood immunization rates," in *Proceedings of the Section on Survey Research Methods* (pp. 175–180). American Statistical Association, Alexandria, VA.

Resnick, Dean, Love, S., Taeuber, C., and Staveley, J.M. (2004). "Analysis of ACS Food Stamp Program Participation Underestimate," paper presented at the 2004 American Statistical Association, Joint Statistical Meetings, Toronto, ON.

Re-interview / Content Evaluation

Definition: Re-interview is a method by which respondents are asked the same question at different points in time to see if their answers "match up." The method makes the assumption that a respondent's answer to a survey question should be stable over time. Re-interview studies are often done by taking a random sub-sample of the original survey sample and administering the questions to them again (Biemer, 2004). The primary goal is generally

to estimate bias due to reporting errors (Willimack, et al., 2004). Re-interviews are also sometimes known as "content evaluations" (Corby, 1984, 1987; Van Nest, 1987).

Re-interview is a method for evaluating the extent of response error in a survey item. When respondents answer the same question differently at different points in time it can show that the item has suspect validity (Willis, 2005; Albright, et al., 2000). Re-interview is often used to check for interviewer falsification but it can also be used to estimate simple response variance and response bias (Forsman and Shriener, 1991). Due to cost and other factors, this method is typically used as a question evaluation procedure, as opposed to a pretesting method (Tourangeau, 2004; Willis, 2005).

In government surveys, re-interviews are typically done with a different interviewer than the one used in the primary survey administration. The second interviewer is often a supervisor and the interview is done via a less expensive mode of data collection, such as computer-assisted telephone interview (CATI). A smaller number of questions is also typically asked than in the original interview. Re-interviews often take place about 2 weeks after the initial interview. High levels of variance are usually considered an indicator that further work on a particular question is needed (Groves, 1999).

Strengths: Re-interview can fill in a gap left by cognitive interviewing, in that it provides a measure of reliability of a survey item (Willis, 2005). Re-interview can also be very useful in estimating and reducing measurement errors and results can lead to significant improvement in data quality (Forsman and Shriener, 1991).

Limitations: In order to examine the issue of bias due to reporting errors, re-interview samples must be sufficiently large (Willimack, et al., 2004). For this reason, the method can be costly. Additional limitations include the fact that respondents' answers may change over time for legitimate reasons and different responses may not always be an indication of poor question performance. Another issue is that when selecting a household as part of a sample, different people may be interviewed across survey administrations. Finally, if the time elapsed between administrations of the question is not long, it is possible that a respondent will remember his or her initial answer and provide the same response accordingly (Willis, 2005).

References:

Albright, K.A., Reichart, J.W., Flores, L.R., Moore, J.C., Hess, J.C., and Pascale, J. (2000). "Using Response Reliability to Guide Questionnaire Design," in *Proceedings of the Section on Survey Methods Research* (pp. 157–152), American Statistical Association, Alexandria, VA.

Corby, C. (1984). *Content Evaluation of the 1977 Economic Censuses (DE-2)*, Statistical Research Report 84/29. Washington, DC: US Bureau of the Census.

Corby, C. 1987. "Content Evaluation of the 1982 Economic Censuses: Petroleum Distributors," in *1982 Economic Censuses and Census of Governments: Evaluation Studies* (pp.27–50). Washington DC, U.S. Bureau of the Census.

Forsman, G., and Schreiner, I. (1991). "The Design and Analysis of Reinterview: An Overview," in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys* (pp. 279–301). New York: John Wiley & Sons.

Groves, R.M. (1999). "Survey Error Models and Cognitive Theories of Response Behavior," in Sirken, M.G., Herrmann, D.J., Schechter, S., Schwartz, N., Tanur, J.M. and Tourangeau, R. (eds.), *Cognition and Survey Research* (pp. 235–250). New York: John Wiley & Sons.

Tourangeau, R. (2004). "Experimental Design Considerations for Testing and Evaluating Questionnaires," in Presser, S., Rothgeb, J.M., Couper, M., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (eds.) *Methods for Testing and Evaluating Survey Questionnaires* (pp. 209–224). Hoboken, NJ: John Wiley & Sons.

Van Nest, J. (1987). "Content Evaluation Pilot Study," in *1982 Economic Censuses and Census of Governments: Evaluation studies*. Washington, DC: US Bureau of the Census.

Willimack, D.K., Lyberg, L., Martin, J., Japec, L., and Whitridge, P. (2004). "Evolution and Adaptation of Questionnaire Development Evaluation, and Testing Methods for Establishment Surveys," in Presser, S., Rothgeb, J.M., Couper, M., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 385–407). Hoboken, NJ: John Wiley & Sons.

