PALGRAVE
HANDBOOKS

THE PALGRAVE HANDBOOK OF SURVEY RESEARCH

Edited by
David L. Vannette & Jon A. Krosnick

The Palgrave Handbook of Survey Research

David L. Vannette • Jon A. Krosnick
Editors

The Palgrave Handbook of Survey Research

palgrave
macmillan

Editors

David L. Vannette
Stanford University
Stanford, California, USA

Qualtrics, LLC
Provo, Utah, USA

Jon A. Krosnick
Stanford University
Stanford, California, USA

ISBN 978-3-319-54394-9
ISBN 978-3-319-54395-6 (eBook)
https://doi.org/10.1007/978-3-319-54395-6

Library of Congress Control Number: 2017953116

Cover illustration: Anna Berkut / Alamy Stock Photo

Printed on acid-free paper

This Palgrave Macmillan imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Section 1 Conventional Survey Research

1 Assessing the Accuracy of Survey Research 3
Jon A. Krosnick

2 The Importance of Probability-Based Sampling Methods for Drawing Valid Inferences 7
Gary Langer

3 Sampling for Single and Multi-Mode Surveys Using Address-Based Sampling 13
Colm O'Muircheartaigh

4 Evidence About the Accuracy of Surveys in the Face of Declining Response Rates 19
Scott Keeter

5 Sampling to Minimize Nonresponse Bias 23
J. Michael Brick

6 Cross-National Issues in Response Rates 29
Vasja Vehovar and Koen Beullens

7 Choosing a Mode of Survey Data Collection 43
Roger Tourangeau

8 **Mixed-Mode and Mixed-Device Surveys** 51
Edith Desiree de Leeuw and Vera Toepoel

9 **The Use and Effects of Incentives in Surveys** 63
Eleanor Singer

10 **Methods for Determining Who Lives in a Household** 71
Kathleen Targowski Ashenfelter

11 **Harmonization for Cross-National Comparative Social Survey Research: A Case Study Using the "Private Household" Variable** 79
Jürgen H. P. Hoffmeyer-Zlotnik and Uwe Warner

12 **Answering for Someone Else: Proxy Reports in Survey Research** 87
Curtiss Cobb

13 **Improving Question Design to Maximize Reliability and Validity** 95
Jon A. Krosnick

14 **Cognitive Interviewing in Survey Design: State of the Science and Future Directions** 103
Gordon Willis

15 **Survey Interviewing: Departures from the Script** 109
Nora Cate Schaeffer

16 **How to Improve Coding for Open-Ended Survey Data: Lessons from the ANES** 121
Arthur Lupia

17 **Applying Human Language Technology in Survey Research** 129
Mark Liberman

18 **Maintaining Respondent Trust and Protecting Their Data** 135
Roger Tourangeau

19 **Tackling Panel Attrition** 143
Peter Lynn

20 Respondent Attrition Versus Data Attrition and Their Reduction 155
Randall J. Olsen

21 Best Practices for Creating Survey Weights 159
Matthew DeBell

Section 2 Opportunities to Expand Data Collection

22 New Kinds of Survey Measurements 165
Carli Lessof and Patrick Sturgis

23 The Role of Surveys in the Era of “Big Data” 175
Mario Callegaro and Yongwei Yang

24 Getting the Most Out of Paradata 193
Frauke Kreuter

25 Open Probability-Based Panel Infrastructures 199
Marcel Das, Arie Kapteyn and Michael Bosnjak

26 Collecting Interviewer Observations to Augment Survey Data 211
Brady T. West

27 “Just One More Thing”: Using Leave-Behind Measurement Supplements to Augment Survey Data Collection 217
Michael W. Link

28 Ecological Momentary Assessment in Survey Research 221
Arthur A. Stone

29 Biomarkers in Representative Population Surveys 227
David Weir

30 Measuring Past Events Using Calendar Aids 235
Robert F. Belli

31 Collecting Social Network Data 241
Tobias H. Stark

Section 3 Linking Survey Data with External Sources

32 Methods of Linking Survey Data to Official Records 257
Joseph W. Sakshaug

33 Linking Individual-Level Survey Data to Consumer File Records 263
Josh Pasek

34 Linking Survey Data to Administrative Records in a Comparative Survey Context 267
Annelies G. Blom and Julie Korbmacher

35 Linking Survey Data with the Catalist Commercial Database 275
Robert Blaemire

36 Challenges and Opportunities in Collecting Election Administration Data 289
Michael P. McDonald

37 Challenges with Validating Survey Data 293
Matthew K. Berent

38 The Promise of Collaborative Data Sharing Across Research Sectors 297
Robert M. Groves

Section 4 Improving Research Transparency and Data Dissemination

39 The Importance of Data Curation 303
Steven Ruggles

40 Improving the Usability of Survey Project Websites 309
David L. Vannette

41 The Role of Transparency in Maintaining the Legitimacy and Credibility of Survey Research 315
Arthur Lupia

42 Evidence-Based Survey Operations: Choosing and Mixing Modes 319
Michael Bosnjak

43 Best Practices for Survey Research 331
David L. Vannette

Section 5 Detailed Chapters

44 Reasons for Optimism About Survey Research 347
Jon A. Krosnick

45 Probability Versus Non-Probability Methods 351
Gary Langer

46 Address-Based and List-Based Sampling 363
Colm O'Muircheartaigh

47 The Impact of Survey Non-response on Survey Accuracy 373
Scott Keeter

48 Optimizing Response Rates 383
J. Michael Brick

49 Data Collection Mode 393
Roger Tourangeau

50 Survey Incentives 405
Eleanor Singer

51 Building Household Rosters Sensibly 417
Kathy Ashenfelter

52 Proxy Reporting 427
Curtiss Cobb

53 Questionnaire Design 439
Jon A. Krosnick

54 Cognitive Evaluation of Survey Instruments 457
Gordon Willis

55 Interviewer Deviations from Scripts 465
Nora Cate Schaeffer

56 Coding Open Responses 473
Arthur Lupia

57 What HLT Can Do for You (and Vice Versa) 489
Mark Liberman

58 Confidentiality, Privacy, and Anonymity 501
Roger Tourangeau

59 Panel Attrition 509
Randall Olsen

60 Computation of Survey Weights 519
Matthew DeBell

61 Paradata 529
Frauke Kreuter

62 Interviewer Observations 537
Brady T. West

63 Leave-Behind Measurement Supplements 549
Michael Link

64 Ecological Momentary Assessment and Experience Sampling 563
Arthur Stone

65 Biomarkers 573
David Weir

66 Specialized Tools for Measuring Past Events 587
Robert F. Belli

67 Linking Survey Data to Official Government Records 597
Joseph W. Sakshaug

68 Linking Knowledge Networks Web Panel Data with External Data 607
Josh Pasek

69 History and Promise and Blending Survey Data with Government Records on Turnout 621
Michael P. McDonald

70 Improving Information Quality and Availability Through Interactions Between Government and Academic, and Industry Survey Research Sectors 627
Robert M. Groves

71 Metadata and Preservation 635
Steven Ruggles

72 Usability of Survey Project Websites 645
David L. Vannette

73 Research Transparency and the Credibility of Survey-Based Social Science 655
Arthur Lupia

Index 667

List of Figures

Fig. 6.1 Response rates in relation to nonresponse bias (left) and costs per unit of accuracy (right) 37
Fig. 7.1 The evolution of survey technology 44
Fig. 7.2 Effects of data collection mode on features of the data 47
Fig. 28.1 Comparison of data from DRM and EMA 224
Fig. 29.1 Comparison of self-reported versus measured height by age and gender 228
Fig. 30.1 An example of a CAPI event history calendar 237
Fig. 30.2 A model of data types, data collection modes, and new tools 238
Fig. 31.1 Drag-and-drop question in GENSI. Answers are given by moving name circles into answer boxes 248
Fig. 31.2 Question for network relationships in GENSI. Relationships are indicated by drawing lines between related network contacts 249
Fig. 39.1 A schematic of common data input and output formats 305

List of Tables

Table 6.1 Response rates (%) in the European Social Survey across countries and seven rounds (R1–R7) 31

Overview

For more than thirty years the National Science Foundation has supported data for research on a wide variety of topics by making awards to three major long-term survey efforts, the American National Elections Studies (ANES), the Panel Study of Income Dynamics (PSID), and the General Social Survey (GSS). In February 2012, the Advisory Committee for the Social, Behavioral and Economic Sciences (SBE) was asked to provide advice about future investments in these surveys and others. The Advisory Committee then charged a subcommittee to provide that advice. The Subcommittee on Advancing SBE Survey Research is comprised of Jon Krosnick (Stanford University, chair), Janet Harkness (University of Nebraska, deceased), Kaye Husbands-Fealing (University of Minnesota), Stanley Presser (University of Maryland), and Steven Ruggles (University of Minnesota).

This book provides guidance for researchers, funding agencies, and organizations engaged in survey research as to how best use their resources to support research through survey data collection. Specifically, the book addresses the following questions, as requested:

1. What are the challenges facing survey-based data collection today (e.g., falling participation rates, rising costs, or coverage of frames)?
2. What innovations in survey methodology have taken place or are on the horizon?
3. How should researchers and organizations think about survey data in the context of the explosion of new digital sources of data? Are there opportunities for blending data or mixed source methods that integrate existing

administrative, commercial, or social media data with existing surveys to answer social science questions?

4. Given current challenges faced by survey research as well as the potential opportunities presented by new approaches to survey research, what types of questions will we be able to address with surveys in the future?

The book addresses these four questions—which are about the current and future status of survey research in general (as opposed to uniquely about NSF funded surveys)—by drawing on the results of research content that we commissioned from leading experts.

We assembled a group of leading scholarly experts to generate rich content on topics that fit into four broad areas. First, challenges being faced in conventional survey research were covered across a broad landscape, including key topics such as: probability versus non-probability sampling methods; multi-mode survey techniques; optimizing response rates and how nonresponse affects survey accuracy; use of incentives in survey collection; survey design, visual displays and cognitive evaluation of survey instruments; proxy reporting; interviewing techniques and challenges; confidentiality, respondent attrition and data attrition; and computation of survey weights.

The second category of exploration focuses on opportunities to expand data collection, including: paradata; the use of leave-behind measurement supplements and biomarkers; and specialized tools for measuring past events. Third, several methods of linking survey data with external sources are studied, specifically: improving government, academic and industry data-sharing opportunities; linking survey data to official government records or with the Catalist Commercial Database; linking knowledge networks web panel data with external data; and the use of election administration data with other datasets. Lastly, there is an emphasis on improving research transparency and data dissemination, with a focus on: data curation; evaluating the usability of survey project websites; and the broader topic of the credibility of survey-based social science. Throughout the book we highlight steps that can be taken to enhance the value of survey methodology to a wide range of users, in academia, government, and the private sector.

This book provides several useful outputs, including: (1) insights about how surveys should be done today to maximize data quality (thereby specifying how major infrastructure surveys should be designed and carried out), (2) important challenges facing the methodology, (3) best practices in data dissemination and data collection procedure documentation, (4) approaches that would be most desirable for large-scale infrastructure surveys to implement, and (5) research questions that merit future investigation.

David Vannette is a Ph.D. Candidate in the Department of Communication at Stanford University, and Principal Research Scientist at Qualtrics, LLC., Davis, CA, USA. Jon Krosnick is the Frederic O. Glover Professor in Humanities and Social Sciences at Stanford University, Stanford, CA, USA, and a University Fellow at Resources for the Future. This work was supported by National Science Foundation Award [1256359]. Address correspondence to David L. Vannette or Jon A. Krosnick, Stanford University, McClatchy Hall, Stanford, CA 94305-2050, USA; email: dave.vannette@gmail.com or krosnick@stanford.edu.

Introduction

Survey research is at a crossroads. The need for information to track the public's behaviors, experiences, needs, and preferences has risen dramatically in recent years. Government agencies, businesses, academics, and others make decisions and create policies based on knowledge of populations, and a great deal of such information is collected via surveys. The unemployment rate, the inflation rate, and many other national indicators used widely in America are generated in this way. Thus, the need for high-quality survey research is great and rising.

At the same time, the challenges of conducting high-quality surveys are substantial. The U.S. federal government remains committed to implementing face-to-face interviewing for many of its most important surveys, and in many countries around the world, face-to-face interviewing is the only way to reach a probability sample of the population without incurring substantial noncoverage. Furthermore, research to date suggests that face-to-face interviewing may be the method most likely to generate the highest response rates, the greatest trust and rapport between the researchers/interviewers and the respondents, the most cognitive effort from respondents in generating answers accurately, and the most honestly when providing reports regarding sensitive topics. But face-to-face interviewing is extremely expensive, and the costs of implementing such efforts well have been rising quickly.

