Methods for Testing and Evaluating Survey Questionnaires

Edited by

STANLEY PRESSER
University of Maryland, College Park, MD

JENNIFER M. ROTHGEB
U.S. Bureau of the Census, Washington, DC

MICK P. COUPER
University of Michigan, Ann Arbor, MI

JUDITH T. LESSLER
Research Triangle Institute, Research Triangle Park, NC

ELIZABETH MARTIN
U.S. Bureau of the Census, Washington, DC

JEAN MARTIN
Office for National Statistics, London, UK

ELEANOR SINGER
University of Michigan, Ann Arbor, MI

A JOHN WILEY & SONS, INC., PUBLICATION

Methods for Testing and Evaluating Survey Questionnaires

WILEY SERIES IN SURVEY METHODOLOGY
Established in Part by WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *Robert M. Groves, Graham Kalton, J. N. K. Rao, Norbert Schwarz, Christopher Skinner*

The ***Wiley Series in Survey Methodology*** covers topics of current research and practical interests in survey methodology and sampling. While the emphasis is on application, theoretical discussion is encouraged when it supports a broader understanding of the subject matter.

The authors are leading academics and researchers in survey methodology and sampling. The readership includes professionals in, and students of, the fields of applied statistics, biostatistics, public policy, and government and corporate enterprises.

BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys
BIEMER and LYBERG · Introduction to Survey Quality
COCHRAN · Sampling Techniques, *Third Edition*
COUPER, BAKER, BETHLEHEM, CLARK, MARTIN, NICHOLLS, and O'REILLY (editors) · Computer Assisted Survey Information Collection
COX, BINDER, CHINNAPPA, CHRISTIANSON, COLLEDGE, and KOTT (editors) · Business Survey Methods
*DEMING · Sample Design in Business Research
DILLMAN · Mail and Internet Surveys: The Tailored Design Method
GROVES and COUPER · Nonresponse in Household Interview Surveys
GROVES · Survey Errors and Survey Costs
GROVES, DILLMAN, ELTINGE, and LITTLE · Survey Nonresponse
GROVES, BIEMER, LYBERG, MASSEY, NICHOLLS, and WAKSBERG · Telephone Survey Methodology
GROVES, FOWLER, COUPER, LEPKOWSKI, SINGER, and TOURANGEAU · Survey Methodology
*HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volume 1: Methods and Applications
*HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volume II: Theory
HARKNESS, VAN DE VIJVER, and MOHLER · Cross-Cultural Survey Methods
HEERINGA and KALTON · Leslie Kish Selected Papers
KISH · Statistical Design for Research
*KISH · Survey Sampling
KORN and GRAUBARD · Analysis of Health Surveys
LESSLER and KALSBEEK · Nonsampling Error in Surveys
LEVY and LEMESHOW · Sampling of Populations: Methods and Applications, *Third Edition*
LYBERG, BIEMER, COLLINS, de LEEUW, DIPPO, SCHWARZ, TREWIN (editors) · Survey Measurement and Process Quality
MAYNARD, HOUTKOOP-STEENSTRA, SCHAEFFER, VAN DER ZOUWEN · Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview
PRESSER, ROTHGEB, COUPER, LESSLER, MARTIN, MARTIN, and SINGER (editors) · Methods for Testing and Evaluating Survey Questionnaires
RAO · Small Area Estimation
SIRKEN, HERRMANN, SCHECHTER, SCHWARZ, TANUR, and TOURANGEAU (editors) · Cognition and Survey Research
VALLIANT, DORFMAN, and ROYALL · Finite Population Sampling and Inference: A Prediction Approach

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Methods for Testing and Evaluating Survey Questionnaires

Edited by

STANLEY PRESSER
University of Maryland, College Park, MD

JENNIFER M. ROTHGEB
U.S. Bureau of the Census, Washington, DC

MICK P. COUPER
University of Michigan, Ann Arbor, MI

JUDITH T. LESSLER
Research Triangle Institute, Research Triangle Park, NC

ELIZABETH MARTIN
U.S. Bureau of the Census, Washington, DC

JEAN MARTIN
Office for National Statistics, London, UK

ELEANOR SINGER
University of Michigan, Ann Arbor, MI

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Methods for testing and evaluating survey questionnaires / Stanley Presser . . . [et al.].
p. cm.—(Wiley series in survey methodology)
Includes bibliographical references and index.
ISBN 0-471-45841-4 (pbk. : alk. paper)
1. Social surveys—Methodology. 2. Questionnaires—Methodology. 3. Social sciences—Research—Methodology. I. Presser, Stanley II. Series.

HM538.M48 2004
300′.72′3—dc22

2003063992

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To the memory of
Charles Cannell and Seymour Sudman,
two pretesting pioneers whose contributions shaped the field
of survey research