Willis (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications, Inc.

Analysis of Paradata

Definition: This method examines auxiliary survey data called paradata. Paradata are data about *the process* of producing a survey statistic, and are distinct from the substantive data the survey primarily is designed to collect. Paradata can be captured at every stage of that process–from sampling frame generation, to case management, through data collection and processing–and provide detailed information (e.g., at the respondent, interviewer, or item level) that can be used to evaluate the quality of the survey operations and the data the survey collects (Kreuter, 2013). Examples of paradata include interviewer observations about respondent, household or neighborhood characteristics, contact history records detailing the timing and outcome of contact attempts, keystroke data from electronic instruments (e.g., recording time stamps, movements between items or screens), edit failures, etc. Paradata are generated within the survey process and tend to be recorded at a highly granular level (e.g., at the contact-attempt level), though more aggregate-level paradata are possible (e.g., case-level notes; interviewer-level response rates).

Paradata analyses support a variety of evaluative objectives, but in the context of question evaluation, subject-matter researchers typically analyze paradata after data collection has ended to identify potential problems with

question design and administration. For example, they use paradata from web-based and computer-assisted interviewer-administered surveys to look for signs that respondents may be having difficulty with an item (e.g., if they request definitional help, trigger error messages, or take an inordinate amount of time before responding), or problems with interviewing protocols (e.g., variation in the use of follow-up probes).

Paradata analyses can support evaluations of question revisions (e.g., through pre-/post-redesign comparisons) or experimental manipulations that vary question protocols, and thus can be used to develop (or refine) survey methods. These analyses could involve question or questionnaire evaluation through a variety of relevant corrective actions such as rephrasing a particular question, reordering of related questions, or changing survey mode. Paradata analyses may be quantitative (e.g., statistical analyses of item response times, number of mouse clicks, number of contact attempts, etc.) or qualitative (e.g., coding and tabulations of interviewer notes). Analysts typically explore paradata through frequency distributions and control charts of data collection processes to monitor survey progress, data quality, and production. Statistical models can include both survey data variables and paradata to investigate predictors of error in survey response.

Increasingly, real-time paradata are collected and used to manage ongoing data collection operations in an effort to locate sources of errors, improve data quality and minimize survey costs (Kreuter, Couper and Lyberg, 2010; Laflamme, Maydan, and Miller, 2008). Depending on the survey design used, adjustments can be applied quickly or at a later phase of data collection (e.g., Groves and Heeringa, 2006; Couper and Wagner, 2011). Typically, these paradata-based adjustments are made to address concerns about nonresponse bias, to ensure adequate representation among target groups, or to manage collection costs, but they may also be aimed at reducing measurement error (e.g., Calinescu, Schouten, and Bhulai, 2012).

Strengths: Paradata provide an additional source of information to enrich the core content of a survey. Publicly available paradata can complement core survey variables in analysis of substantive areas of interest by survey data users. One other strong feature of paradata rests on the ease of producing them in large quantity, and with minimal burden to the survey respondent. Most surveys now use computerized data collection systems that automatically record information about the survey process such as time stamps and audit trails of keystrokes during each interview session.

Limitations: It can be difficult to organize and structure the vast amounts of paradata collected, particularly from computer-assisted surveys. Moreover, analysis of paradata can be a challenge because they often are generated at a finer level than the survey data (e.g., there can be multiple call attempts recorded for a single case, or multiple respondent actions taken to produce an answer to a single item on a web survey) and differences may exist across paradata within a given survey in their levels of granularity. This makes it difficult to know the optimal level of aggregation that should be used when preparing those data for analysis (Kreuter, 2013). Some empirical work has been done in this area (e.g., Yan and Olson, 2013), but additional research and practical application is needed. Another challenge in the analysis of paradata is the quality of data themselves, particularly those that involve

manual entry or subjective judgments. Although a comparative analysis of three federal surveys concluded some confidence on paradata, the evidence is confined to the contact history variables examined (Bates, N., Dahlhamer, J., Phipps, P., Safir, A., and Tan, L., 2010). Further study of paradata quality is warranted (see, e.g., West and Sinibaldi, 2013). Finally, to the extent that these data are released to the public, security concerns exist (Nicolaas, 2011). Data confidentiality is an issue since there are interviewers and respondents involved. As the amount of core data and paradata released to the public increases, the level of risk for respondent identification also increases (Deang and Davies, 2009). This limitation is distinct in itself although it is related to data organization and data structure in files intended and released for public use.