In that light, alternative methods of data collection for surveys are appealing. Although telephone interviewing rose in popularity greatly in the 1970s as a more practical alternative to face-to-face interviewing, this method's response rates have been dropping in recent years, and costs have been rising. Remarkably, the accuracy of random-digit dial telephone surveys appears to remain high, but respondents are less likely to be cognitively effortful and honest when being interviewed over the phone than when being interviewed face-to-face.

The rising costs of telephone interviewing set the stage for Internet data collection to become popular. And it has become so. Indeed, billions of dollars are spent annually around the world collecting survey data via the Internet. And some comparison studies have suggested that answering questions via a computer enhances cognitive performance and honesty relative to oral interviewing by telephone. When done with probability samples, Internet surveys seem to be a very promising avenue for effective and efficient data collection.

However, although a number of countries in addition to the United States now have commercial firms or academic institutions collecting survey data from probability samples of the population via the Internet (e.g., the Netherlands, Germany, France, Iceland, Norway, Sweden), this methodology has yet to catch on broadly across the world. Instead, most Internet survey data collection is done from nonprobability samples of people who volunteer to complete surveys for money. Alternative vehicles, such as Amazon's Mechanical Turk, allow for such data collection from individuals who have not signed up to join a survey panel, and Google's survey platform allows for survey data collection from people who surf to a newspaper website and wish to continue reading a news story at no cost to them. Studies in the United States and abroad suggest that such data collection does not yield samples that reflect the population as accurately as do standard probability sampling methods.

But nonprobability sample surveys implemented via the Internet have had tremendous appeal to researchers inside and outside of academia because of their practicality, especially their affordability. Thus, modes of data collection are in flux and in a state of tension. On the one hand, traditional, reliable methods are becoming increasingly costly. And on the other hand, new methods have obvious limitations in terms of their potential to produce generalizable results. At the same time, researchers are increasingly aware of another challenge in survey research: questionnaire design. For nearly a century, survey researchers have, for the most part, designed questionnaires in an intuition-driven, *ad hoc* fashion. As a result, there is tremendous heterogeneity in the design of questions across surveys and even within a single survey. Consider, for example, the ubiquitous rating scale, which has been used in countless surveys. The design of rating scales has no standardization across surveys – scales differ in terms of the number of points offered, the number of points that have verbal labels versus numeric labels versus no labels, the particular labels chosen, and the order in which the points are presented to respondents.

From this heterogeneity, an outside observer might conclude that there is no optimal way to design rating scales or, indeed, to make any other decisions when designing questionnaires. Instead, perhaps all question designs work equally well – as long as respondents can understand a question, they can answer it accurately, one might imagine.

But 70 years of research across the social sciences suggest that this is not true. In fact, hundreds, if not thousands, of studies provide guidance on how to design questions to maximize measurement reliability and validity, how to maximize the uniformity of respondent interpretations of questions, and how to minimize the cognitive demands made of respondents during the process of interpreting questions and answering them. But this information has yet to be disseminated and put into practice consistently across the nation's most important continuing and new surveys. Yet as practitioners' awareness of these best practices grows, so does concern about the value of data collected by questionnaires not conforming to these principles of optimizing measurement accuracy.

Furthermore, as the cost of survey data collection rises, other forms of data are increasingly available in the form of official records that some observers perceive to be potential replacements for survey data. That is, observers ask, "Why ask people whether they were victims of a crime when researchers can consult the electronic records of police departments to assess crime rates?" or "Why ask people how much they paid for milk when researchers can consult scanner data collected and retained by supermarkets?" The answers to these questions are actually quite simple in many cases: as appealing as these uses of official records are, those records are inadequate for many applications where survey data can serve the purpose effectively. For example, many crimes are not reported to the police, and some crimes reported to police officers are not recorded in official records. So efforts to explore the full frequency of crimes require reports from people who experience them. Likewise, although supermarkets track purchases of products and can even link some purchases to the households that made the purchases, many purchases of food items are not made in such settings, and it is not yet possible to link purchasing behavior by a single individual across the full wide array of purchase settings without asking the individual via surveys. Thus, official records do not yet appear to be a viable replacement for all survey data.

Official records do appear to offer potential value in a different way: as a supplement to survey data. Consider, for example, the measurement of voter turnout. Agencies in almost all states in the country make available to researchers official records of who voted in each election, and some states provide a little additional information about the individuals, such as their gender and age. Furthermore, commercial companies offer services whereby they provide additional, in depth information purportedly about each individual on such lists, which can also be gathered from other publicly available records. These sorts of records are thought to be matchable to data collected from survey respondents to enrich understanding of these individuals with what are presumed to be accurate official records about them.

This accuracy hinges on the accuracy of the process by which a survey respondent is matched to official records purportedly about the same individual. This process of matching is being implemented by a number of commercial firms, but these firms consider the matching processes to be proprietary, so scientists cannot full observe the process and assess the accuracy of the results. It is possible that this matching process can be accomplished effectively using highly confidential federal government data obtainable via Census Data Centers because individuals can be matched using their social security numbers. Advances in computer algorithms and computing power make this type of sophisticated and resource-intensive research increasingly achievable. However, very little research has actually examined these opportunities. Thus, the notion of enriching survey data with data from official records is both appealing and increasingly possible.

Another growing challenge in the survey research arena is the maintenance of records documenting how survey data were collected. In recent years, survey professionals have become increasingly sensitive to the importance of documenting absolutely all details of the process by which data collection occurs, to allow researchers to understand the data and differences in results obtained by different data collection methods. This includes show cards displayed to respondents, interviewer training manuals, text of open-ended questions, detailed field reports to permit calculation of response rates using various contemporary methods, and much more information collected by survey researchers and often not retained or disseminated in ways that allow for in-depth, accurate understanding by scholars. Recently, the dissemination of survey data and survey data collection documentation has advanced considerably. But most survey research organizations are not collecting and disseminating information about their surveys optimally. As a result, analysts are handicapped, uninformed about important aspects of the process by which data were generated (and therefore unable to tailor analysis accordingly), and unable to explore important design issues that might impact findings.

Section 1

Conventional Survey Research

1

Assessing the Accuracy of Survey Research

Jon A. Krosnick

Although research on the accuracy of surveys is important, it has not received the attention it deserves. Many articles and books have focused on survey errors resulting from issues relating to coverage, sampling, non-response, and measurement, but very little work has comprehensively evaluated survey accuracy.

Research on survey accuracy may be scarce because it requires having an external measure of the “true” values of a variable in order to be able to judge how well that value is measured by a survey question. For example, in the area of voting behavior, self-reports of turnout are often collected in surveys and compared with the official turnout statistics provided by the Federal Election Commission (FEC) after an election. When these sources yielded different rates, the errors have usually been assumed to be in the self-reports; the FEC numbers are assumed to document the truth.

Studies that have assessed survey accuracy have not yet been integrated into a single comprehensive review. Chang et al. (working paper) conducted such a review, the results of which constitute the first-ever meta-analysis of survey accuracy. The authors identified four principal methods for assessing the accuracy of survey results and collected published studies using each method. These studies assessed accuracy in a wide range of domains,

J.A. Krosnick (✉)
Departments of Communication, Political Science, and Psychology,
Stanford University, Stanford, CA, USA
e-mail: krosnick@stanford.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_1

including behaviors in the arenas of healthcare utilization, crime, voting, media use, and smoking, and measures of respondent characteristics such as demographics, height, and weight.

First, the authors identified 555 studies that matched each respondent's self-report data with objective individual records of the same phenomena, resulting in a dataset of over 520,000 individual matches. This method of verification indicated that for more than 85 percent of the measurements, there was perfect agreement between the survey data and the objective records or measures. Second, the investigators found 399 studies that matched one-time aggregate survey percentages and means with available benchmarks from non-survey data. These studies involved different units of measurement, such as percentages, means in centimeters, kilograms, days, hours, drinks, etc. This assessment method indicated that survey measures matched benchmarks exactly in 8 percent of the instances, 38 percent manifested almost perfect matches (less than one-unit difference), and 73 percent manifested very close matches (less than five-unit difference). Third, the authors found 168 instances in which studies correlated individuals' self-reports in surveys with secondary objective data. The results from this method indicated generally strong associations between the self-reports and the secondary data. Specific results and estimates are shown in the PowerPoint materials. The authors identified six studies that correlated trends over time in self-reports and with trends in objective benchmarks. This approach documented very strong associations between the self-report survey data and trends in the objective benchmarks. Thus, in this meta-analysis, Chang and her colleagues examined over 1000 published comparisons gauging the validated accuracy of survey data, and the vast majority of survey measurements of objective phenomena were found to be extremely accurate.

When differences do occur between survey estimates and objective benchmarks, it is important to consider exactly how these differences may have arisen, rather than immediately discounting the survey data. For example, researchers tend to assume that surveys overestimate voter turnout because of respondent lying. That is, respondents are thought to believe that voting is socially desirable, and so people who didn't vote may claim to have voted in order to look presentable. However, the accumulating literature suggests instead that individual survey reports may be remarkably accurate, and the problem may be that people who participate in elections also over-participate in surveys. If so, the disagreement between aggregate rates of turnout according to surveys vs. government statistics may not be due to inaccurate respondent reporting.

These findings should give survey producers, consumers, and funding agencies considerable optimism about the continued accuracy of surveys as a method of collecting data. The findings also indicate that survey research deserves its role as one of the most used and trusted methods for data collection in the social sciences.

Jon A. Krosnick is the Frederic O. Glover Professor in Humanities and Social Sciences at Stanford University, Stanford, CA, USA and a University Fellow at Resources for the Future. This work was supported by National Science Foundation Award [1256359].

2

The Importance of Probability-Based Sampling Methods for Drawing Valid Inferences

Gary Langer

Before 1936, data on populations generally were collected either via a census of the entire population or "convenience" sampling, such as straw polls. The latter, while quick and inexpensive, lacked a scientific, theoretical basis that would justify generalization to a broader population. Using such methods, the Literary Digest correctly predicted presidential elections from 1916 to 1932 – but the approach collapsed in 1936. The magazine sent postcards to 10 million individuals selected from subscriptions, phone books, and automobile registration records. Through sampling and self-selection bias, the 2.4 million responses disproportionately included Republicans, and the poll predicted an easy win for the losing candidate, Alf Landon.

George Gallup used quota sampling in the same election to draw a miniature of the target population in terms of demographics and partisanship. Using a much smaller sample, Gallup correctly predicted Franklin D. Roosevelt's win. This set the stage for systematic sampling methods to become standard in polling and survey research. (See, e.g., Gallup and Rae 1940.)

But quota sampling turned out not to be a panacea. The approach suffered a mortal blow in the 1948 presidential election, when Gallup and others erroneously predicted victory for Thomas Dewey over Harry Truman. While a variety of factors was responsible, close study clarified the shortcomings of quota sampling. Replicating the U.S. population in terms of

G. Langer (✉)
Langer Research Associates, New York, USA
e-mail: glanger@langerresearch.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_2

cross-tabulations by ethnicity, race, education, age, region, and income, using standard categories, would require 9,600 cells, indicating a need for enormous sample sizes. Further, "The microcosm idea will rarely work in a complicated social problem because we always have additional variables that may have important consequences for the outcome" (Gilbert et al. 1977). And bias can be introduced through interviewers' purposive selection of respondents within each quota group.

After spirited debate, survey researchers coalesced around probability sampling as a scientifically rigorous method for efficiently and cost-effectively drawing a representative sample of the population. In this technique, each individual has a known and ideally non-zero probability of selection, placing the method on firmly within the theoretical framework of inferential statistics. As put by the sampling statistician Leslie Kish, "(1) Its measurability leads to objective statistical inference, in contrast to the subjective inference from judgment sampling, and (2) Like any scientific method, it permits cumulative improvement through the separation and objective appraisal of its sources of errors" (Kish 1965).

In modern times, high-quality surveys continue to rely on probability sampling. But new non-probability methods have come forward, offering data collection via social media postings and most prominently though opt-in online samples. These often are accompanied by ill-disclosed sampling, data collection, and weighting techniques, yet also with routine claims that they produce highly accurate data. Such claims need close scrutiny, on theoretical and empirical bases alike.

Opt-in surveys typically are conducted among individuals who sign up to click through questionnaires on the Internet in exchange for points redeemable for cash or gifts. Opportunities for falsification are rife, as is the risk of a cottage industry of professional survey respondents. One study (Fulgoni 2006) found that among the 10 largest opt-in survey panels, 10 percent of panelists produced 81 percent of survey responses, and 1 percent of panelists accounted for 24 percent of responses.