Contents

Contributors

Reginald P. Baker, Market Strategies, Inc., Livonia, MI

Paul Beatty, National Center for Health Statistics, Hyattsville, MD

Paul Biemer, Research Triangle Institute, Research Triangle Park, NC

Johnny Blair, Abt Associates, Washington, DC

Natacha Borgers, Utrecht University, The Netherlands

Anna Chan, U.S. Bureau of the Census, Washington, DC

Frederick G. Conrad, University of Michigan, Ann Arbor, MI

Mick P. Couper, University of Michigan, Ann Arbor, MI

Scott Crawford, Market Strategies, Inc., Livonia, MI

Edith de Leeuw, Utrecht University, The Netherlands

Theresa J. DeMaio, U.S. Bureau of the Census, Washington, DC

Wil Dijkstra, Free University, Amsterdam, The Netherlands

Don A. Dillman, Washington State University, Pullman, WA

Pat Doyle, U.S. Bureau of the Census, Washington, DC

Stasja Draisma, Free University, Amsterdam, The Netherlands

Jennifer Dykema, University of Wisconsin, Madison, WI

Barbara Forsyth, Westat, Rockville, MD

Floyd Jackson Fowler, Jr., University of Massachusetts, Boston, MA

Irmtraud Gallhofer, University of Amsterdam, The Netherlands

Julia Klein Griffiths, U.S. Bureau of the Census, Washington, DC

Sue Ellen Hansen, University of Michigan, Ann Arbor, MI

Janet Harkness, Zentrum fur Umfragen Methoden und Analysen, Mannheim, Germany

John P. Hoehn, Michigan State University, East Lansing, MI
Lilli Japec, Statistics Sweden, Stockholm, Sweden
Michael D. Kaplowitz, Michigan State University, East Lansing, MI
Ashley Landreth, U.S. Bureau of the Census, Washington, DC
Judith T. Lessler, Research Triangle Institute, Research Triangle Park, NC
Frank Lupi, Michigan State University, East Lansing, MI
Lars Lyberg, Statistics Sweden, Stockholm, Sweden
Elizabeth Martin, U.S. Bureau of the Census, Washington, DC
Jean Martin, Office for National Statistics, London, United Kingdom
Louise C. Mâsse, National Cancer Institute, Bethesda, MD
Danna L. Moore, Washington State University, Pullman, WA
Jeffrey Moore, U.S. Bureau of the Census, Washington, DC
Joanne Pascale, U.S. Bureau of the Census, Washington, DC
Beth-Ellen Pennell, University of Michigan, Ann Arbor, MI
Stanley Presser, University of Maryland, College Park, MD
Cleo D. Redline, National Science Foundation, Arlington, VA
Bryce B. Reeve, National Cancer Institute, Bethesda, MD
Jennifer M. Rothgeb, U.S. Bureau of the Census, Washington, DC
Willem E. Saris, University of Amsterdam, The Netherlands
Nora Cate Schaeffer, University of Wisconsin, Madison, WI
Alisú Schoua-Glusberg, Research Support Services, Evanston, IL
Eleanor Singer, University of Michigan, Ann Arbor, MI
Johannes H. Smit, Free University, Amsterdam, The Netherlands
Tom W. Smith, National Opinion Research Center, Chicago, IL
Astrid Smits, Statistics Netherlands, Heerlen, The Netherlands
Janice Swinehart, Market Strategies, Inc., Livonia, MI
John Tarnai, Washington State University, Pullman, WA
Roger Tourangeau, University of Michigan, Ann Arbor, MI
Diane K. Willimack, U.S. Bureau of the Census, Washington, DC
William van der Veld, University of Amsterdam, The Netherlands
Johannes van der Zouwen, Free University, Amsterdam, The Netherlands
Patricia Whitridge, Statistics Canada, Ottawa, Ontario, Canada
Gordon B. Willis, National Cancer Institute, Bethesda, MD

Preface

During the past 20 years, methods for testing and evaluating survey questionnaires have changed dramatically. New methods have been developed and are being applied and refined, and old methods have been adapted from other uses. Some of these changes were due to the application of theory and methods from cognitive science and others to an increasing appreciation of the benefits offered by more rigorous testing. Research has begun to evaluate the strengths and weaknesses of the various testing and evaluation methods and to examine the reliability and validity of the methods' results. Although these developments have been the focus of many conference sessions and the subject of several book chapters, until the 2002 International Conference on Questionnaire Development, Evaluation and Testing Methods, and the publication of this monograph, there was no conference or book dedicated exclusively to question testing and evaluation.

Jennifer Rothgeb initially proposed the conference at the spring 1999 Questionnaire Evaluation Standards International Work Group meeting in London. The Work Group members responded enthusiastically and encouraged the submission of a formal proposal to the organizations that had sponsored prior international conferences on survey methodology. One member, Seymour Sudman, provided invaluable help in turning the idea into a reality and agreed to join the organizing committee, on which he served until his death in May 2000. Shortly after the London meeting, Rothgeb enlisted Stanley Presser for the organizing committee, as they flew home from that year's annual meetings of the American Association for Public Opinion Research (and, later, persuaded him to chair the monograph committee). The members of the final organizing committee, chaired by Rothgeb, were Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Stanley Presser, Eleanor Singer, and Gordon B. Willis.

The conference was sponsored by four organizations: the American Statistical Association (Survey Research Methods Section), the American Association for Public Opinion Research, the Council of American Survey Research Organizations, and the International Association of Survey Statisticians. These organizations provided funds to support the development of both the conference and the monograph. Additional financial support was provided by:

Abt Associates
Arbitron Company
Australian Bureau of Statistics
Iowa State University
Mathematica Policy Research
National Opinion Research Center
National Science Foundation
Nielsen Media Research
Office for National Statistics (United Kingdom)
Research Triangle Institute
Schulman, Ronca & Bucuvalas, Inc.
Statistics Sweden
University of Michigan
U.S. Bureau of Justice Statistics
U.S. Bureau of Labor Statistics
U.S. Bureau of the Census
U.S. Bureau of Transportation Statistics
U.S. Energy Information Administration
U.S. National Agricultural Statistics Service
U.S. National Center for Health Statistics
Washington State University
Westat, Inc.

Without the support of these organizations, neither the conference nor the monograph would have been possible.