References:

Bates, N., Dahlhamer, J., Phipps, P., Safir, A., and Tan, L. (2010). "Assessing Contact History Paradata Quality Across Several Federal Surveys," Section on Survey Research Methods, Joint Statistical Meetings, Vancouver, BC.

Calinescu, M., Schouten, B. and Bhulai, S. (2012). "Adaptive Survey Designs That Minimize Nonresponse and Measurement Risk," discussion paper, The Hague, Netherlands: Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/0515BE43-D5D4-42B9-8B26-755E1E10CC72/0/201224x10pub.pdf.

Couper, M. P., and Wagner, J. (2011). "Using Paradata and Responsive Design to Manage Survey Nonresponse." Ann Arbor, MI: University of Michigan Institute for Social Research.

Deang, L.P., and Davies, P.S. (2009). "Access Restrictions and Confidentiality Protections in the Health and Retirement Study," Research Note No. 2009-01, Office of Research, Evaluation, and Statistics, US Social Security Administration, paper presented at the 2008 Federal Committee on Statistical Methodology Statistical Policy Seminar, "Beyond 2010: Confronting the Challenges," November 18–19, 2008, Washington DC.

Groves, R.M., and Heeringa, S.G. (2006). "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Nonresponse and Costs," *Journal of the Royal Statistical Society: Series A (Statistics in Society*), 169(3):439–457.

Kreuter, F. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. John Wiley & Sons.

Kreuter, F., Couper, M., and Lyberg, L. (2010). "The Use of Paradata to Monitor and Manage Survey Data Collection," Section on Survey Research Methods, Joint Statistical Meetings Vancouver, BC.

Laflamme, F., Maydan, M., and Miller, A. (2008). "Using Paradata to Actively Manage the Data Collection Survey Process," Section on Survey Research Methods, Joint Statistical Meetings, Denver. CO.

Nicolaas, Gerry. (2011). "Survey Paradata: A Review," ESRC National Centre for Research, methods review paper. National Centre for Research Methods, Swindon, UK.

West, B., and Sinibaldi, J. (2013). "The Quality of Paradata: A Literature Review," in Kreuter, F. (ed.), Improving Surveys with Paradata: Analytic Uses of Process Information. New York: John Wiley & Sons.

Yan, T., and Olson, K. (2013). "Analyzing Paradata to Investigate Measurement Error," in Kreuter, F. (ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information*. New York: John Wiley & Sons.

Response Quality Indicators

Definition: Statistical process indicators (SPIs) refer to quality measures based on actual survey responses. Though not an exhaustive list, SPIs include item response distributions, item response rates, breakoff rates, plausibility indices, response consistency, edit failure rates (both within the instrument and during data processing), response differentiation, and completeness of open-ended responses (for subsequent coding). These measures can be used to evaluate the quality of collected data and the performance of individual questions.

SPIs can be used with pretests or field tests to identify problematic items prior to the main data collection, as well as during and immediately following the main data collection. The most common approach, especially in field pretests, is to review item response distributions. Is there enough variation in responses? Are scale items highly skewed? Is there a high rate of "other" responses? Is the item producing a high rate of "refusals" and/or "don't knows" (or "blanks" if a self-administered survey)? Any problems of this type would trigger further investigation (Presser et al., 2004). More sophisticated multivariate approaches, such as item response theory (see contribution on Item Response Theory; Reeve, 2011) and latent class modeling (see Miller, this volume; Biemer and Berzofsky, 2011), could also be used.

A common approach for ongoing data collections is to compare current responses to prior year(s) data. Unexpected deviations in SPIs or the survey estimates themselves may suggest a problem with the question(s). A split-ballot field experiment can also be performed to test proposed design changes to counter increasing refusal responses (Pleis, Dahlhamer, and Meyer, 2006). This work also demonstrates the importance of decomposing item nonresponse. Analyses of the relative contribution of different types of nonresponse (e.g., "refusals," "don't knows," "blanks") may also provide insights into possible reduction strategies and their potential impact on measurement error (Dahlhamer et al., 2004).

Other examples of process indicators include:

- Breakoff rates which measure the extent to which a respondent starts the survey but stops providing answers at some point during the survey (Peytchev, 2006). Rates can be measured at the survey (how many respondents broke off from the survey) or at the item level (how many respondents broke off at a particular question). Questions generating a high breakoff rate should be investigated.