An example of further challenges in opt-in online surveys is their common and generally undisclosed use of routers to maximize efficiency of administration, albeit at the cost of coverage. As an illustration, participants may be asked if they are smokers; if so, are routed to a smoking survey. If not smokers, they may be asked next if they chew gum. If yes, they are routed to a gum-chewers survey. If not, they may next be asked if they use spearmint toothpaste, and so on. Unbeknownst to sponsors of the toothpaste study, smokers and gum chewers are systematically excluded from their sample.

The approach, then, raises many questions. Who joins these poll-taking clubs, what are their characteristics, and what do we know about the reliability and validity of their responses? Are respondent identities verified? Are responses validated? What sorts of quality control measures are put in place? What survey weights are applied, how were they obtained, and what is their effect? What claims are made about the quality of these data, and how are these claims justified?

Purveyors of opt-in online sampling often point to the declining response rates and increasing costs of probability-based telephone surveys, topics that are addressed later in this book. But these arguments are hardly a constructive defense of alternative methodologies, nor do they recognize the wealth of research identifying response rates as a poor indicator of data quality. Rather than pointing toward potential deficiencies in existing methods, it seems incumbent on the proponents of alternative non-probability methods to construct a reasoned defense of the approach, including a theoretical basis for its validity.

Empirical research consistently has found validity in scientific probabilistic sampling methods. Results for non-probability opt-in panels have been more concerning. An extensive review of existing literature, the AAPOR Report on Online Panels, published by the American Association for Public Opinion Research in 2010, recommended that "researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values." This report concluded, "There currently is no generally accepted theoretical basis from which to claim that survey results using samples from nonprobability online panels are projectable to the general population. Thus, claims of 'representativeness' should be avoided when using these sample sources" (Baker et al. 2010). (Subsequent to this presentation, an AAPOR report on non-probability sampling, in 2013, again noted the absence of a theoretical framework that would support statistical inference.)

In a large empirical study in 2011, Yeager and his colleagues compared seven opt-in online sample surveys with two probability sample surveys, finding that the probability surveys were "consistently highly accurate" while the opt-in samples were "always less accurate … and less consistent in their level of accuracy" (Yeager et al. 2011). The authors also found little empirical support for the claim that some non-probability panels are consistently more accurate others. They reported that weighting did not always improve accuracy of these panels, and they found no indication that higher completion rates produce greater accuracy. A report on data produced for a study by the Advertising Research Foundation found similar problems, as did an independent analysis of 45 individual data quality studies (Baker 2009; Callegaro et al. 2012). These confirm the fundamental issue: the absence of theory that would predict accurate, reliable results from non-probability samples.

Even if they can't be used to generalize about a broader population, it has been suggested that non-probability approaches are sufficient for evaluating associations among variables and for tracking trends over time. However, an empirical study on propensity to complete the U.S. Census, comparing otherwise identical probability-based and non-probability surveys, indicated otherwise. It found "systematic and often sizable differences between probability sample telephone data and non-probability Internet data in terms of demographic representativeness of the samples, the proportion of respondents reporting various opinions and behaviors, the predictors of intent to complete the Census form and actual completion of the form, changes over time in responses, and relations between variables" (Pasek and Krosnick 2010). More study is warranted, but the picture to date is bleak.

Another recent trend is to evaluate information made publicly available on social networks such as Facebook and Twitter. The appeal of these datasets is their size and scope. Data can be collected on a minute-by-minute basis in vast quantities on nearly any topic imaginable. While these forms of data may hold great potential for social scientists, they also present unique challenges. For example, it may be assumed that a Twitter or Facebook post represents one individual expressing his or her actual opinion on something once. In fact some users may post multiple times, using a single account or multiple accounts. Postings may not reflect the self-initiated expression of actual attitudes, but rather may be part of orchestrated campaigns. Accounts may be created by interest groups, corporations, or paid public relations agents. Posts may be produced by automated computer programs known as "bots." Fake accounts can be purchased in bulk. All of these forms of information exist within the same datasets.

Regardless of their source, selecting relevant postings and extracting meaning from them are further challenges. Many postings include slang, irony, sarcasm, abbreviations, acronyms and emoticons, or lack identifiable context. Tests of automated coding systems indicate highly inconsistent results. And again we face the lack of theoretical justification to make inferences about a broader population.

What does the future hold for non-probability samples? Can they be "fixed"? Some researchers suggest the use of Bayesian adjustment, or a return to sample matching. While further research is welcome, what has been lacking to date is the required transparency that must underlie any such evaluation. Non-probability methods should be held to the same analytical standards and evaluated on the same basis as probability samples with regard to claims of accuracy, validity, and reliability. Full disclosure of methods and data quality metrics is crucially important. And the Holy Grail remains the development of an online sampling frame with known probabilities of selection, bringing the enterprise into harmony with sampling theory.

Probability sampling requires ongoing evaluation as well. Some organizations implement poor-quality sampling designs and suboptimal execution and analysis. Coverage is an ongoing concern, and the potential impact of declining response rates needs continuing examination. So does work on probability-based alternatives to traditional telephone methods, such as address-based sampling, mixed-mode designs, and others that may be developed in the future.

Areas for future research:

- Expanded empirical research into the validity and reliability of non-probability survey data
- Efforts to develop a theoretical framework under which such samples may support inference
- Improved assessment and methods of analysis of social media data
- Continued examination of probability-based methods
- Development and implementation of transparency standards

References and Further Reading

Baker, R. (2009). Finally, the Real Issue? Retrieved 2009, from http://regbaker.typepad.com/regs_blog/2009/07/finally-the-real-issue.html

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., et al. (2010). Research Synthesis: AAPOR Report On Online Panels. *Public Opinion Quarterly*, *74*(4), 711–781. http://doi.org/10.1093/poq/nfq048

Callegaro, M., Villar, A., Krosnick, J. A., & Yeager, D. S. (2012). A Systematic Review of Studies Investigating the Quality of Data Obtained with Online Panels. Presented at the Annual Meeting of the American Association for Public Opinion Research, Orlando, FL.

Gallup, G. H., & Rae, S. F. (1940). The Pulse of Democracy: The Public Opinion Poll and How it Works. New York: Simon and Schuster.

Gilbert, J. P., Light, L. R., & Mosteller, F. (1977). Assessing Social Innovations: An Empirical Base for Policy. In W. B. Fairley (Ed.), *Statistics and Public Policy*. Reading, MA: Addison-Wesley Pub Co.

Kish, L. (1965). Survey Sampling. New York: John Wiley & Sons, Inc.

Pasek, J., & Krosnick, J. A. (2010). Measuring intent to participate and participation in the 2010 Census and their correlates and trends: Comparisons of RDD telephone and nonprobability sample. *Survey Methodology*.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly*, *75*(4), 709–747. http://doi.org/10.1093/poq/nfr020

Gary Langer is a survey research practitioner. He is president of Langer Research Associates and former long-time director of polling at ABC Network News. Langer is a member of the Board of Directors of the Roper Center for Public Opinion Research, a trustee of the National Council of Public Polls, and former president of the New York Chapter of the American Association for Public Opinion Research. His work has been recognized with 2 News Emmy awards, 10 Emmy nominations, and AAPOR's Policy Impact Award. Langer has written and lectured widely on the measurement and meaning of public opinion.

3

Sampling for Single and Multi-Mode Surveys Using Address-Based Sampling

Colm O'Muircheartaigh

A combination of factors in recent years has provided an opportunity to transform the application of survey sampling methodology in the USA; the potential is greatest in the case of multi-mode surveys. Survey sampling at its most basic level involves identifying and selecting potential sample members from a population using a sampling frame; the sampling frame comprises the set of materials – lists, maps, etc. – that best covers the population that we wish to describe or analyze – the population of inference.

The perfect sampling frame should include each element in the population once, and only once, and include only eligible population elements. A sampling frame can suffer from two primary types of deficiencies: overcoverage or undercoverage. Undercoverage has traditionally been the greater problem; here the sampling frame fails to include eligible population elements. Consider the case of a survey being conducted by telephone. A frame of listed landline telephone numbers would fail to include unlisted landline numbers and cellphone numbers, and thus exclude those whose telephone access was from an unlisted number or by cellphone-only. Overcoverage occurs when the frame includes elements that appear to be members of the population, but turn out not to be. If we chose instead to include all possible 10-digit numbers in the frame (or even all seven-digit numbers within existing area codes), the great majority of the numbers in the frame would

C. O'Muircheartaigh (✉)
University of Chicago, Chicago, USA
e-mail: caomuirc@uchicago.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_3

have no working telephone associated with them. These redundant numbers would constitute frame overcoverage.

Consider the general problem of frame construction for a survey of the household-residing U.S. population. We want to develop a set of materials that will identify enable us first to select a sample of residents (or households), and then recruit them to participate in a survey. Every element of the household-residing population can be associated with the location (and address) of their household. If we could construct a frame containing all these addresses, we would have the foundation of a very robust sampling frame. Recent developments have made this a real (though still imperfect) prospect.

Nearly all households in the USA have a mailing address that is used by the United States Postal Service (USPS) to deliver mail. This full set of addresses (technically a very close facsimile) can be obtained in the form of the USPS Computerized Delivery Sequence File (DSF or CDSF). The DSF has been designed as an organizational tool for the USPS and enables them to effectively route the delivery of mail at every subdivision of the organization down to the individual mail carriers. The DSF is continuously updated by individual mail carriers via “edit books” that allow them to keep their routes updated with any changes such as new or deleted addresses. As it is in the interest of the mail carrier to have an accurate list of addresses, we have considerable confidence in the general quality of the carrier route data; methodological research suggests that this belief is well founded (Brick et al., 2011; English et al., 2011; Iannacchione, 2011; Link et al., 2008; O’Muircheartaigh et al., 2003). The entire DSF is updated on a monthly basis and can be obtained through third-party vendors who have licensing arrangements with the USPS.

The DSF has nearly 100 percent coverage of households in the USA. The DSF does contain addresses that are not easily identified or located; some examples are Post Office boxes, rural route boxes, drop points, and vacant dwellings; though the incidence of such cases is decreasing, they still present non-trivial challenges.

As a potential sampling frame, the DSF has excellent coverage properties and is greatly superior to the alternatives. The DSF can be used as a frame for many modes given the proper conditions. There are four modes of survey data collection; two involve self-completion (mail and web) and two are interviewer-administered (telephone and face-to-face). For mail it is extremely efficient and effective as the frame is designed for the mail system. For face-to-face (in-person) surveys the DSF has displaced the traditional method of field listing of housing units except in specail subpopulations [e.g. rural addresses without street designation and urban dwellings with multiple informal partitions]. For telephone sampling it is effective only when the address sampled can be readily matched to a telephone number for that sample unit (a sample unit may be an organization, household, or individual). As the basis for a web sample, the DSF frame has limited current value as it is considerably more challenging to match e-mail addresses to mailing addresses; however, in the future new databases may enable address-based lists to form the basis of a frame for web addresses (Messer and Dillman, 2011).

The general name given to sample selection based on the DSF is *address-based sampling* (ABS). Given the versatility of ABS for different modes, it is a particularly powerful approach when multi-mode methods are appropriate. Multi-mode surveys are increasingly common and important in survey research due to the coverage and nonresponse issues with telephone surveys and the escalating costs of face-to-face interviewing. The renaissance in mail surveys has also been a factor. Multi-mode approaches are becoming more common in an attempt to overcome, or bypass, the problems associated with single-mode methods.

In order to be able to conduct a multi-mode survey, the sample of respondents needs to be drawn from identical or at least comparable frames and the sample design needs to permit the transfer of cases across modes. The system must also have the capacity to track cases that are transferred. These requirements make the DSF a promising frame for multi-mode studies, particularly when telephone numbers can be matched to the DSF. Furthermore, advances in technology have made compatibility of case management systems across modes more feasible and these systems are now more capable of handling the complex decision rules and branching involved. Multi-mode approaches are not without their own important drawbacks, however, including an increase in the complexity of data collection and analyses and susceptibility to mode effects.

While the DSF has presented a great opportunity to expand the use of ABS over the past 10 years, there are a number of limitations that need to be addressed by future research. Over- and undercoverage are the main limitations. With regard to overcoverage, identifying units that no longer exist, discarding businesses that have been misclassified as residential, and correcting misgeocoded addresses are problems that continue to present challenges. Undercoverage is a more critical problem, especially when listings need to be matched to the frame for another mode – telephone number for telephone interviewing or geolocation for in-person interviewing. Drop points, rural vacancies, new construction, simplified addresses that lack the standard formatting needed for sampling, and incomplete address fields such as ZIP codes, all present significant problems for surveys targeting specific subpopulations.