In 2000, the monograph committee, composed of the editors of this volume, issued a call for abstracts. Fifty-three were received. Authors of 23 of the abstracts were asked to provide detailed chapter outlines that met specified goals. After receiving feedback on the outlines, authors were then asked to submit first drafts. Second drafts, taking into account the editors' comments on the initial drafts, were due shortly before the conference in November 2002. Final revisions were discussed with authors at the conference, and additional editorial work took place after the conference.

A contributed papers subcommittee, chaired by Gordon Willis and including Luigi Fabbris, Eleanor Gerber, Karen Goldenberg, Jaki McCarthy, and Johannes van der Zouwen, issued a call for submissions in 2001. One hundred five were received and 66 chosen. Two of the contributed papers later became monograph chapters.

The International Conference on Questionnaire Development, Evaluation and Testing Methods—dedicated to the memory of Seymour Sudman—was held in Charleston, South Carolina, November 14–17, 2002. There were 338 attendees, with more than one-fifth from outside the United States, representing 23 countries on six continents. The Survey Research Methods Section of the American Statistical Association funded 12 conference fellows from South Africa, Kenya, the Philippines, Slovenia, Italy, and Korea, and a National Science Foundation grant funded 10 conference fellows, most of whom were U.S. graduate students.

Over half of the conference participants attended at least one of the four short courses that were offered: Methods for Questionnaire Appraisal and Expert Review by Barbara Forsyth and Gordon Willis; Cognitive Interviewing by Eleanor Gerber; Question Testing for Establishment Surveys by Kristin Stettler and Fran Featherston; and Behavior Coding: Tool for Questionnaire Evaluation by Nancy Mathiowetz. Norman Bradburn gave the keynote address, "The Future of Questionnaire Research," which was organized around three themes: the importance of exploiting technological advances, the increasing challenges posed by multicultural, multilanguage populations, and the relevance of recent research in sociolinguistics. The main conference program included 32 sessions with 76 papers and 15 poster presentations.

Conference planning and on-site activities were assisted by Linda Minor, of the American Statistical Association (ASA), and Carol McDaniel, Shelley Moody, and Safiya Hamid, of the U.S. Bureau of the Census. Adam Kelley and Pamela Ricks, of the Joint Program in Survey Methodology, developed and maintained the conference Web site, and Robert Groves, Brenda Cox, Daniel Kasprzyk and Lars Lyberg, successive chairs of the Survey Research Methods Section of the ASA, helped to promote the conference. We thank all these people for their support.

The goal of this monograph is a state-of-the-field review of question evaluation and testing methods. The publication marks a waypoint rather than an ending. Although the chapters show great strides have been made in the development of methods for improving survey instruments, much more work needs to be done. Our aim is for the volume to serve both as a record of the many accomplishments in this area, and as a pointer to the many challenges that remain.

We hope the book will be valuable to students training to become the next generation of survey professionals, to survey researchers seeking guidance on current best practices in questionnaire evaluation and testing, and to survey methodologists designing research to advance the field and render the current chapters out of date.

After an overview in Chapter 1 of both the field and of the chapters that follow, the volume is divided into seven parts

I. Cognitive Interviews: Chapters 2 to 5
II. Supplements to Conventional Pretests: Chapters 6 to 8
III. Experiments: Chapters 9 to 11
IV. Statistical Modeling: Chapters 12 to 14
V. Mode of Administration: Chapters 15 to 18
VI. Special Populations: Chapters 19 to 22
VII. Multimethod Applications: Chapters 23 to 25

Each of the coeditors served as a primary editor for several chapters: Rothgeb for 3 to 5; Singer for 6 to 8; Couper for 9, 10, 15, 16, and 18; Lessler for 11, 12, 14, and 17; E. Martin for 2, 13, and 19 to 22; and J. Martin for 23 to 25. In

addition, each coeditor served as a secondary editor for several other chapters. We are grateful to the chapter authors for their patience during the lengthy process of review and revision, and for the diligence with which they pursued the task.

We are also indebted to Rupa Jethwa, of the Joint Program in Survey Methodology (JPSM), for indefatigable assistance in creating a final manuscript from materials provided by dozens of different authors, and to Robin Gentry, also of JPSM, for expert help in checking references and preparing the index.

Finally, for supporting our work during the more than four years it took to produce the conference and book, we thank our employing organizations: the University of Maryland, U.S. Bureau of the Census, University of Michigan, Research Triangle Institute, and U.K. Office for National Statistics.

August 2003

STANLEY PRESSER
JENNIFER M. ROTHGEB
MICK P. COUPER
JUDITH T. LESSLER
ELIZABETH MARTIN
JEAN MARTIN
ELEANOR SINGER

CHAPTER 1

Methods for Testing and Evaluating Survey Questions

Stanley Presser
University of Maryland

Mick P. Couper
University of Michigan

Judith T. Lessler
Research Triangle Institute

Elizabeth Martin
U.S. Bureau of the Census

Jean Martin
Office for National Statistics, United Kingdom

Jennifer M. Rothgeb
U. S. Bureau of the Census

Eleanor Singer
University of Michigan

1.1 INTRODUCTION

An examination of survey pretesting reveals a paradox. On the one hand, pretesting is the only way to evaluate in advance whether a questionnaire causes problems for interviewers or respondents. Consequently, both elementary textbooks

Methods for Testing and Evaluating Survey Questionnaires, Edited by Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer
ISBN 0-471-45841-4

and experienced researchers declare pretesting indispensable. On the other hand, most textbooks offer minimal, if any, guidance about pretesting methods, and published survey reports usually provide no information about whether questionnaires were pretested and, if so, how, and with what results. Moreover, until recently, there was relatively little methodological research on pretesting. Thus, pretesting's universally acknowledged importance has been honored more in the breach than in the practice, and not a great deal is known about many aspects of pretesting, including the extent to which pretests serve their intended purpose and lead to improved questionnaires.