- Edit failures rates which are measures of the extent to which survey responses are deemed implausible based on a set of editing rules (Biemer and Lyberg, 2003). High rates of edit failures

(post-collection processing) or instrument error messages (during data collection) may highlight items that are difficult to answer or where a design issue exists that inhibits recording answers in a correct manner. However, edit failures may be difficult to interpret as some edits look for associations among multiple items or with previously reported data, making it difficult to isolate error sources (Tuttle, et al., 2010).

- Response consistency measures are often used when a re-interview is conducted (see section on re-interviews) but response consistency across survey modules can also be explored. Inconsistent responses to questions in different parts of a survey instrument that are intended to measure the same construct may indicate that questions are not working as intended; respondents are not giving their full effort in completing the task, or interviewers are not administrating the items properly

- Response differentiation (or straightlining) is a measure that looks at the number of different scale points used by a respondent over a set of questions. A formalized measure ranging from 0 (no differentiation; i.e., respondent answered the same way for all questions) to 1 (respondent used all possible scale points) was proposed by McCarty and Shrum (2000). While less differentiation may be indicative of error from interviewers, respondents, or mode of collection, poor question design may also be plausible.

More sophisticated comparative approaches such as plausibility indices have been used with business surveys, although the method could be adapted for some household surveys. A plausibility index is defined as "an index of deviation from expected values that are computed from [administrative] data, data from previous years and data from comparable firms" (Giesen and Hak, 2005).

Strengths: SPIs are cost-effective when they can be incorporated into ongoing data collections since they are based on actual survey responses; no additional data collection is necessary. Results are directly applicable since they reflect actual survey behavior under real survey conditions. If SPIs are based on a random sample, they will not suffer from nonrepresentativeness which plagues some field- and laboratory-based evaluations. Statistical inferences would also be supported.

Limitations: Unless SPIs are analyzed as part of testing done prior to fielding of the data collection, problems will not be identified until after production administration has begun. Also, when comparing data over time, the analyst must take care in interpretation as changes in responses/estimates may be an artifact of changing social and/or economic conditions. Hence, when working with SPIs it can be difficult to determine when a problem truly exists. Finally, SPIs may uncover question problems (i.e., demonstrate diagnostic utility), but provide little insight into solutions to those problems (i.e., design utility). A question may suffer from high item nonresponse, and statistical modeling may identify subgroups for which the item is problematic, but the actual revisions needed may not be

readily apparent. It is important that SPIs are paired with question evaluation methods that can identify the source or nature of those problems.

References:

Biemer, P.P., and Berzofsky, M. (2011). "Some Issues in the Application of Latent Class Models for Questionnaire Design, in Madans, J., Miller, K., Maitland, A., and Willis, G. (eds.), *Question Evaluation Methods: Contributing to the Science of Data Quality* (pp. 153–186). Hoboken, NJ: John Wiley & Sons.

Biemer, P.P., and Lyberg, L.E. (2003). *Introduction to Survey Quality*. Hoboken, NJ: John Wiley & Sons.

Couper, M.P., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley.

Dahlhamer, J., Dixon, J., Doyle, P., Eargle, J., and McGovern, P. (2004). "Quality at the Item Level: Decomposing Item and Concept Response Rates," *Proceedings of the European Conference on Quality and Methodology in Official Statistics (Q2004)*, Mainz, Germany, May 24–26.

Giesen, D., and Hak, T. (2005). "Revising the Structural Business Survey: From a Multi-method Evaluation to Design," *Proceedings of the FCSM Research Conference*, Federal Committee on Statistical Methodology, Washington DC.

McCarty, J.A., and Shrum, L.J. (2000). "The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures," *Public Opinion Quarterly, 64*:271–298.

Peytchev, A. (2006). "A Framework for Survey Breakoffs," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Pleis, J.R., Dahlhamer, J.M., and Meyer, P.S. (2006). "Unfolding the Answers? Income Nonresponse and Income Brackets in the National Health Interview Survey," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J.M., and Singer, E. (2004). "Methods for Testing and Evaluating Survey Questions in Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, Hoboken, NJ: Wiley.

Reeve, B.B. (2011). "Applying Item Response Theory for Questionnaire Evaluation," in Madans, J., Miller, K., Maitland, A., and Willis, G. (eds.), *Question Evaluation Methods: Contributing to the Science of Data Quality* (pp. 105–124). Hoboken, NJ: John Wiley & Sons.

Tuttle, A.D., Morrison, R.L., and Willimack, D.K. (2010). "From Start to Pilot: A Multi-method Approach to the Comprehensive Redesign of an Economic Survey Questionnaire," *Journal of Official Statistics*, 26(1): 87–103.