Not all problems affect the DSF uniformly across the frame. An important example is the issue of "drop points" – these are addresses where the mail carrier drops off the mail but the distribution to the individual residences is carried out by a third party. The residences all have the same address on the frame, though the frame will usually note how many residences are contained at the address. Drop points are most commonly multi-unit residences such as apartment complexes or small multi-apartment buildings. Drop points represent only about 2 percent of the total addresses on the DSF frame, but in the cities where these drop points tend to be clustered, they present a significant problem. In Chicago approximately 15 percent of addresses are at drop points and in New York City the percentage is close to 20 percent. As the DSF does not specify names associated with an address, it is currently impossible to conduct mail surveys at drop points, and as the mail identifier is only at the drop point level it becomes almost impossible to match any other data to individual households residing at that drop point – making the matching of telephone numbers to the particular household impossible also, unless supplementary data are available. For face-to-face surveys, the problem is remediable but only by training and empowering the interviewer to list (and sample from) the residences at the drop point.

Future research should focus on augmenting the DSF in ways that may be useful to survey researchers. For example, vendors such as InfoUSA and Experian routinely add demographic and other information at the address level – this can include household composition, race/ethnicity, and income, along with market-derived data such as car ownership, magazine subscriptions, and other purchasing behavior. With the caveat that such supplementary information may be subject to serious measurement errors, any information that can be obtained outside the survey could enrich both the design (through stratification for instance) and the analysis (by providing ancillary variables for control or explanation). Identifying novel approaches to augmenting the DSF with external databases promises to be an extremely useful and fruitful area of future research.

Research on reconciling frames in order to make them more compatible will allow ABS to become even more useful as the basis for multi-mode approaches. Linking frames will enable better coverage and make sampling easier, making the entire survey process more flexible and versatile. It is also important to develop hierarchical and relational data structures within which it will be easier to switch modes and survey instruments dynamically during data collection, even within units, without invalidating the analyses. Linking frames with sophisticated database management approaches will enable rapid responses to changes when a survey is in the field. Building in capacity to use an adaptive approach to partially completed cases or requests within cases to change modes could help boost response rates and reduce expense. In the multi-mode context, where ABS is most useful, it is important to continue work on questionnaire design and the development of comparable stimuli for different modes; this will include work on mode effects of question form and question order sequences, and their implications for the reliability and validity of data collected across modes.

Areas for future research:

- Developing robust multi-mode question forms and question wording
- Augmenting the DSF with ancillary data
- Addressing the problem of drop points on the sampling frame
- Identifying improved methods for reconciling and linking disparate sampling frames

References and Further Reading

O'Muircheartaigh, C., Eckman, S., & Weiss, C. (2003) Traditional and Enhanced Field Listing for Probability Sampling. 2002 Proceedings of the Section on Survey Research Methods of the American Statistical Association.

English, N., O'Muircheartaigh, C., Dekker, K., & Fiorio, L. (2011). Qualities of Coverage: Who is Included or Excluded by Definitions of Frame Composition. *2010 Proceedings of the American Statistical Association, Section on Survey Research Methods [CD ROM]*. Alexandria, VA: American Statistical Association.

Brick, J. M., Williams, D., & Montaquila, J. M. (2011). Address-Based Sampling for Subpopulation Surveys. *Public Opinion Quarterly*, *75*(3), 409–428. http://doi.org/10.1093/poq/nfr023

Iannacchione, V. G. (2011). The Changing Role of Address-Based Sampling in Survey Research. *Public Opinion Quarterly*, *75*(3), 556–575. http://doi.org/10.1093/poq/nfr017

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys. *Public Opinion Quarterly*, *72*(1), 6–27. http://doi.org/10.1093/poq/nfn003

Messer, B. L., & Dillman, D. A. (2011). Surveying the General Public over the Internet Using Address-Based Sampling and Mail Contact Procedures. *Public Opinion Quarterly*, *75*(3), 429–457. http://doi.org/10.1093/poq/nfr021

Colm A. O'Muircheartaigh is professor at the University of Chicago in the Harris School of Public Policy and the College; he served as dean of the Harris School from 2009 to 2014. His research encompasses survey sample design, measurement errors in surveys, cognitive aspects of question wording, and latent variable models for nonresponse. He is a senior fellow in the National Opinion Research Center (NORC), where he is responsible for the development of methodological innovations in sample design.

O'Muircheartaigh is co-principal investigator on NSF's Center for Advancing Research and Communication in Science, Technology, Engineering, and Mathematics (ARC-STEM) and on the National Institute on Aging's National Social Life Health and Aging Project (NSHAP). He is a member of the Committee on National Statistics of the National Academies (CNSTAT) and of the Federal Economic Statistics Advisory Committee (FESAC), and serves on the board of Chapin Hall Center for Children.

4

Evidence About the Accuracy of Surveys in the Face of Declining Response Rates

Scott Keeter

Over the past 25 years, there has been a consistent and significant decline in response rates to surveys (Kohut et al. 2012). This can be tracked most reliably in long-term surveys that have decades of year-over-year data on the response rates that they have achieved. For example, the National Household Education Survey has gone from a response rate of 60 percent down to nearly 30 percent in the 11-year period from 1996 to 2007. In the 1970s and 1980s conventional wisdom suggested that the quality of inference from survey data would decline rapidly once response rates dropped below 50 percent. Yet today response rates above 60 percent are the exception rather than the rule, even for the most important national surveys such as the "Big 3" funded by the National Science Foundation (NSF): the American National Election Studies, General Social Survey, and Panel Study of Income Dynamics.

Fortunately, survey results have maintained a remarkable level of reliability despite declining response rates. The consistent high level of performance by high-quality surveys in the face of this challenge is a testament to the robustness of the survey research paradigm. This robustness has done little to quell the perennial fears associated with declining response rates in the survey research community. The great fear is that at some unknown point response rates will degrade to the level where no amount post hoc adjustment will reduce nonresponse bias enough to use the data for population inference.

S. Keeter (✉)
Pew Research Center, Washington, DC, USA
e-mail: skeeter@pewresearch.org

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_4

However, a very compelling meta-analysis by Groves and Peytcheva (2008) demonstrated that at any given level of nonresponse the level of bias varies considerably, meaning that nonresponse itself does not reliably predict nonresponse bias. Experimental research has supported this notion by demonstrating that higher response rates achieved by increasing expense and effort in reducing nonresponse via refusal conversion did not provide different estimates than the same surveys conducted at lower effort/cost/response rates (Keeter et al. 2000; Keeter et al. 2006; Kohut et al. 2012).

While this is good news for the validity of survey data in the face of declining response rates, it does raise another problem: Within any given survey at any level of nonresponse, there can be significant variability in the amount of nonresponse bias for individual measures (Groves et al. 2006). This presents a serious concern because it means that researchers have to figure out what kinds of measures have the greatest likelihood of being biased by nonresponse. This is further complicated by the fact that the nonresponse bias can be caused by a number of different factors including survey design features and characteristics of respondents, both of which may interact to create another layer of causal complexity to disentangle. This makes predicting nonresponse bias incredibly difficult and indeed there is no comprehensive theory of survey response that can generate reliable predictions about when nonresponse bias will occur (Peytchev 2013). Because nonresponse bias is so unpredictable it has also been very difficult to generate remedies or a set of best practices aimed at preventing it from occurring.

There are a few areas for future research in need of attention from researchers and funding agencies that will help develop a better understanding of the impact that declining response rates have on the accuracy of survey data. First, survey researchers should take a concerted look at smarter ways to invest in reducing nonresponse bias. Rather than focusing on propping up unsustainable and largely irrelevant nominal response rates, we should be asking what promising areas of survey design or opportunities for data integration and matching across databases might provide a better return on investment? Second, more basic research is needed into the correlates of nonresponse bias both in terms of survey design and respondent characteristics. One promising but underutilized design in this regard is seeding samples with households that have known characteristics and then observing their response propensities to provide more precise estimates of nonresponse bias. Finally, future research should examine the promise of developing novel weighting schemes based on different known characteristics of respondents and nonrespondents. For example, volunteering behavior has been associated

with response propensity and could be used as an important variable when creating post-stratification weights. If more variables like this can be identified then nonresponse bias in a greater variety of outcome variables can be estimated more precisely and corrected for more adequately.
Areas for future research:

- Identifying targeted approaches to combating nonresponse during survey administration (e.g., adaptive design) that may yield better insurance against nonresponse bias than simply applying comparable effort toward all nonresponding cases (Wagner 2012).
- Identifying novel weighting approaches based on known nondemographic characteristics of respondents and nonrespondents

References and Further Reading

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, *72*(2), 167–189. http://doi.org/10.2307/25167621?ref=search-gateway:e66a229203abccca25b1fead2553bc60

Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in Producing Nonresponse Bias. *Public Opinion Quarterly*, *70*(5), 720–736. http://doi.org/10.2307/4124223?ref=search-gateway:1b71a1db2d93b4619290899f7ac87955

Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006). Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey. *Public Opinion Quarterly*, *70*(5), 759–779. http://doi.org/10.1093/poq/nfl035

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. *Public Opinion Quarterly*, *64*(2), 125–148. http://doi.org/10.1086/317759

Kohut, A., Keeter, S., Doherty, C., & Dimock, M. (2012). Assessing the representativeness of public opinion surveys.

Peytchev, A. (2013). Consequences of Survey Nonresponse. *The Annals of the American Academy of Political and Social Science*, *645*(1), 88–111. http://doi.org/10.1177/0002716212461748

Wagner, J. (2012). A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly*, *76*(3), 555–575. http://doi.org/10.1093/poq/nfs032

Scott Keeter is Senior Survey Advisor for the Pew Research Center in Washington, DC. He is a past president of the American Association for Public Opinion Research (AAPOR) and the recipient of the AAPOR Award for Exceptionally Distinguished Achievement. Since 1980, he has been an election night analyst of exit polls for NBC News. Keeter's published work includes books and articles on public opinion, political participation and civic engagement, religion and politics, American elections, and survey methodology. A native of North Carolina, he attended Davidson College as an undergraduate and received a PhD in political science from the University of North Carolina at Chapel Hill. He has taught at George Mason University, Rutgers University and Virginia Commonwealth University.

5

Sampling to Minimize Nonresponse Bias

J. Michael Brick

As discussed in the Abstract, response rates are a major concern for all researchers conducting surveys. Considerable work has been done to seek ways to stem the tide of declining response rates and also to examine the impact of response rates on data quality. The federal government often requests that funded projects obtain "optimal response rates" in Requests for Proposals, but the very notion of an optimal response rate is ill defined. There are two broad conceptualizations that we might infer from this wording: (1) simply maximize the response rate or (2) achieve the response rate that minimizes nonresponse bias.

In the first conceptualization, researchers are interested in identifying survey design features that will maximize overall response rates within a fixed data collection cost. Key factors in this regard are the content and salience, the sponsor of the survey, and the mode in which the survey is conducted. The content and sponsor of the survey may be relevant to some respondents, influencing the salience of the survey and their willingness to participate. Mode is often one of the most important decisions that influences both cost and response rates. However, these factors are usually challenging to alter, salience is often idiosyncratic to particular respondents, and it is cost that determines the mode more often than the desired response rate.

J.M. Brick (✉)
Westat, 1600 Research Blvd, Rockville, Maryland 20850, USA
e-mail: mikebrick@westat.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_5

Beyond these largely fixed factors, there are some design features that may aid researchers in maximizing response rates. Refusal conversion is particularly important regardless of survey mode; refusal conversion refers to the process of attempting to convert sampled individuals who have previously refused to participate in the survey to become respondents. The number of contacts made to a sampled household is another key to increasing the response rate. Staff training, whether interviewers or other survey staff, is also important. Interviewer training has been demonstrated to increase response rates. For mail and web surveys appropriate design of the questionnaires and other ancillary survey materials such as letters and invitations can aid in maximizing response rates. The relative importance of each of the factors outlined before may vary depending on the type of survey and the population being sampled. For example, households may differ on many of the design features and dimensions that influence response rates when compared with organizational surveys or surveys of other specialized populations such as teachers or doctors.

For cross-sectional household surveys, there are a few best practices for achieving the biggest increase in response per dollar spent. First, a token monetary incentive can substantially increase response rates. The demands of the survey and burden on the respondent should be considered when determining the amount of the incentive; for example, a survey collecting medical specimens such as blood or saliva should provide larger incentives, as should surveys that require a significant amount of the respondent's time to complete.