Pretesting dates either to the founding of the modern sample survey in the mid-1930s or to shortly thereafter. The earliest references in scholarly journals are from 1940, by which time pretests apparently were well established. In that year, Katz reported: "The American Institute of Public Opinion [i.e., Gallup] and *Fortune* [i.e., Roper] pretest their questions to avoid phrasings which will be unintelligible to the public and to avoid issues unknown to the man on the street" (1940, p. 279).

Although the absence of documentation means we cannot be certain, our impression is that for much of survey research's history, there has been one conventional form of pretest. Conventional pretesting is essentially a dress rehearsal in which interviewers receive training like that for the main survey and administer a questionnaire as they would during a survey proper. After each interviewer completes a handful of interviews, response distributions (generally univariate, occasionally bivariate or multivariate) may be tallied, and there is a debriefing in which the interviewers relate their experiences with the questionnaire and offer their views about the questionnaire's problems.

Survey researchers have shown remarkable confidence in this approach. According to one leading expert: "It usually takes no more than 12–25 cases to reveal the major difficulties and weaknesses in a pretest questionnaire" (Sheatsley, 1983, p. 226), a judgment similar to that of another prominent methodologist, who maintained that "20–50 cases is usually sufficient to discover the major flaws in a questionnaire" (Sudman, 1983, p. 181).

This faith in conventional pretesting was probably based on the common experience that a small number of conventional interviews often reveals numerous problems, such as questions that contain unwarranted suppositions, awkward wordings, or missing response categories. But there is no scientific evidence justifying the confidence that this type of pretesting identifies the major problems in a questionnaire.

Conventional pretests are based on the assumption that questionnaire problems will be signaled either by the answers that the questions elicit (e.g., don't knows or refusals), which will show up in response tallies, or by some other visible consequence of asking the questions (e.g., hesitation or discomfort in responding), which interviewers can describe during debriefing. However, as Cannell and Kahn (1953, p. 353) noted: "There are no exact tests for these characteristics." They go on to say that "the help of experienced interviewers is most useful at this

point in obtaining subjective evaluations of the questionnaire." Similarly, Moser and Kalton (1971, p. 50) judged that "almost the most useful evidence of all on the adequacy of a questionnaire is the individual fieldworker's [i.e., interviewer's] report on how the interviews went, what difficulties were encountered, what alterations should be made, and so forth." This emphasis on interviewer perceptions is nicely illustrated in Sudman and Bradburn's (1982, p. 49) advice for detecting unexpected word meanings: "A careful pilot test conducted by *sensitive* interviewers is the most direct way of discovering these problem words" (emphasis added).

Yet even if interviewers were trained extensively in recognizing problems with questions (as compared with receiving no special training at all, which is typical), conventional pretesting would still be ill suited to uncovering many questionnaire problems. This is because certain kinds of problems will not be apparent from observing respondent behavior, and the respondents themselves may be unaware of the problems. For instance, respondents can misunderstand a closed question's intent without providing any indication of having done so. And because conventional pretests are almost always "undeclared" to the respondent, as opposed to "participating" (in which respondents are informed of the pretest's purpose; see Converse and Presser, 1986), respondents are usually not asked directly about their interpretations or other problems the questions may cause. As a result, undeclared conventional pretesting seems better designed to identify problems the questionnaire poses for interviewers, who know the purpose of the testing, than for respondents, who do not.

Furthermore, when conventional pretest interviewers do describe respondent problems, there are no rules for assessing their descriptions or for determining which problems that are identified ought to be addressed. Researchers typically rely on intuition and experience in judging the seriousness of problems and deciding how to revise questions that are thought to have flaws.

In recent decades, a growing awareness of conventional pretesting's drawbacks has led to two interrelated changes. First, there has been a subtle shift in the goals of testing, from an exclusive focus on identifying and fixing overt problems experienced by interviewers and respondents to a broader concern for improving data quality so that measurements meet a survey's objectives. Second, new testing methods have been developed or adapted from other uses. These include cognitive interviews (the subject of Part I of this volume), behavior coding, response latency, vignette analysis, and formal respondent debriefings (all of which are treated in Part II), experiments (covered in Part III), and statistical modeling (Part IV). In addition, new modes of administration pose special challenges for pretesting (the focus of Part V), as do surveys of special populations, such as children, establishments, and those requiring questionnaires in more than one language (all of which are dealt with in Part VI). Finally, the development of new pretesting methods raises issues of how they might best be used in combination, as well as whether they in fact lead to improvements in survey measurement (the topics of Part VII).

1.2 COGNITIVE INTERVIEWS

Ordinary interviews focus on producing codable responses to the questions. Cognitive interviews, by contrast, focus on providing a view of the processes elicited by the questions. Concurrent or retrospective *think-alouds* and/or probes are used to produce reports of the thoughts that respondents have either as they answer the survey questions or immediately after. The objective is to reveal the thought processes involved in interpreting a question and arriving at an answer. These thoughts are then analyzed to diagnose problems with the question.