Item Response Theory

Definition: Item response theory (IRT) is a statistical modeling tool that analyzes the relationship and function of individual items within a set of questions measuring a common construct. IRT models are often referred to as *latent trait models*. The term *latent* is used to emphasize that the measures of interest are hypothesized traits, constructs, or attributes which are not directly observable and hence must be inferred from discrete item responses. Therefore, IRT is only appropriate for examination of multiple survey questions that are decidedly understood as measuring a particular phenomenon.

IRT can be used to examine subgroup differences in responding to survey questions and to develop item banks that serve as the foundation for computerized adaptive testing (CAT). More specifically, IRT models an individual's response to a survey question as a function of his or her standing on the construct. Each item is fit with its own IRT model composed of a set of parameters that describe how the item performs for measuring different levels of the measured construct. Parametric models, such as the IRT two-parameter logistic (2PL) model, typically include an item discrimination, or slope, parameter and an item difficulty, severity, or threshold parameter. The item discrimination parameter indicates the magnitude of an item's ability to differentiate among people at different levels along the trait continuum. Larger discrimination parameters indicate stronger relationships between the item and the latent construct, or trait, and contribute more to determining a person's score. When an item has just two responses, e.g. True/False, the threshold parameter is equal to the point on the trait continuum at which a person has a 50 percent probability of responding "true" to the item. For items such as multiple choice questions with a correct answer, a "guessing" parameter is often incorporated into the model. The "guessing" parameter is the probability of a positive response to the item if the person does not know the answer.

Parametric IRT models make three key assumptions: unidimensionality, local independence, and monotonicity. Unidimensionality posits that the set of items measure a single continuous latent trait/construct. Local independence asserts that the only systematic relationship among the items is explained by their conditional relationship with the latent trait/construct. High inter-item correlations or residual correlations in a factor model would suggest a violation of the local independence assumption. Lastly, monotonicity means the probability of endorsing or selecting an item response indicative of higher levels of the trait/construct should increase as the underlying level of the trait/construct increases.

Strengths: An important feature of IRT models is the information function, an index that indicates the range over the trait continuum for which an item is most useful for discriminating among individuals. The item information functions can be used to create shorter surveys by allowing the questionnaire developers to select the most informative group of questions for measuring the underlying trait/construct in the target population. Furthermore, under the assumption of local independence, the item information functions can be summed across multiple items

creating a test information function. The test information function evaluates the tests, or scales, reliability over the trait/construct continuum.

Limitations: One of the largest limitations of IRT is that the actual meaning and composition of the latent construct is a hypothesis, and is not proven during analysis. Additionally, IRT generally requires a survey to have a large sample size and can only be conducted following data collection. Because IRT must be conducted using multiple items that purport to measure the same construct, the questionnaire becomes longer and increases respondent burden.

Reference

Madans, Jennifer, et al. (2011). *Question Evaluation Methods: Contributing to the Science of Data Quality*, vol. 567. New York: John Wiley & Sons.

Latent Class Analysis

Definition: A method for assigning persons to classes (categories) based on some underlying latent construct measured using mutually exclusive categorical items or indicators (Feick, 1989). Latent class analysis (LCA) is a type of structural equation modeling similar to factor analysis that instead focuses specifically on categorical inputs and outcomes (Feick, 1989; Lazarsfeld & Henry, 1968). LCA is typically used to model an unobserved, categorical latent variable with a hypothesized number of latent classes or groups (or levels, if ordered) that explains the relationship between a set of independent observed variables (Feick, 1989). LCA can be used both to predict membership to a given class and/or the probability of a specific response to an item given class membership (Feick, 1989). LCA is directly analogous to factor analysis for categorical inputs and outputs.

While LCA had become increasing popular as a means to evaluate measurement error in survey data, applications for directly evaluating questions are more limited (Kreuter et al., 2008; Biemer, 2004; Biemer and Weisen, 2002). The most direct application of LCA requires multiple measurements of the concept of interest by the same respondent. Few federal surveys incorporate multiple measures of the concept of interest within the same interview, but many employ panel designs that can be exploited using LCA models with some assumptions. LCA models can incorporate classification variables, such as individual demographics or family characteristics.