The design of the survey materials is often overlooked but can have a significant impact on survey response rates. Poorly designed surveys may increase breakoffs, as may surveys that require respondents to perform a very uninteresting task at the outset such as an extensive household roster. Other best practices include developing plans for training interviewers, the number and protocol for contact attempts, and refusal conversion. A protocol that includes multiple contact attempts is critical to obtaining higher response rates. Advance letters informing the respondent of their impending invitation to participate in the survey have also been demonstrated to improve response rates.

To maximize overall response rates, it is important to realize that choices of these influencing design factors are interrelated. The total survey design approach of Don Dillman and the leverage-salience theory of survey participation suggest that the factors may influence different respondents in varied ways. Thus, the point is not to define a set of rigid rules defining the survey recruitment process, but to recognize that many factors are at work and different combinations may be necessary to maximize overall response rates.

If the goal is to maximize overall response rates, then a likely strategy involves "cherry picking" respondents, that is, targeting respondents with the highest propensity to respond. While the survey organization may not explicitly have "cherry picking" as an objective, when the main message to interviewers is to increase response rates the way the interviewers may respond is to choose the easiest households to reach this goal. This strategy results in obtaining more responses from people in the following demographic groups: older, female, homeowners, English speakers, and upper-middle income. These demographics are typically associated with the highest willingness to participate for the lowest amount of effort. To increase response rates it is possible interviewers may target more of these people. However, this sort of approach is unlikely to reduce nonresponse bias, which should be the primary concern when considering response rates.

Often survey clients, including government organizations, will request a particular response rate without acknowledging that higher response rates may actually increase nonresponse bias. This type of fixed response rate requirement places survey providers in the position of recruiting more of the same types of people that typically respond to survey. The survey provider must meet the requirement, even if this is likely to exacerbate nonresponse bias.

A more scientifically valid approach to response rates is to focus on achieving the response rates that minimize nonresponse bias. This is a more difficult construct and harder to incorporate in a fair way into the survey requirements. However, there is considerable evidence demonstrating that response rates do not reliably predict nonresponse bias. Thus, instead of simply trying to maximize response rates or specifying a particular arbitrary and high response rate, survey clients should shift their focus to minimizing nonresponse bias. This is not an easy task and it is much harder to evaluate.

Early research on nonresponse bias modeled it as a deterministic function of the differences between respondent and nonrespondent characteristics and the nonresponse rate. However, this model depends on having measured nonrespondent characteristics, which may be expensive, impractical, and is not commonly possible. A more modern best practice involves modeling nonresponse bias as a stochastic process based on the association between response propensity and the characteristic being estimated. In this model, nonresponse bias cannot be viewed as simply a function of nonresponse rates. It is a much more nuanced and complex problem involving (1) response propensities of units, (2) the type of estimates being derived, (3) the specific estimator, and (4) the auxiliary data used as covariates to model the nonresponse bias.

Future research is needed to understand the reasons that nonresponse bias occurs and what the causal mechanisms are. There may be particular indicators of nonresponse bias that researchers should regularly measure and document. These could be features of the survey such as sponsorship, content, mode, questionnaire design, etc., or they could be characteristics of respondents, such as altruism.

Because surveys produce many types of estimates and bias is a function of the particular statistic, future research to improve methods for optimizing nonresponse bias is not a simple task. A large body of research has looked at the impact of declining response rates on nonresponse error and described many features of nonresponse bias. One area that deserves more attention is to predict nonresponse bias quantitatively – specifically when bias will occur and its magnitude. Existing theory provides very little insight into the underlying measurable causes of nonresponse bias, and empirical research has been largely unable to produce high levels of bias even under conditions where it is expected. One possible explanation is that the unpredictability is related to the dependencies associated with design features affecting response. Thus, future research should aim to develop a more comprehensive theory of nonresponse bias that generates testable hypotheses and reliably predicts the conditions under which nonresponse bias will be observed in the real-world context.

Concrete steps for future research should include comparative analysis of respondents and nonrespondents, with a focus on features that can be experimentally manipulated. Nonresponse bias is likely to continue to be a significant and intractable problem until researchers are able to reliably produce it in an experimental context. Additionally, theories of survey response should be tested in practice, producing differential response rates that are in line with theory are a necessary place to start. It would be helpful if such research begins with low-cost factors such as the design of survey materials. On the analysis side, more research is needed that evaluates potential links between statistical adjustments that are performed post hoc and data collection procedures. It would be extremely beneficial to know if steps can be taken in the data collection process that reduces the need for statistical adjustments later on.

Progress toward understanding and reducing nonresponse bias is likely to remain limited without the development of a broad program of research aimed specifically at this area. Individual studies that are not linked by a central program of research are not likely to be sufficient. National Science

Foundation should consider funding a cohesive set of projects or a larger program of research aimed at understanding nonresponse in a more comprehensive and systematic manner.
Areas for future research:

- Developing a comprehensive program of research aimed at understanding and measuring nonresponse bias
- Comparative analysis of respondents and nonrespondents focused on features that can be experimentally manipulated
- Procedures to test nonresponse bias theories in practice:
  - Programs that aim to produce differential response rates for domains as predicted
  - Programs that aim to change domain response rates based on a sequence of actions
  - Focus on low-cost factors such as material design
- Evaluations linking statistical adjustments and data collection procedures

Further Reading

Brick, J. M., Dipko, S., Presser, S., Tucker, C., & Yuan, Y. (2006). Nonresponse Bias in a Dual Frame Sample of Cell and Landline Numbers. *Public Opinion Quarterly*, *70*(5), 780–793. http://doi.org/10.2307/4124226?ref=search-gateway:1b71a1db2d93b4619290899f7ac87955

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, Phone, Mail, and Mixed-Mode Surveys. Hoboken, N.J.:John Wiley & Sons.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, *70*(5), 646–675. http://doi.org/10.2307/4124220?ref=search-gateway:b0e25f3209c1797c08f7d6e562ea4fc2

Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in Producing Nonresponse Bias. *Public Opinion Quarterly*, *70*(5), 720–736. http://doi.org/10.2307/4124223?ref=search-gateway:1b71a1db2d93b4619290899f7ac87955

Groves, R. M., Presser, S., & Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. *Public Opinion Quarterly*, *68*(1), 2–31.

Groves, R. M., Singer, E., & Corning, A. D. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly, 64*(3), 299–308.

Olson, K. (2006). Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly*, *70*(5), 737–758. http://doi.org/10.1093/poq/nfl038

Peytcheva, E., & Groves, R. M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. *Journal of Official Statistics*, *25*(2), 193–201.

Stinchcombe, A. L., Jones, C., & Sheatsley, P. (1981). Nonresponse Bias for Attitude Questions. *Public Opinion Quarterly*, *45*(3), 359–375. http://doi.org/10.2307/2748611?ref=search-gateway:1b71a1db2d93b4619290899f7ac87955

Dr. J. Michael Brick is Vice President at Westat where he is Co-Director of the Survey Methods Unit and Associate Director of the Statistical Staff. He has more than 40 years of experience in survey research, with special expertise in sample design and estimation for large surveys and in the investigation of nonsampling errors in surveys. Dr. Brick has published in numerous journals and is a Fellow of the American Statistical Association, and an elected member of the International Statistical Institute.

6

Cross-National Issues in Response Rates

Vasja Vehovar and Koen Beullens

Introduction

It is commonly agreed that response rates express the ratio between the responding and eligible units in a survey. The American Association for Public Opinion Research (AAPOR) Standard Definitions (2016) set the standards here, including formulas for response rate calculations. In this chapter, we overview contemporary issues concerning response rates. We first illustrate the existing level of response rates in general population surveys. Next, we overview complications with response rate calculations that are emerging due to the expanded use of web surveys and other technological changes. Finally, we address the context of survey costs, which we believe is one of the biggest response rate research challenges.

V. Vehovar (✉)
University of Ljubljana, Ljubljana, Slovenia
e-mail: Vasja.Vehovar@fdv.uni-lj.si

K. Beullens
KU Leuven, Leuven, Belgium
e-mail: koen.beullens@kuleuven.be

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_6

Response Rates in the European Social Survey

We illustrate the response rate levels and trends from a European perspective, which, however, is very typical of all developed countries. For brevity, we further narrow the focus to general population surveys and the academic context. The European Social Survey (ESS) is perhaps the best case for this purpose given its standardized face-to-face survey mode and that it requires at least four contact attempts. In addition, ambitious methodological targets have been set: non-contacts no greater than 3 percent and response rates no less than 70 percent of the sampled persons. Nonresponse issues are studied very closely, including the use of contact forms (e.g. interviewer observations of neighborhood conditions). Response rates are also comprehensively documented[1] and extensively researched (e.g. Stoop et al. 2010; Billiet and Vehovar 2009).

Response rates (AAPOR standard RR1) for all countries and for all rounds, from Round 1 in 2002 (R1) to Round 7 in 2014 (R7), can be seen in Table 6.1. We can observe that in some countries the response rates are steadily declining (Germany, Norway, Slovenia, Sweden), but sometimes the opposite trend can also appear (France, Spain, Switzerland), which is mainly due to increased efforts in these countries. When interpreting these figures, we need to be aware that the change (e.g. decline) in respondents' willingness to cooperate can hardly be separated from the year-to-year variations in the level of the survey efforts and fieldwork procedures, such as a change in the survey agency, respondent incentives, interviewer rewards, advanced letters, refusal conversion, etc. However, the overall impression is that response rates converge to between 50 percent and 60 percent. Alternatively, we could say that the countries make an effort to assure response rates of between 50 percent and 60 percent. Of course, such a claim might also suggest that further efforts – which could bring the response rates up closer to 70 percent or higher – are not reasonable since the gains in data quality would be too small given the increase in costs. We address this very intriguing issue in final sections, where we discuss the costs.

Response Rate Variations Across Survey Types

Let us illustrate the trends and variations in response rates across survey types. We restrict ourselves here only to probability surveys of the general population (we talk about nonprobability surveys in the next section). We

[1] http://www.europeansocialsurvey.org/

Table 6.1 Response rates (%) in the European Social Survey across countries and seven rounds (R1–R7)

Country	R1 (2002)	R2 (2004)	R3 (2006)	R4 (2008)	R5 (2010)	R6 (2012)	R7 (2014)
Albania	–	–	–	–	–	79	–
Austria	60	62	64	62	–	–	52
Belgium	59	62	62	59	54	59	57
Bulgaria	–	–	66	76	82	75	–
Croatia	–	–	–	64	55	–	–
Cyprus	–	–	67	82	71	77	–
Czech	54	71	–	71	71	69	68
Denmark	68	65	51	54	55	49	52
Estonia	–	79	65	63	56	68	60
Finland	73	71	64	68	60	67	63
France	43	44	47	50	48	53	52
Germany	56	53	55	48	32	34	31
Greece	80	79	–	74	69	–	–
Hungary	70	70	66	62	61	65	53
Iceland	–	51	–	–	–	55	–
Ireland	64	63	57	53	65	68	61
Israel	71	–	–	85	73	78	74
Italy	44	61	–	–	–	37	–
Kosovo	–	–	–	–	–	67	–
Latvia	–	–	71	68	–	–	–
Lithuania	–	–	–	52	45	77	69
Luxembourg	44	52	–	–	–	–	–
Netherlands	68	64	60	50	60	56	59
Norway	65	66	66	60	58	55	54
Poland	73	74	70	71	70	75	66

(*continued*)

Table 6.1 (continued)

Country	R1 (2002)	R2 (2004)	R3 (2006)	R4 (2008)	R5 (2010)	R6 (2012)	R7 (2014)
Portugal	69	71	73	76	67	77	43
Romania	–	–	72	68	–	–	–
Russia	–	–	70	68	67	67	–
Slovakia	–	63	73	73	75	74	–
Slovenia	72	70	65	59	65	58	52
Spain	53	56	66	67	69	71	68
Sweden	69	66	67	63	51	53	51
Switzerland	33	47	52	50	54	52	53
Turkey	–	54	–	67	–	–	–
Ukraine	–	67	66	62	64	59	–
UK	56	51	55	56	56	53	44

Note: The dash sign “-” denotes countries which were not included in the corresponding round further narrow the illustration to the case of Slovenia, which is a typical European Union country with respect to various socio-economic indicators, including survey participation. The directions of the response rate variations presented next are thus very likely to also be found in other countries. Unless stated otherwise, the estimates are for 2015.