Although he is not commonly associated with cognitive interviewing, William Belson (1981) pioneered a version of this approach. In the mid-1960s, Belson designed "intensive" interviews to explore seven questions that respondents had been asked the preceding day during a regular interview administered by a separate interviewer. Respondents were first reminded of the exact question and the answer they had given to it. The interviewer then inquired: "When you were asked that question yesterday, exactly what did you think the question meant?" After nondirectively probing to clarify what the question meant to the respondent, interviewers asked, "Now tell me exactly how you worked out your answer from that question. Think it out for me just as you did yesterday—only this time say it aloud for me." Then, after nondirectively probing to illuminate how the answer was worked out, interviewers posed scripted probes about various aspects of the question. These probes differed across the seven questions and were devised to test hypotheses about problems particular to each of the questions. Finally, after listening to the focal question once more, respondents were requested to say how they would now answer it. If their answer differed from the one they had given the preceding day, they were asked to explain why. Six interviewers, who received two weeks of training, conducted 265 audiotaped, intensive interviews with a cross-section sample of residents of London, England. Four analysts listened to the tapes and coded the incidence of various problems.

These intensive interviews differed in a critical way from today's cognitive interview, which integrates the original and follow-up interviews in a single administration with one interviewer. Belson assumed that respondents could accurately reconstruct their thoughts from an interview conducted the previous day, which is inconsistent with what we now know about the validity of self-reported cognitive processes (see Chapter 2). However, in many respects, Belson moved considerably beyond earlier work, such as Cantril and Fried (1944), which used just one or two scripted probes to assess respondent interpretations of survey questions. Thus, it is ironic that his approach had little impact on pretesting practices, an outcome possibly due to its being so labor intensive.

The pivotal development leading to a role for cognitive interviews in pretesting did not come until two decades later with the Cognitive Aspects of Survey Methodology (CASM) conference (Jabine et al., 1984). Particularly influential was Loftus's (1984) postconference analysis of how respondents answered survey questions about past events, in which she drew on the think-aloud technique used by Herbert Simon and his colleagues to study problem solving (Ericsson

and Simon, 1980). Subsequently, a grant from Murray Aborn's program at the National Science Foundation to Monroe Sirken supported both research on the technique's utility for understanding responses to survey questions (Lessler et al., 1989) and the creation at the National Center for Health Statistics (NCHS) in 1985 of the first "cognitive laboratory," where the technique could routinely be drawn on to pretest questionnaires (e.g., Royston and Bercini, 1987).

Similar laboratories were soon established by other U.S. statistical agencies and survey organizations.[1] The labs' principal, but not exclusive activity involved cognitive interviewing to pretest questionnaires. Facilitated by special exemptions from Office of Management and Budget survey clearance requirements, pretesting for U.S. government surveys increased dramatically through the 1990s (Martin et al., 1999). At the same time, the labs took tentative steps toward standardizing and codifying their practices in training manuals (e.g., Willis, 1994) or protocols for pretesting (e.g., DeMaio et al., 1993).

Although there is now general agreement about the value of cognitive interviewing, no consensus has emerged about best practices, such as whether (or when) to use think-alouds versus probes, whether to employ concurrent or retrospective reporting, and how to analyze and evaluate results. In part, this is due to the paucity of methodological research examining these issues, but it is also due to lack of attention to the theoretical foundation for applying cognitive interviews to survey pretesting.

In Chapter 2, Gordon Willis addresses this theoretical issue, and in the process contributes to the resolution of key methodological issues. Willis reviews the theoretical underpinnings of Ericsson and Simon's original application of think-aloud interviews to problem-solving tasks and considers the theoretical justifications for applying cognitive interviewing to survey tasks. Ericsson and Simon concluded that verbal reports can be veridical if they involve information a person has available in short-term memory, and the verbalization itself does not fundamentally alter thought processes (e.g., does not involve further explanation). Willis concludes that some survey tasks (for instance, nontrivial forms of information retrieval) may be well suited to elucidation in a think-aloud interview. However, he cautions that the *general* use of verbal report methods to target cognitive processes involved in answering survey questions is difficult to justify, especially for tasks (such as term comprehension) that do not satisfy the conditions for valid verbal reports. He also notes that the social interaction involved in interviewer administered cognitive interviews may violate a key assumption posited by Ericsson and Simon for use of the method.

Willis not only helps us see that cognitive interviews may be better suited for studying certain types of survey tasks than others, but also sheds light on the different ways of conducting the interviews: for instance, using think-alouds versus

[1]Laboratory research to evaluate self-administered questionnaires was already under way at the Census Bureau before the 1980 census (Rothwell, 1983, 1985). Although inspired by marketing research rather than cognitive psychology, this work foreshadowed cognitive interviewing. For example, observers asked respondents to talk aloud as they filled out questionnaires. See also Hunt et al. (1982).

probes. Indeed, with Willis as a guide we can see more clearly that concurrent think-alouds may fail to reveal how respondents interpret (or misinterpret) word meanings, and that targeted verbal probes should be more effective for this purpose. More generally, Willis's emphasis on the theoretical foundation of testing procedures is a much-needed corrective in a field that often slights such concerns.

Chapter 3, by Paul Beatty, bears out Willis's concern about the reactivity of aspects of cognitive interviewing. Beatty describes NCHS cognitive interviews which showed that respondents had considerable difficulty answering a series of health assessment items that had produced no apparent problems in a continuing survey. Many researchers might see this as evidence of the power of cognitive interviews to detect problems that are invisible in surveys. Instead, Beatty investigated whether features of the cognitive interviews might have created the problems, problems that the respondents would not otherwise have had.