Where LCA has been applied, the estimates are shown to be successful in determining misclassification or, at least, differences in classification by question type. Because the "true" value of the underlying construct of interest is typically unknown, results from these models have to be accepted with some caution. However, the alternative, where one question or type of response is intrinsically favored (e.g., a larger expenditure report), is generally less robust. In an experiment, Kreuter et al. (2008) showed that estimates from LCA consistently approximated known true values. Even when models are under-identified (where there is too little observed information to estimate the

latent constructs under classical LCA theory [see Goodman, 1974]), and in cases where some assumptions are violated, the LCA estimates are fairly robust. LCA is useful for the evaluation of questions in panel surveys or where multiple measures are available, especially if no gold standard is available and all current measures are thought to measure the concept of interest with some error.

Strengths: LCA is a statistical test that can be employed when a true value or a gold standard of the concept of interest is not available. Unlike factor analysis, LCA is well suited to the analysis of categorical data that is commonly found in federal surveys. Model estimates have shown to be robust even when some assumptions are violated and when the number of indicators are too few to fully identify the model (underidentification). LCA models can not only evaluate the extent of misclassification (or other type of measurement error) associated with survey items, but the characteristics of participants who are especially susceptible to measurement error can be identified.

Limitations: LCA only predicts categorical outcomes that are mutually exclusive and exhaustive. The models require multiple measurements of the concept of interest. LCA models frequently require strong model assumptions, especially if there are too few indicators for the model to be fully identified. Estimation of these models by maximum likelihood can be difficult and requires multiple iterations of the estimation procedure to avoid common pitfalls (e.g., local minima, boundary issues, poor selection of starting values, and data sparseness [Biemer 2011]).

References:

Biemer, P.P. (2011). *Latent Class Analysis of Survey Error.* Hoboken, New Jersey: Wiley.

Biemer, P.P. (2004). "An Analysis of Classification Error for the Revised Current Population Survey Employment Questions," *Survey Methodology* 30(2):127–140.

Biemer, P.P., and Wiesen, C. (2002). "Measurement Error Evaluation of Self-Reported Drug Use: A Latent Class Analysis of the US National Household Survey on Drug Abuse," *Journal of the Royal Statistical Society* 165(1):97–119.

Feick, L.F. (1989). "Latent Class Analysis of Survey Questions That Include Don't Know Responses," *Public Opinion Quarterly* 53(4): 525–547.

Goodman, L.A. (1974). "Exploratory Latent Structure Analysis Using both Identifiable and Unidentifiable Models," *Biometrika* 61:215–231.

Griffin, J. (2013). "On the Use of Latent Variable Models to Detect Differences in the Interpretation of Vague Quantifiers," *Public Opinion Quarterly* 77(1):124–144.

Kreuter, F., Ting, Y., and Tourangeau, R. (2008). "Good Item or Bad: Can Latent Class Analysis Tell? The Utility of Latent Class Analysis for the Evaluation of Survey Questions," *Journal of the Royal Statistical Society: Series A* 171(3):723–738.

Lazarsfeld, P.F., and Henry, N.W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

McCutcheon, A.L. (1985). “A Latent Class Analysis of Tolerance and Nonconformity in the American Public,” *Public Opinion Quarterly* 49(4):474–488.

Attachment A: Multi-Method Examples

Multi-Method Questionnaire Evaluation and Testing for the Census of Agriculture

1. Expert Reviews:

In early 2009 a National Agricultural Statistics Service (NASS) team began initial work reviewing input from a number of groups of experts. This work included recommendations from an external panel review of the census program conducted by the Council on Food, Agricultural and Resource Economics, recommendations from the NASS advisory committee, suggestions from both formal and informal contacts with commodity organizations, data users, subject matter experts, and routinely collected feedback from NASS staff following the previous (2007) Census of Agriculture (COA). Each of these expert groups provided both specific and general comments on potential improvements to the form and the processes used in data collection.

2. Evaluation of Historical Data:

Initial work by the testing team also included a review of the data that had been collected in the 2007 COA. For each item on the questionnaire, the number of times the item had been edited, either by the processing programs or by an analyst, was tabulated, as was the number of times the item had been missing and imputed. The items then were rank ordered by both the number and percent of edits and imputations. This clearly indicated items that were good candidates for evaluation, redesign and improvement. The testing team also reviewed the record of calls that had been made to the toll-free telephone help line during the 2007 COA. Information about what section of the form respondents were asking about was collected, along with a narrative comment was collected. In this way, a ranking of the sections where most help was requested was developed along with information on why the respondent had called.