Face-to-face surveys of a general population are best illustrated by the Slovenian general social survey (called SJM), one of the longest-running academic surveys in the world. It started in 1968 with a response rate close to 100 percent, before dropping to 92 percent (1980) and 86 percent (1992) (Štebe 1995). Further declines have more recently led to response rates similar to those for the ESS (Table 6.1), from 72 percent (2002) to 53 percent (2014). The same trend can also be observed with the OECD survey Programme of the International Assessment of Adult Competencies[2] where the response rate dropped from 70 percent in 1998 to 62 percent in 2014. In official statistics,[3] the response rates are slightly higher: 69 percent for European Union Statistics on Income and Living Conditions and 68 percent for the Labour Force Survey, while Household Budget Surveys have a response rate of 56 percent for the general part and 49 percent for the diary part. On the other side, the most elaborated commercial surveys with contact strategies similar to the ESS (e.g. National Readership Survey) struggle to achieve 30 percent response rates (Slavec and Vehovar 2011).

Let us also provide some expert estimates of response rates for other survey modes in Slovenia:

- The *telephone surveys* in official statistics (e.g. consumer attitude surveys, tourism travels of domestic populations, household energy consumption) typically obtain a response rate of 40–50 percent. In academic surveys, the response rates are around 30–40 percent, while for commercial ones they are around 10 percent. However, due to public telephone directories' low coverage of just 50 percent of the target population, the overall "reach" is only less than half of that, that is, below 25 percent.
- The *mail surveys* (without incentives) from government statistical offices can obtain response rates of up to about 50 percent, in academic surveys they range from 20 percent to 40 percent, while commercial surveys are typically in single digits.

[2] http://www.oecd.org/site/piaac/

[3] Figures related to official statistical surveys are taken from the website of the Statistical Office of Republic of Slovenia, www.stat.si.

- The *web surveys with mail invitations* in official statistics have response rates of up to 25 percent. (In 2015 the Internet household penetration rate in Slovenia was 78 percent.[4]) Use of mixed-mode data collection (a web questionnaire followed by a mail questionnaire) can move this above 30 percent and it has been demonstrated that prepaid incentives (€5) can even boost it above 70 percent (Berzelak et al. 2015).

Response Rates and Technological Evolution

The changes brought by technology, predominantly reflected in the expansion of web surveys, are strongly altering the response rate landscape. On one side, they introduce various problems for response rate calculations:

- With *standard web surveys* (Callegaro et al. 2015) definition problems appear due to the changing nature of breakoffs, as well as with usable, unusable, complete, and partial units. For example, AAPOR (2016, p. 15) still relies on a classification and terminology based on the face-to-face situation in which a break-off has usually meant an unusable interview with the result that completeness statuses are still classified as complete, partial, or breakoff interviews. However, this does not reflect the specific understanding of breakoffs in web surveys, where breakoffs can still have a complete or partial status (e.g. when a respondent leaves the survey one question before the last one). Similarly, web surveys can hardly be called interviews because typically no interviewer is involved.
- In *online probability panels* serious complications appear (DiSogra and Callegaro 2016) when calculating the response rate due to the increasing complexity of various stages of recruitment (AAPOR 2016, p. 23).
- Surveys are also conducted with dedicated *mobile survey apps* where respondents answer the questionnaire in offline mode. This introduces new issues for response rate calculations. For example, a unit may have problems installing the app or a unit might provide all answers, but they

[4] Source Eurostat: http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tin00134&plugin=1. Available May 10, 2016.

were not successfully transmitted, so such cases should be separated from non-cooperation and implicit refusal. Currently, it is unclear exactly how to do this.

- Various *new survey devices* have emerged, from tablets and smart phones to gaming consoles, and unobtrusive wearable devices. With the "Internet of things," almost any device linked to the Internet can also serve as a survey device. However, for any given device, the technical specifics will potentially require adaptations in response rate calculations because of the unique situations that can arise there, similar to those discussed earlier for mobile apps.
- Technology is also accelerating the introduction of *mixed-mode* surveys, particularly various combinations of inexpensive web surveys and traditional modes (face-to-face, telephone, mail). However, with mixed modes the response rates are becoming complicated to calculate, complex to decompose across the modes, and difficult to compare. They are also becoming less and less informative, as we demonstrate in the next section on survey costs.

On the other side, the changes brought by technology are also closely linked to the prevalent use of nonprobability samples where units are included in a survey with unknown (or zero) probabilities (Vehovar et al. 2016). As a consequence, the initial purpose and meaning of response rates is lost. Namely, without knowing the probabilities of inclusion, the measures developed in probability sampling – including response rates and confidence intervals – cannot and should not be calculated because these calculations rely on the corresponding probabilities.

Despite this fact, many practitioners routinely misapply such calculations also in the nonprobability context. However, these response rates, even if calculated, do not represent the response behavior of the target population, but only measure the recruitment efficacy in a specific subset of the target population. Consequently, for some types of nonprobability sampling (e.g. nonprobability online panels, river sampling) the AAPOR Standard Definitions (2016) recommend labeling the ratios between the number of respondents and the number of invited units as "participation rates" and not as response rates. For some other sample types (e.g. quota and convenience samples), we neither count nor care about the nonresponding units so response rate calculations are not possible at all. The new dominance of nonprobability sampling – which is increasingly considered even in official statistics (Cooper and Greenaway 2015; Rendtel and Amarov 2015) – is thus radically changing the traditional meaning of the term "response rate" We should add that recent AAPOR (2015) code also extends the acceptance of nonprobability samples by (conditionally) allowing the calculation of confidence intervals.

Response Rates Within the Context of Survey Costs

The emerging technological changes and nonprobability samples are not only revolutionizing the survey format and the role of response rate calculations, but they highlight the importance of costs, which have been a somewhat neglected research topic in survey methodology. More specifically, it is very much true that response rates have been extensively discussed in general textbooks, dedicated monographs, papers, book chapters and workshops, such as the Household Survey Nonresponse workshop (1990–2015).[5] However, the overall impression is that this research predominantly focuses on isolated aspects related to trends, levels, mediating factors, prevention measures, and post-survey adjustments. On the other side, response rates are rarely observed in relation to nonresponse bias (which is defined as the difference between the true value and expected value of the variable in a sample survey). Namely, when respondents differ from nonrespondents (e.g. respondents may have higher incomes than nonrespondents), this can result in incorrect estimates (e.g. the reported mean income is too high) which produces nonresponse bias. To observe this relationship, we need to run experiments or simulations within single surveys, a research approach we encounter surprisingly rarely (e.g. Fuchs et al. 2013). Meta-analysis of different surveys (e.g. Groves and Peytcheva 2008) cannot help much here because the uncontrolled variables and selection bias create spurious effects (i.e. ecological fallacies) known in aggregated analyses. For example, surveys with variables, which are vulnerable to nonresponse bias, already make additional efforts to achieve high response rates so as to prevent nonresponse bias, which then contributes to the (false) impression that there is no relationship between response rates and nonresponse bias. To study this relationship, the effect of the increased efforts needs to be observed in a single survey where we can control all the other survey characteristics. For this purpose, we provide the following illustration.

Three hypothetical but plausible situations are illustrated for a single variable, where – all else equal – we observe the effects of increased efforts

[5] http://www.nonresponse.org/

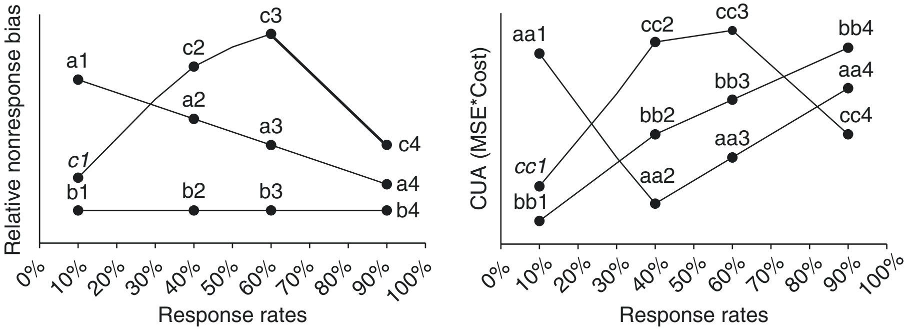


Fig. 6.1 Response rates in relation to nonresponse bias (left) and costs per unit of accuracy (right)

to obtain high response rates (e.g. more contacts, higher incentives, refusal conversions) on the nonresponse bias. We can assume that initial response rates of 10 percent would be obtained, say, with a single contact attempt, 40 percent with 3 contacts, 60 percent with 5, and 90 percent with 10 contacts. In Fig. 6.1 (left panel), we present three potential patterns showing the relationship between response rates and nonresponse bias:

- *Line a* shows the most commonly expected situation where increased efforts actually improve the bias; the corresponding points a1 (response rate 10 percent), a2 (40 percent), a3 (60 percent), and a4 (90 percent) show a linearly declining bias.
- However, it is often true, particularly for variables in marketing research, that response rates have little effect on the bias. Fuchs et al. (2013) also demonstrated for many variables in the ESS that increased response rates (due to more contact attempts) do not change the estimates, so little relation was found with the bias. *Line b* shows this very clearly because efforts to move response rates from 10 percent → 40 percent → 60 percent → 90 percent have no effect on the bias (b1 = b2 = b3 = b4).
- Intriguingly, situations also exist where increased response rate efforts attract even more of the unrepresentative population (Vehovar et al. 2010), so the initial increase in the response rate is counterproductive. We have this situation with *line c* where the rise in the response rate from 10 percent to 40 percent and then to 60 percent further increases the bias (c1→c2→c3). It is only when efforts push the response rate beyond 60 percent (c3) and toward 90 percent (c4) that the bias starts to decrease.

These hypothetical examples reveal that a high response rate is not always desirable. Still, this is often an implicit assumption, although there is no empirical evidence showing this is generally true. Instead, as we demonstrated before, the relationship can be very complex. The situation further changes when we observe the entire context of the total survey error, which is often reduced to accuracy (Vehovar et al. 2010). The latter is measured with the inverse of the mean squared error (MSE), which typically integrates bias and sampling variance (MSE = Bias^2 + Var).

Greater differences appear when we integrate not only nonresponse bias and accuracy, but also the survey costs, for example, Vehovar et al. (2010), Vannieuwenhuyze (2014), Tourangeau et al. (2016), and Roberts et al. (2014). This also puts the problem in a real setting the practitioner faces: which survey design and nonresponse strategy provides the "best buy," that is, the best information (highest data quality) for my costs (efforts)? For this purpose, we observe the costs per unit of accuracy (CUA), which can be calculated as the product of total survey costs and MSE. Based on the CUA, Fig. 6.1 (right panel) further expands the three examples:

- With *line a*, where an increasing response rate reduces the bias, the decreased bias (a1→a3) initially (10 percent→40 percent) outweighs the increased costs. As a consequence, the CUA first decrease (aa1→aa2), too. However, later (at 60 percent and 90 percent) the costs of achieving higher response rates outweigh the gains (a2→a4) of the reduced bias. The best buy (i.e. the lowest CUA) thus remains at 40 percent (aa2).
- With *line b*, where increased response rates are of no help in reducing the bias, this same effect also manifests in a steady increase of the corresponding CUA (bb1→bb2→bb3→bb4), so bb1 at response rate of 10 percent remains the optimal decision.
- With *line c*, the initial efforts to increase the response rates (10 percent→40 percent→60 percent) are counterproductive, because – besides increasing the costs – they further increase the bias (c1→c2→c3). As expected, this also increases the CUA (cc1→cc2→cc3). Only after 60 percent do the efforts to increase the response rates (toward 90 percent) finally become beneficial and they decrease the CUA (cc4). However, the best buy still remains at 10 percent (cc1).

These hypothetical, yet realistic, illustrations show how risky it is to focus only on response rates because in surveys we wish to obtain (accurate) information with given resources, and we do not necessarily focus on high response rates or low nonresponse bias. Adding the context of costs can thus radically change the conclusion from an analysis based on the isolated treatment of response rates. For example, when Lozar Manfreda et al. (2008) observed in their meta-analysis of response rates that web surveys have 10 percent lower response rates compared to mail surveys, this is a very limited and partial insight because the results would likely change if the key metric was the CUA. In this case, the cost savings from using a web survey could be invested, say, in incentives, which would further increase the corresponding response rate, thereby dramatically changing the response rate comparisons.

The decreased informative value of response rates can already be observed in the context of mixed-mode surveys, where the response rates are usually lower than in an alternative surveys based only on a single traditional mode. This is typically true when web is combined with face-to-face mode (Ainsaar et al. 2013) or when web is combined with mail (Dillman 2015). However, despite the lower response rates, researchers (and clients) still prefer mixed-mode designs because of their better cost-error performance.