Transcriptions from the taped cognitive interviews were analyzed for evidence that respondent difficulty was related to the interviewer's behavior, in particular the types of probes posed. The results generally indicated that respondents who received reorienting probes had little difficulty choosing an answer, whereas those who received elaborating probes had considerable difficulty. During a further round of cognitive interviews in which elaborating probes were restricted to the post-questionnaire debriefing, respondents had minimal difficulty choosing an answer. This is a dramatic finding, although Beatty cautions that it does not mean that the questions were entirely unproblematic, as some respondents expressed reservations about their answers during the debriefing.

Elaborating and reorienting probes accounted for only a small fraction of the interviewers' contribution to these cognitive interviews, and in the second part of his chapter, Beatty examines the distribution of all the interviewers' utterances aside from reading the questions. He distinguishes between cognitive probes (those traditionally associated with cognitive interviews, such as "What were you thinking . . .?" "How did you come up with that . . .?" "What does [term] mean to you?"); confirmatory probes (repeating something the respondent said in a request for confirmation); expansive probes (requests for elaboration, such as "Tell me more about that"); functional remarks (repetition or clarification of the question, which included all reorienting probes); and feedback (e.g., "Thanks; that's what I want to know" or "I know what you mean"). Surprisingly, cognitive probes, the heart of the method, accounted for less than 10% of interviewer utterances. In fact, there were fewer cognitive probes than utterances in any of the other categories.

Taken together, Beatty's findings suggest that cognitive interview results are importantly shaped by the interviewers' contributions, which may not be well focused in ways that support the inquiry. He concludes that cognitive interviews would be improved by training interviewers to recognize distinctions among probes and the situations in which each ought to be employed.

In Chapter 4, Frederick Conrad and Johnny Blair argue that (1) the raw material produced by cognitive interviews consists of verbal reports; (2) the different techniques used to conduct cognitive interviews may affect the quality of these

verbal reports; (3) verbal report quality should be assessed in terms of problem detection and problem repair, as they are the central goals of cognitive interviewing; and (4) the most valuable assessment data come from experiments in which the independent variable varies the interview techniques and the dependent variables are problem detection and repair.

In line with these recommendations, they carried out an experimental comparison of two different cognitive interviewing approaches. One was uncontrolled, using the unstandardized practices of four experienced cognitive interviewers; the other, more controlled, used four less-experienced interviewers, who were trained to probe only when there were explicit indications that the respondent was experiencing a problem. The authors found that the conventional cognitive interviews identified many more problems than the conditional probe interviews.

As with Beatty's study, however, more problems did not mean higher-quality results. Conrad and Blair assessed the reliability of problem identification in two ways: by interrater agreement among a set of trained coders who reviewed transcriptions of the taped interviews, and by agreement between coders and interviewers. Overall, agreement was quite low, consistent with the finding of some other researchers about the reliability of cognitive interview data (Presser and Blair, 1994). But reliability was higher for the conditional probe interviews than for the conventional ones. (This may be due partly to the conditional probe interviewers having received some training in what should be considered "a problem," compared to the conventional interviewers, who were provided no definition of what constituted a "problem.") Furthermore, as expected, conditional interviewers probed much less than conventional interviewers, but more of their probes were in cases associated with the identification of a problem. Thus, Conrad and Blair, like Willis and Beatty, suggest that we rethink what interviewers do in cognitive interviews.

Chapter 5, by Theresa DeMaio and Ashley Landreth, describes an experiment in which three different organizations were commissioned to have two interviewers each conduct five cognitive interviews of the same questionnaire using whatever methods were typical for the organization, and then deliver a report identifying problems in the questionnaire and a revised questionnaire addressing the problems (as well as audiotapes for all the interviews). In addition, expert reviews of the original questionnaire were obtained from three people who were not involved in the cognitive interviews. Finally, another set of cognitive interviews was conducted by a fourth organization to test both the original and three revised questionnaires.

The three organizations reported considerable diversity on many aspects of the interviews, including location (respondent's home versus research lab), interviewer characteristics (field interviewer versus research staff), question strategy (think-aloud versus probes), and data source (review of audiotapes versus interviewer notes and recollections). This heterogeneity is consistent with the findings of Blair and Presser (1993) but is even more striking given the many intervening years in which some uniformity of practice might have emerged. It does,

however, mean that differences in the results of these cognitive interviews across organization cannot be attributed unambiguously to any one factor.

There was variation across the organizations in both the number of questions identified as having problems and the total number of problems identified. Moreover, there was only modest overlap in the particular problems diagnosed (i.e., the organizations tended to report unique problems). Similarly, the cognitive interviews and the expert reviews overlapped much more in identifying which questions had problems than in identifying what the problems were. The organization that identified the fewest problems (both overall and in terms of number of questions) also showed the lowest agreement with the expert panel. This organization was the only one that did not review the audiotapes, and DeMaio and Landreth suggest that relying solely on interviewer notes and memory leads to error.[2] However, the findings from the tests of the revised questionnaires did not identify one organization as consistently better or worse than the others.

All four of these chapters argue that the methods used to conduct cognitive interviews shape the data they produce. This is a fundamental principle of survey methodology, yet it may be easier to ignore in the context of cognitive interviews than in the broader context of survey research. The challenge of improving the quality of verbal reports from cognitive interviews will not be easily met, but it is akin to the challenge of improving data more generally, and these chapters bring us closer to meeting it.

1.3 SUPPLEMENTS TO CONVENTIONAL PRETESTS

Unlike cognitive interviews, which are completely distinct from conventional pretests, other testing methods that have been developed may be implemented as add-ons to conventional pretests (or as additions to a survey proper). These include behavior coding, response latency, formal respondent debriefings, and vignettes.