3. Cognitive Testing:

Once questionnaires had been revised and newly proposed content had been added, the draft questionnaires were tested in cognitive interviews. As is usual in cognitive testing, the samples were small, approximately 40 interviews, with respondents chosen who were thought to be representative of other census respondents. The respondents reported data in the new sections to identify items considered to be problematic. During these interviews, respondents reported the requested data items and then answered follow up questions about their reporting processes and interpretation of terms. Interviews were conducted in multiple states, with multiple kinds of respondents and with specially trained interviewers.

4. Field Testing:

Following the review of 2007 data and initial cognitive testing, initial revisions were made to the form and a small field test was conducted in early 2010. Approximately 5,000 forms were mailed out using procedures similar to normal operational data collection. A larger field test was also conducted in early 2011, with a sample of 30,000

mailed forms after additional revisions. Each field test included several split sample comparisons. These included possible changes to the data collection techniques such as alternative cover letters or alternative reminders. After each round of the field testing, response rates and the reported data were examined. Analysis looked for data obviously reported in error (e.g., inconsistency between items, unreasonable values), rates of missing data and other suspicious data.

5. Follow Up Interviews:

Following each round of the field test, a subset of approximately 200 respondents was re-contacted, asked to review their reported data, and then asked to verify or expand on their reports. The objective of these interviews was to explain questionable data, and verify that reasonable data were accurate. For example, respondents who reported land rented but did not report paying any cash rent were asked if they were renting the land under some other type of rental arrangement. In addition to verifying data, follow-up interview respondents were also asked about their use of the instruction sheet, their reaction to online reporting, their overall experience, etc.

Results from each round of testing were used to inform questionnaire and procedure revisions. Questionnaires were revised several times during this process before they were finalized for the 2012 data collection. Additional reviews of data collected in the 2012 Census of Agriculture will provide a final evaluation of the performance of the questionnaires.

Multi-Method Questionnaire Development and Evaluation for National Postsecondary Student Aid and Beginning Postsecondary Student Longitudinal Studies

Every 4 years, the National Center for Education Statistics (NCES) fields the National Postsecondary Student Aid Study (NPSAS), a nationally representative sample survey focused on how students and their families pay for postsecondary education. Alternating cycles of NPSAS serve as the base year of either the Beginning Postsecondary Student Longitudinal Study (BPS) or the Baccalaureate and Beyond Longitudinal Study (B&B). The former tracks a cohort of first-time students in postsecondary education for a period of six years and focuses on college completion, while the latter follows a cohort of baccalaureate graduates to learn more about the relationship between attainment of a bachelor's degree and students' life experiences up to 10 years after completing college (e.g., labor market outcomes, enrollment in graduate education).

By 2008, senior NCES leadership had grown concerned that one of the longitudinal studies–BPS–was becoming somewhat outdated. Although it continued to provide useful descriptive estimates of persistence and attainment across a wide range of student and institutional characteristics, its capacity to support modeling aimed at understanding more complex relationships among those characteristics and other elements of the student experience was diminishing. It was noted that the study lacked a coherent conceptual framework, resulting in a NPSAS base-year (and BPS follow-up) student interview that was somewhat unfocused, containing legacy items of uncertain utility and missing items hypothesized to be related to student persistence behaviors while in college. To address these concerns, the decision was made to rebuild key components of the base-year BPS interview in NPSAS.

In 2009, NCES commissioned two nationally known scholars of higher education to propose one or more conceptual frameworks for the NPSAS/BPS redesign and, as appropriate, relevant methodological approaches to data collection. The resulting document offered an explanatory model of college persistence and attainment informed by theories of human capital development and behavior economics. The conceptual framework was based on Becker's Human Capital Model, which posits that students weigh the costs and benefits of education in order to decide whether to continue in college, but the explanatory model also borrows from more contemporary theories about the cognitive and behavioral constraints surrounding this decision process. The document also included a series of constructs to be measured and a literature review summarizing relevant prior findings for each. To operationalize the model, NCES undertook the following steps:

1. The face validity of existing BPS items was checked against the new framework, and all items that were clearly unrelated to the study's new focus were removed;
2. Potentially valid items were aligned to the list of new measurement constructs, and priorities for new item development were determined;