Conclusions

We can summarize that with probability samples the response rates for academic face-to-face surveys of the general population have roughly stabilized (at least in Europe) in the range of 50–60 percent. Of course, this is achieved – and continuously preserved – only at the cost of increased efforts. In addition, there are considerable variations in response rates depending on type of survey and other circumstances. On the other hand, we can observe that technological changes, emerging nonprobability samples and mixed-mode surveys are creating serious problems for response rate calculations and also for their perceptions.

We also demonstrated in this chapter that costs are an extremely important issue for future response rate research, which is a much neglected research area. The explicit modeling of the relationships between response rates, response bias, accuracy, and survey costs can thus bring about important insights here, which can help practitioners in deciding whether to increase the efforts to achieve high response rates or not.

Directions that are also important for future response rate research are the efforts to provide improved definitions and calculations, as well as strategies for observing and comparing response rates. With respect to the latter, a very important challenge stems from studying and understanding different factors that may influence response rates in international surveys and comparative research.

References and Further Reading

AAPOR. (2015). The Code of Professional Ethics and Practices. Retrieved from http://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics/AAPOR_Code_Accepted_Version_11302015.aspx

AAPOR. (2016). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. Retrieved from http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Ainsaar, M., Lilleoja, L., Lumiste, K., & Roots, A. (2013). ESS mixed Mode Experiment Results in Estonia (CAWI and CAPI Mode Sequential Design). Institute of Sociology and Social Policy, University of Tartu. Retrieved from http://www.yti.ut.ee/sites/default/files/ssi/ess_dace_mixed_mode_ee_report.pdf

Berzelak, J., Vehovar, V., & Lozar Manfreda, K. (2015). Web Mode as Part of Mixed – Mode Surveys of the General Population: An Approach to the Evaluation of Costs and Errors. *Metodološki zvezki* 12(2), 45–68.

Billiet, J., & Vehovar, V. (2009). Non-Response Bias in Cross-National Surveys: Designs for Detection and Adjustment in the ESS. *Ask* 18(1), 3–43.

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web Survey Methodology*. London: Sage.

Cooper, D., & Greenaway, M. (2015). Non-probability Survey Sampling in Official Statistics. Office for National Statistics. Retrieved from http://mi.ris.org/uploadi/editor/doc/1462090294nonprobabilitysurveysamplinginofficialstatistics_tcm77-4077571.pdf

Dillman, D. (2015). On Climbing Stairs Many Steps at a Time: The New Normal in Survey Methodology. SES seminar at The School of Economics, Washington State University, September 18, 2015. Retrieved from http://ses.wsu.edu/wp-content/uploads/2015/09/DILLMAN-talk-Sept-18-2015.pdf

DiSogra, C., & Callegaro, M. (2016). Metrics and Design Tool for Building and Evaluating Probability-Based Online Panels. *Social Science Computer Review* 34(1), 26–40.

Fuchs, M., Bossert, D., & Stukowski, S. (2013). Response Rate and Nonresponse Bias -Impact of the Number of Contact Attempts on Data Quality in the European Social Survey. *Bulletin de Methodologie Sociologique* 117(1), 26–45.

Groves, R. M., & Peytcheva, E. (2008). The impact of Nonresponse Rates on Nonresponse Bias. *Public Opinion Quarterly* 72(2), 67–189.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Hass, I., & Vehovar, V. (2008). Web Surveys Versus Other Survey Modes: A Meta-Analysis Comparing Response rates. *International Journal of Market Research* 50(1), 79–104.

Rendtel, U., & Amarov, B. (2015). The Access Panel of German Official Statistics as a Selection Frame. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis

(Eds.), *Improving Survey Methods: Lessons from Recent Research* (pp. 236–249). New York and London: Routledge Taylor & Francis Group.

Roberts, C., Vandenplas, C., & Stähli, M. E. (2014). Evaluating the Impact of Response Enhancement Methods on the Risk of Nonresponse Bias and Survey costs. *Survey Research Methods* 8(2), 67–80.

Slavec, A., & Vehovar, V. (2011). Nonresponse Bias in Readership Surveys. Paper presented at 22nd International Workshop on Household Survey Nonresponse, Bilbao, September 5–7, 2011.

Štebe, J. (1995). Nonresponse in Slovene Public Opinion Survey. *Metodološki zvezki* 10, 21–37.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). Response and Nonresponse Rates in the European Social Survey. In Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (Eds.), *Improving Survey Response: Lessons Learned from the European Social Survey* (pp. 89–113). Chichester, UK: John Wiley & Sons, Ltd.

Tourangeau, R., Brick, J. M., Lohr, S., & Li, J. (2016). Adaptive and Responsive Survey Designs: A Review and Assessment. *Journal of the Royal Statistical Society. Series A.* Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/rssa.12186/full

Vannieuwenhuyze, J. T. A. (2014). On the Relative Advantage of Mixed-Mode versus Single-Mode Surveys. *Survey Research Methods* 8(1), 31–42.

Vehovar, V., Berzelak, J., & Lozar Manfreda, K. (2010). Mobile Phones in an Environment of Competing Survey Modes: Applying Metric for Evaluation of Costs and Errors. *Social Science Computer Review* 28(3), 303–318.

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). Nonprobability Sampling. In C. Wolf, D. Joye, T. W. Smith, & Y.-C. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (pp. 329–345). London: Sage.

Vasja Vehovar is Professor of Statistics at the Faculty of Social Sciences, University of Ljubljana, Slovenia.

His initial interest is in survey sampling and nonresponse, but he works in broader social science methodology issues, particularly those related to social informatics. Since the last decade, however, he predominantly focuses on web survey methodology.

Within this context he had run (1996) one of the first scientific experiments with web survey. In 1998 he co-established WebSM website, which is today the leading global resource on web survey methodology and it also received American Associating for Public Opinion Research (AAPOR) innovators award in 2009. In 2008 he launched open-source web survey software 1KA to support web survey research. In 2015 he co-authored a monograph Web Survey Methodology (book.websm.org).

Koen Beullens (PhD) is a Senior Researcher at the Centre for Sociological Research at the KU Leuven. His main research topics are nonresponse error and interviewer variance in survey research. He is also a member of central coordination team of the European Social Survey, contributing to the quality assessment of this survey, both the output as well as the process approach.

7

Choosing a Mode of Survey Data Collection

Roger Tourangeau

Survey research began in the late 1930s and the early 1940s, and, for the first several decades of its history, virtually all surveys used just two modes of data collection: mail questionnaires or face-to-face (FTF) interviews. As telephone coverage of the population improved over time, different sectors of the survey research industry began adopting telephone interviewing as a way to improve response rates relative to mail surveys and to decrease costs relative to FTF surveys. Different sectors of the survey research industry adopted telephone interviewing at different times, with the surveys done by or for the federal statistical system being the last to adopt the mode in the 1970s.

The evolution of survey technology, shown in Fig. 7.1, which is adapted from a figure by Mick Couper, shows the progression in data collection modes, beginning in the 1970s with the three traditional modes – mail, telephone, and FTF. Since then, there have been two major waves of change. In the first, computers began replacing paper as the basic medium on which surveys were conducted. For example, with telephone surveys, by the mid-1970s parts of the survey research industry had already begun switching to computer-assisted telephone interviewing (CATI). This method became very popular and CATI essentially came to replace paper telephone surveys almost entirely. Similarly, as computers got smaller and lighter, some survey

R. Tourangeau (✉)
Rockville, Maryland, USA
e-mail: RogerTourangeau@westat.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_7

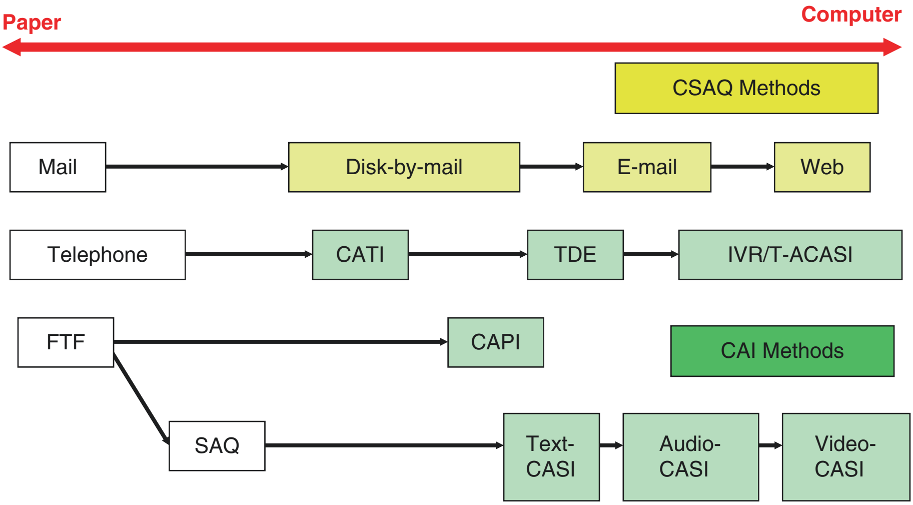


Fig. 7.1 The evolution of survey technology

organizations began sending interviewers out with computers for use with FTF surveys, an approach that came to be known as computer-assisted personal interviewing (CAPI) and it eventually supplanted paper FTF interviews almost completely as well. Surveys that had traditionally used mail were slower to adapt to computerization, but this is likely due to the relatively later advent of the Internet and e-mail.

The second wave of technological innovation came as computers displaced interviewers in favor of self-administered questionnaires. Even in the context of FTF interviewing respondents frequently interact directly with the computer when answering sensitive questions. Considerable evidence indicates that this approach leads to more accurate reports of sensitive information.

Both research and practice have demonstrated that computerization provides a major advantage across modes, enabling researchers to implement much more complex questionnaires while reducing respondent and interviewer burden. Automating the route that each respondent takes through the survey ensures that researchers are able to ask questions that are relevant to each respondent without the interviewer or respondent needing to figure out how to follow complex skip instructions (for questions that branch from core questions). Thus, computerization reduces the burden on both interviewers and respondents while providing relevant data for researchers.

Web surveys became popular as Internet coverage improved. Sometimes these surveys were adjuncts to traditional survey modes where, for example, a mail survey might invite the respondent to provide their answers online instead of on paper. Web-only survey data collection has a considerable history of controversy given its close association with non-probability sampling. Thus, none of the major National Science Foundation (NSF)-funded surveys (and few government-sponsored surveys more generally) are web-only.

Data collection decisions often imply other survey design decisions. For example, a particular method of data collection is typically yoked with a specific sampling approach; thus, most CATI surveys are conducted with random-digit dial samples. Further, with national FTF designs, cost constraints make clustered designs necessary; thus, FTF interviews are often done with area probability samples. So the mode of data collection and the sampling frame are typically bundled in a package. These bundles of features influence the entire spectrum of survey error sources, as well as the survey's cost and timeliness:

Non-observation error:

- Coverage (since each mode is linked with a sampling frame and access)
- Non-response
- Random sampling error (clustering, stratification, sample size)
- Sampling bias (e.g., with non-probability web panels)

Observation errors:

- Random measurement error
- Response order effects (primacy/recency)
- Interviewer effects on respondent reports (none in mail, some in telephone, many in FTF)
- Social desirability bias (the tendency of respondents to provide inaccurate reports to present themselves in a more favorable light)

Given the central importance of the choice of survey mode, this choice may reflect a number of important features of the survey. The first and foremost of these features is cost; FTF interviewing is extremely expensive but often organizations will sacrifice sample size for the higher response rate and better data quality that are often associated with in-person surveys. Second, different sampling frames have their own particular issues. With FTF interviews, the most common approach is area sampling due to need to cluster the sample geographically; with telephone, list-assisted frames of telephone numbers are most common; and with self-administered surveys, the frame determines whether the administration mode is mail or web, although the U.S. Postal Service's address list is coming to be used for both. Generally, web surveys are hampered by the lack of suitable sampling frames; this has prevented the bundling of mode and frame since no standard approach to sampling Internet users has yet emerged.

Coverage error is often the second most important consideration when selecting a mode of data collection. This error arises when not every unit in the population is represented on the sampling frame. If there are differences between the units that are on the frame and those that are omitted, then coverage error becomes a problem. For example, web surveys exclude those who do not have Internet access and landline-only telephone surveys exclude those who only have cell phones. Considerable evidence indicates that coverage error is a significant concern for web surveys. This coverage error manifests itself in the "digital divide" – the large number of substantial demographic and non-demographic differences between people with Internet access and those without.