Behavior coding was developed in the 1960s by Charles Cannell and his colleagues at the University of Michigan Survey Research Center and can be used to evaluate both interviewers and questions. Its early applications were almost entirely focused on interviewers, so it had no immediate impact on pretesting practices. In the late 1970s and early 1980s, a few European researchers adopted behavior coding to study questions, but it was not applied to pretesting in the United States until the late 1980s (Oksenberg et al.'s 1991 article describes it as one of two "new strategies for pretesting questions").

Behavior coding involves monitoring interviews or reviewing taped interviews (or transcripts) for a subset of the interviewer's and respondent's verbal behavior in the question asking and answering interaction. Questions marked by high frequencies of certain behaviors (e.g., the interviewer did not read the question verbatim or the respondent requested clarification) are seen as needing repair.

[2]Bolton and Bronkhorst (1996) describe a computerized approach to evaluating cognitive interview results, which should reduce error even further.

Behavior coding may be extended in various ways. In Chapter 6, Johannes van der Zouwen and Johannes Smit describe an extension that draws on the sequence of interviewer and respondent behaviors, not just the frequency of the individual behaviors. Based on the sequence of a question's behavior codes, an interaction is coded as either paradigmatic (the interviewer read the question correctly, the respondent chose one of the alternatives offered, and the interviewer coded the answer correctly), problematic (the sequence was nonparadigmatic but the problem was solved, e.g., the respondent asked for clarification and then chose one of the alternatives offered), or inadequate (the sequence was nonparadigmatic and the problem was not solved). Questions with a high proportion of nonparadigmatic sequences are identified as needing revision.

Van der Zouwen and Smit analyzed a series of items from a survey of the elderly to illustrate this approach as well as to compare the findings it produced to those from basic behavior coding and from four *ex-ante methods*, that is, methods not entailing data collection: a review by five methodology experts; reviews by the authors guided by two different questionnaire appraisal systems; and the quality predictor developed by Saris and his colleagues (Chapter 14), which we describe in Section 1.5. The two methods based on behavior codes produced very similar results, as did three of the four ex ante methods—but the two sets of methods identified very different problems. As van der Zouwen and Smit observe, the ex-ante methods point out what *could* go wrong with the questionnaire, whereas the behavior codes and sequence analyses reveal what actually *did* go wrong.

Another testing method based on observing behavior involves the measurement of response latency, the time it takes a respondent to answer a question. Since most questions are answered rapidly, latency measurement requires the kind of precision (to fractions of a second) that is almost impossible without computers. Thus, it was not until after the widespread diffusion of computer-assisted survey administration in the 1990s that the measurement of response latency was introduced as a testing tool (Bassili and Scott, 1996).

In Chapter 7, Stasja Draisma and Wil Dijkstra use response latency to evaluate the accuracy of respondents' answers, and therefore, indirectly to evaluate the questions themselves. As they operationalize it, *latency* refers to the delay between the end of an interviewer's reading of a question and the beginning of the respondent's answer. The authors reason that longer delays signal respondent uncertainty, and they test this idea by comparing the latency of accurate and inaccurate answers (with accuracy determined by information from another source). In addition, they compare the performance of response latency to that of several other indicators of uncertainty.

In a multivariate analysis, both longer response latencies and the respondents' expressions of greater uncertainty about their answers were associated with inaccurate responses. Other work (Chapters 8 and 23), which we discuss below, reports no relationship (or even, an inverse relationship) between respondents' confidence or certainty and the accuracy of their answers. Thus, future research needs to develop a more precise specification of the conditions in which different measures of respondent uncertainty are useful in predicting response error.

Despite the fact that the interpretation of response latency is less straightforward than that of other measures of question problems (lengthy times may indicate careful processing, as opposed to difficulty), the method shows sufficient promise to encourage its further use. This is especially so, as the ease of collecting latency information means that it could be routinely included in computer-assisted surveys at very low cost. The resulting collection of data across many different surveys would facilitate improved understanding of the meaning and consequences of response latency and of how it might best be combined with other testing methods, such as behavior coding, to enhance the diagnosis of questionnaire problems.

Chapter 8, by Elizabeth Martin, is about vignettes and respondent debriefing. Unlike behavior coding and response latency, which are "undeclared" testing methods, respondent debriefings are a "participating" method, which informs the respondent about the purpose of the inquiry. Such debriefings have long been recommended as a supplement to conventional pretest interviews (Kornhauser, 1951, p. 430), although they most commonly have been conducted as unstructured inquiries improvised by interviewers. Martin shows how implementing them in a standardized manner can reveal both the meanings of questions and the reactions that respondents have to the questions. In addition, she demonstrates how debriefings can be used to measure the extent to which questions lead to missed or misreported information.

Vignette analysis, the other method Martin discusses, may be incorporated in either undeclared or participating pretests. Vignettes—hypothetical scenarios that respondents evaluate—may be used to (1) explore how people think about concepts, (2) test whether respondents' interpretations of concepts are consistent with those that are intended, (3) analyze the dimensionality of concepts, and (4) diagnose other question wording problems. Martin provides examples of each of these applications and offers evidence of the validity of vignette analysis by drawing on evaluations of questionnaire changes made on the basis of the method.