3. Using the literature review identified by the external consultants, NCES and data collection contractor staff developed potential new items for each measurement construct that was not previously represented in the instrument or that was represented but could be improved;
4. A series of exploratory focus groups comprised of first-time beginning undergraduates (n = 50) enrolled at all types of institutions were held, seeking to determine the terminology students used to describe phenomena of interest, develop appropriate response sets, gain an initial understanding of how students were interpreting questions and terms, and assess cognitive burden of certain question types and response formats.
5. The results from focus groups were used to further the development and redesign of survey questions formulated to measure the identified new measurement constructs. The resulting items were then tested in cognitive interviews (n = 48).
6. Using feedback from the cognitive labs, proposed items were revised, and framework designers and other federal and nonfederal subject matter experts were chosen for a technical review panel (TRP);
7. The TRP was convened to reach consensus about whether each new item demonstrated construct validity (i.e., that it measured what the conceptual framework demanded) and that, overall, the complete item set demonstrated content validity (i.e., that construct-valid items existed to fully cover all elements in the proposed framework);
8. Using feedback from the TRP, proposed items were further revised and plans made for another round of cognitive interviewing;
9. A series of in-person and over-the-phone cognitive interviews were conducted with first-time beginning undergraduates (n = 25) enrolled at all types of institutions, in which interviewers prompted respondents to "think aloud" as they completed questions, gauged respondents' understanding of key terms, and identified sources of confusions or undue burden;
10. Using feedback from the second round of cognitive labs, proposed items were again revised, a field-test questionnaire prepared, and instrumentation experiments designed to compare visual analog scales to traditional categorical response options and to test different approaches to capturing a new construct identified by the conceptual framework;
11. The field test was administered (n = 3,860) to test the survey and conduct the experiments;
12. At the conclusion of the field test, responses were analyzed to: establish convergent or divergent validity (as appropriate) between items that represented key constructs, identify potentially problematic response sets or items that elicited floor or ceiling effects, identify appropriate instrumentation techniques, and assess respondent burden (i.e., form timings, break-off).
13. In preparation for the development of the full-scale questionnaire, the process repeated itself starting at step (4) above, re-testing items with students in focus groups, sharing analysis results with TRP participants, and conducting final cognitive lab work prior to finalizing the questionnaires and finalizing the questionnaires for the web-based instrument used in the national data collection.

Multi-Method Questionnaire Development and Evaluation of a Sexual Identity Question

In 2013, the National Center for Health Statistics worked to develop and test a sexual identity question for the National Health Interview Survey (NHIS). Both qualitative and quantitative methods were used to determine design problems with existing questions and to develop and test a modified version.

1. Analysis of existing data sets. Data from various surveys were analyzed. A primary concern regarding the existing measurement of sexual identity was that the combined frequency of missing categories (i.e., Other and Don't Know/Refused) was approximately the same as that of the sexual minority categories (Gay, Lesbian, Bisexual). Even more problematic, missing data were not randomly distributed. Those more likely to have missing data were more likely to be women with less education, Hispanics and Spanish speakers.

2. Cognitive interviewing study. Cognitive interviews were conducted to determine the explanation for the missing data. Analysis of cognitive interviews revealed that respondents who do not identify as lesbian, gay, or bisexual, do not necessarily identify as "straight" or "heterosexual," and that some of these respondents were confused by the terminology. For example, it was found that respondents can confuse the words "homosexual" and "heterosexual," believing that "heterosexual" is the equivalent of being gay and that "homosexual" is the equivalent of being straight. Additionally, some cognitive interviewing respondents, not knowing the terminology, surmised that the term "bisexual" means "heterosexual," concluding that "bi" means two: one man and one woman. Spanish interviews indicated that terminology was even more of a problem for Spanish-only speakers.

In designing a new question, the 2006 National Survey of Family Growth (NSFG) version was used as a point of departure because it was regarded as the best performing question to date on a survey. The goals to improve upon the 2006 NSFG version were 1) reduce misclassification of nonminority respondents, 2) reduce rates of "don't know" and "something else," and 3) particularly for those respondents who do fall into "something else," be able to sort nonminority from minority sexual identity cases.

3. Specialized field tests. The revised question was fielded in two separate studies: the first with 500 cases, and the second with 2,000 cases. The analytic plan for the field test was twofold. First, missing data (i.e., cases where the respondent was not classified within a sexual identity category) were analyzed in relation to the key demographic variables. Second, in order to understand the reason and extent to which respondents were not classified, data from the follow-up questions were analyzed. Additionally, a split sample experiment was conducted in the 2000 case field test. Half of the sample received the question as part of an ACASI (Audio Computer Assisted Survey Interview) module while the other half were asked the question in a face-to-face interview.

Analysis of the field test data suggests that the modified sexual identity question is an improved version because the percentage of missing data is decreased, and, importantly, there are no significant relationships between missing

cases and any demographic variable. No differences in terms of prevalence estimates, item non-response or cut-offs were found between the ACASI and face-to-face interview modes.