The three chapters in this part suggest that testing methods differ in the types of problems they are suited to identify, their potential for diagnosing the nature of a problem and thereby for fashioning appropriate revisions, the reliability of their results, and the resources needed to conduct them. It appears, for instance, that formal respondent debriefings and vignette analysis are more apt than behavior coding and response latency to identify certain types of comprehension problems. Yet we do not have good estimates of many of the ways in which the methods differ. The implication is not only that we need research explicitly designed to make such comparisons, but also that multiple testing methods are probably required in many cases to ensure that respondents understand the concepts underlying questions and are able and willing to answer them accurately.

1.4 EXPERIMENTS

Both supplemental methods to conventional pretests and cognitive interviews identify questionnaire problems and lead to revisions designed to address the

problems. To determine whether the revisions are improvements, however, there is no substitute for experimental comparisons of the original and revised items. Such experiments are of two kinds. First, the original and revised items can be compared using the testing method(s) that identified the problem(s). Thus, if cognitive interviews showed that respondents had difficulty with an item, the item and its revision can be tested in another round of cognitive interviews to confirm that the revision shows fewer such problems than the original. The interpretation of results from this type of experiment is usually straightforward, although there is no assurance that observed differences will have any effect on survey estimates.

Second, original and revised items can be tested to examine what, if any, difference they make for a survey's estimates. The interpretation from this kind of experiment is sometimes less straightforward, but such split-sample experiments have a long history in pretesting. Indeed, they were the subject of one of the earliest articles devoted to pretesting (Sletto, 1950), although the experiments that it described dealt with the impact on cooperation to mail surveys of administrative matters such as questionnaire length, nature of the cover letter's appeal, use of follow-up postcards, and questionnaire layout. None of the examples concerned question wording.

In Chapter 9, Floyd Fowler describes three ways to evaluate the results of experiments that compare question wordings: differences in response distributions, validation against a standard, and usability, as measured, for instance, by behavior coding. He provides six case studies that illustrate how cognitive interviews and experiments are complementary. For each, he outlines the problems that the cognitive interviews detected and the nature of the remedy proposed. He then presents a comparison of the original and revised questions from split-sample experiments that were behavior coded. As he argues, this type of experimental evidence is essential in estimating whether different question wordings affect survey results, and if so, by how much.

All of Fowler's examples compare single items that vary in only one way. Experiments can also be employed to test versions of entire questionnaires that vary in multiple, complex ways. This type of experiment is described in Chapter 10, by Jeffrey Moore, Joanne Pascale, Pat Doyle, Anna Chan, and Julia Klein Griffiths with data from SIPP, the Survey of Income and Program Participation, a large U.S. Bureau of the Census survey that has been conducted on a continuing basis for nearly 20 years. The authors revised the SIPP questionnaire to meet three major objectives: minimize response burden and thereby decrease both unit and item nonresponse, reduce seam bias reporting errors, and introduce questions about new topics. Then, to assess the effects of the revisions before switching to the new questionnaire, an experiment was conducted in which respondents were randomly assigned to either the new or old version.

Both item nonresponse and seam bias were lower with the new questionnaire, and with one exception, the overall estimates of income and assets (key measures in the survey) did not differ between versions. On the other hand, unit nonresponse reductions were not obtained (in fact, in initial waves, nonresponse was

higher for the revised version) and the new questionnaire took longer to administer. Moore et al. note that these latter results may have been caused by two complicating features of the experimental design. First, experienced SIPP interviewers were used for both the old and new instruments. The interviewers' greater comfort level with the old questionnaire (some reported being able to "administer it in their sleep") may have contributed to their administering it more quickly than the new questionnaire and persuading more respondents to cooperate with it. Second, the addition of new content to the revised instrument may have more than offset the changes that were introduced to shorten the interview.

In Chapter 11, Roger Tourangeau argues that the practical consideration that leads many experimental designs to compare packages of variables, as in the SIPP case, hampers the science of questionnaire design. Because it experimented with a package of variables, the SIPP research could estimate the overall effect of the redesign, which is vital to the SIPP sponsors, but not estimate the effects of individual changes, which is vital to an understanding of the effects of questionnaire features (and therefore to sponsors of other surveys making design changes). As Tourangeau outlines, relative to designs comparing packages of variables, factorial designs allow inference not only about the effects of particular variables, but about the effects of interactions between variables as well. In addition, he debunks common misunderstandings about factorial designs: for instance, that they must have equal-sized cells and that their statistical power depends on the cell size.

Other issues that Tourangeau considers are complete randomization versus randomized block designs (e.g., should one assign the same interviewers to all the conditions, or different interviewers to different versions of the questionnaire?), conducting experiments in a laboratory setting as opposed to the field, and statistical power, each of which affects importantly the inferences drawn from experiments. Particularly notable is his argument in favor of more laboratory experiments, but his discussion of all these matters will help researchers make more informed choices in designing experiments to test questionnaires.

1.5 STATISTICAL MODELING

Questionnaire design and statistical modeling are usually thought of as being worlds apart. Researchers who specialize in questionnaires tend to have rudimentary statistical understanding, and those who specialize in statistical modeling generally have little appreciation for question wording. This is unfortunate, as the two should work in tandem for survey research to progress. Moreover, the two-worlds problem is not inevitable. In the early days of survey research, Paul Lazarsfeld, Samuel Stouffer, and their colleagues made fundamental contributions to both questionnaire design and statistical analysis. Thus, it is fitting that the first of our three chapters on statistical modeling to evaluate questionnaires draws on a technique, latent class analysis, rooted in Lazarsfeld's work. In Chapter 12, Paul Biemer shows how estimates of the error associated with questions may be made when the questions have been asked of the same respondents two or more times.