The key statistical consequences of non-observation error (encompassing sampling error, coverage error, and non-response) are inflated variance and bias. Unadjusted estimates, such as means or proportions, from non-probability samples are likely to be biased estimates; similarly, estimates from any sample affected by non-response bias or coverage bias will, by definition, produce at least some biased estimates. The size and direction of the bias depend on two factors: one reflecting the proportion of the population with no chance of inclusion in the sample, and the second reflecting differences in the inclusion probabilities among different members of the sample who could in principle complete the survey. Measurement error is also influenced by mode; Couper (Chapter 5, in Groves et al. 2004) has proposed five features that help explain the impact of mode on measurement error:

1. The degree of interviewer involvement (e.g., mail and web feature low levels; CAPI high levels);
2. The degree of interaction with the respondent (e.g., eliciting opinions from a respondent vs. abstracting information from respondent records);
3. The degree of privacy (e.g., presence of interviewer, third parties, etc.);
4. Channels of communication (e.g., how questions are communicated to respondents and how they respond); and
5. Technology use (paper vs. computer).

Models of this type can be thought of as proposing mechanisms through which differences in mode may result in differences in the data. Similarly, Tourangeau and Smith (1996) have developed a model for how the data

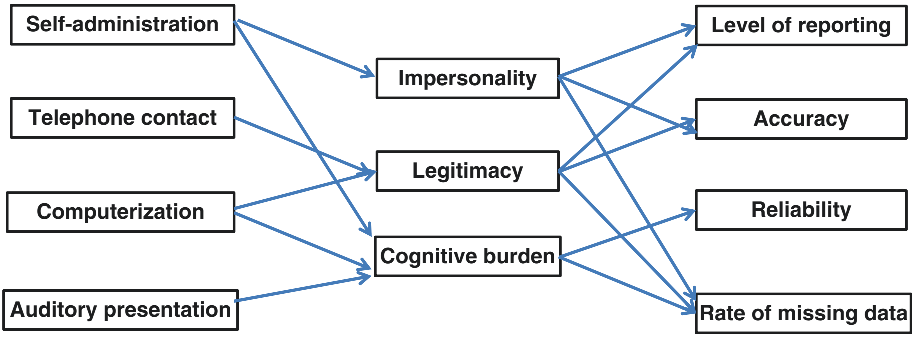


Fig. 7.2 Effects of data collection mode on features of the data

collection mode can affect various features of the data; the following version was adapted from Tourangeau et al. (2000) (Fig. 7.2).

This model implicates three psychological variables that mediate the association between data collection mode and differences in the resulting data. The first is impersonality, which is how personal or impersonal the respondent perceives the survey interaction and context to be. The second is legitimacy, which is whether or not the survey and/or sponsor seem legitimate to the respondent. And the third variable is the cognitive burden of the survey response process for the respondent.

There has been considerable discussion about potential mode differences among self-administered modes and some research has been conducted examining these potential differences. In a meta-analysis of these studies, Tourangeau and Yan (2007) found that there is no significant effect of computerization on response. This is good news for researchers because it means that they can take advantage of the full range of self-administration options without great concern about differences in the resulting data.

Another important area of data collection mode research has been on the use of mixed-mode designs. These often represent best practices with regard to reducing costs and improving response rates. For example, surveys may begin with a less expensive mode of data collection (to reduce cost) and then switch to more expensive modes (to reduce nonresponse). The last few decennial censuses in the USA have followed this model, starting with mail and following up with mail non-respondents using FTF data collection. There are a number of other ways that these mixed mode designs have been done. In some cases, one mode may be used for sampling and recruitment and another for data collection; for example, mail may be used to sample and recruit people to participate in a web survey. Other surveys have used one mode at the start of a panel survey and then changed modes in later waves. For example, the Panel Survey of Income Dynamics used FTF interviewing for the initial wave and then switched to telephone thereafter. Another design that has received considerable research attention uses different modes for different segments of the population, for example, web surveys for those with Internet access and mail for those without. Sometimes, one mode will be used to follow-up for another mode, as with the decennial censuses. In this case, data collection often begins with a cheap mode, such as mail, to recruit willing respondents and then switches to a more expensive mode, such as telephone, to recruit reluctant respondents. Longitudinal surveys may reverse this process, starting with the expensive, high-response-rate mode first to maximize recruitment and then transitioning to a less expensive mode for latter waves. Less common are approaches that implement different modes to collect different types of data from the same respondents.

One goal for some surveys using mixed-mode designs is maximizing comparability between modes; the goal is that the same person should provide the same responses to a survey conducted by any mode. This has brought about a design approach known as unimode designs. The notion behind the unimode design is that mode effects should be minimized at all costs. When implementing a unimode design, there are a number of considerations that arise that are not reflected in single-mode designs. For example, instead of optimizing the survey features for a single mode, questionnaires and procedures would need to be designed to ensure equivalence across modes. If mail is one of the modes used, then any computerized mode should attempt to mimic the design and flow of the mail questionnaire as closely as possible. This means that researchers are unable to take advantage of many of the design features that computerization permits, such as complex skip patterns and branching. Likewise, show cards should not be used in an FTF mode if that mode will be paired with a telephone survey or some other method in which the respondent cannot be presented with the show card.

However, not all researchers agree that this is the best way to conceive of mode effects, particularly when maximizing comparability between modes is a secondary concern to minimizing total error. This alternative way of thinking conceptualizes mode effects as differential measurement error. One model for better understanding the differential measurement error framework is shown in the following formula:

$$wb_A + (1 - w)\ b_{\mathrm{B}}$$

where *b* is a measurement effect, *A* and *B* are different modes, and *w* is the fraction of the sample completing the survey in mode A.

Using this model it becomes clear that making the error in *A* match the error in *B* may not result in the smallest total error. Instead, error in both modes should be minimized in order to minimize overall error. This is a different objective from the one adopted by the unimode approach; the goal is not necessarily maximizing comparability but minimizing overall error. So, under this approach, a researcher using telephone and FTF surveys would certainly want to use a show card in the FTF survey if it might reduce measurement error in that mode, even if there is no analogue to the show card in the telephone mode.

Best practices in data collection are reasonably well understood in the context of traditional survey methods. However, there may be opportunities or challenges that still need to be addressed. This is particularly true when implementing newer approaches to data collection such as cell phone surveys and surveys on mobile devices or when trying to balance comparability with minimum error in the mixed-mode context.

While differences between data collection modes have generated a lot of research attention, there is still more research that needs to be conducted to further develop our understanding of mode effects. Measuring mode effects is important but can be costly, and as a result, it is not regularly done outside of the academic context. While some work has been done on developing models to distinguish non-observation errors from observation errors in mode effects studies, more research is warranted in this area. Future research should also take advantage of opportunities to compare with gold standards such as administrative records and also should make greater use of within-subject designs. Following is a list of proposed data collection mode research topics that researchers and funding agencies should consider pursuing.

Areas for future research:

- Funding research aimed at developing methods and best practices for optimizing data collection across modes rather than mode comparison studies that are simply descriptive
- Mail/Address-Based Sampling versus Telephone/Random Digit Dialing in the changing landscape of falling response rates for telephone surveys
- Minimizing measurement error versus unimode designs for mixed mode studies

- Disentangling observation and non-observation differences in mode comparisons
- Reducing measurement error in self-administered surveys
- Identifying challenges and opportunities in new and changing modes such as cell phones, tablets, and other mobile devices

References and Further Reading

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. 2nd ed. New York: John Wiley & Sons, 2009.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tourangeau, R., & Smith, T. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, 60, 275–304.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 859–883.

Roger Tourangeau is a Vice President in the Statistics Group and co-director of Westat's Survey Methods Group. Tourangeau is known for his research on survey methods, especially on different modes of data collection and on the cognitive processes underlying survey responses. He is the lead author of *The Psychology of Survey Response*, which received the 2006 AAPOR Book Award, and he was given the 2002 Helen Dinerman Award, the highest honor of the World Association for Public Opinion Research for his work on cognitive aspects of survey methodology. He is also the lead author of *The Science of Web Surveys*. Before coming to Westat, Dr. Tourangeau worked at NORC, the Gallup Organization, and the University of Michigan; while at the University of Michigan, he directed the Joint Program in Survey Methodology for nine years. He has a Ph.D. from Yale University and is Fellow of the American Statistical Association.

8

Mixed-Mode and Mixed-Device Surveys

Edith Desiree de Leeuw and Vera Toepoel

Introduction

Mixed-mode surveys are not new and can be traced back to the early 1960s. In a mixed-mode design, researchers combine multiple data collection methods to meet the challenges of single mode surveys and improve coverage of the intended population, to increase response rates, and to reduce survey costs. Examples of these early applications of mixed-mode designs include mail surveys with a telephone follow-up to increase (single mode) mail survey response at affordable costs and face-to-face and telephone mixes to compensate for undercoverage of telephone owners in single mode telephone interviews. Mixed-mode designs really increased in popularity with the advent of online survey data collection. Web surveys have now become one of the most prominent survey data collection methods in Europe and the USA. Web surveys and especially online panels are very cost effective, have a short turnover time, and combine the advantages of self-administration with computer technology. As a result data quality in well-designed online surveys is high, especially when sensitive questions are asked. However, some major disadvantages of single mode online research are undercoverage, as not everyone has Internet access, and high rates of nonresponse. To overcome these problems, and still enjoy the advantages of web surveys, a mixed-mode

E.D. de Leeuw (✉) · V. Toepoel
Utrecht University, Utrecht, The Netherlands
e-mail: E.D.deLeeuw@uu.nl; V.Toepoel@uu.nl

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_8

approach with web surveys as one of the data collection methods in the mix is an attractive option (De Leeuw and Berzelak 2016; Tourangeau 2017).

While a mixed-mode approach may solve major coverage and nonresponse problems of online surveys, a new technological challenge is facing survey designers as mobile devices, such as, smartphones and tablets, are increasingly being used to access the Internet. Web surveys are now morphing from a computer-oriented (i.e., desktop or laptop PC) into a multi-device (i.e., PC, smartphone, and tablet)-oriented concept (Buskirk 2015; Couper et al. 2017). Many researchers doing web surveys do not necessarily think of themselves as doing mixed-device surveys and rarely account for the different types of devices that respondents are using when assessing survey errors. A mixed-device survey is not a mixed-mode survey in the traditional sense of the word. In a mixed-mode approach two disparate data collection methods (e.g., a self-administered online survey and an interviewer administered telephone survey) are combined. In a mixed-device survey, we have one overall data collection principle: a self-administered, computer-assisted (online) survey. However, respondents may choose to respond through a variety of devices. These devices not only widely vary in screen sizes, but also in data entry interface (e.g., keyboard and mouse, touchscreen, on screen keyboard), and the question arises whether or not answers obtained via smartphone and tablet are comparable to answers obtained from pc or laptop. Excluding mobile respondents may lead to serious coverage errors (see Peterson et al., 2017) and researchers should design optimal surveys to accommodate for different devices (e.g., Buskirk, 2015)

In the next sections, we first discuss the most common mixed-mode approaches and summarize the empirical findings on reducing coverage, nonresponse, and measurement error and the implications for design and analysis. We will then review the main issues in mixed-device surveys, again focusing on empirical knowledge and optimal design. We will end with recommendations and a research agenda for the future.

Mixed-Mode Surveys: Design and Implications

There are many forms of mixed-mode designs; researchers may mix contact strategies (e.g., a postal mail prenotification letter, potentially including an incentive for a web survey), or they may mix the actual data collection procedures (e.g., a web and a paper mail survey); for a detailed overview see De Leeuw (2005). Here we will discuss mixed-mode design in its strictest sense: the use of multiple methods of data collection within a survey. Two main implementation strategies can be applied: concurrent and sequential mixed-mode surveys. In a concurrent mixed-mode design, two or more data collection methods are offered at the same time; for instance a web survey offered together with a paper mail survey or a telephone interview. The main reason for a concurrent mixed-mode approach is to overcome coverage problems and include those not on the Internet (e.g., elderly, lower educated). A special form of concurrent mixed-mode is encountered in international studies, as different countries have different survey traditions and a mixed-mode design across countries is the only practical solution. In many cases, standardization and restriction to a single mode of data collection may result in a sub-optimal design (e.g., poor sampling method) for some countries, which may even threaten comparability. A good example of the need for a mixed-mode approach across countries is the International Social Survey Program that started out as a single mode self-administered paper questionnaire, but when more countries joined in a mixed-mode design was implemented allowing face-to-face interviews for low literacy countries.

In a sequential mixed-mode survey, one data collection method is offered after another, in order to improve coverage and response. The most common sequential mixed-mode design starts with the least expensive mode (e.g., mail or web) and follows up with more expensive modes (telephone and/or face-to-face). A well-known example is the American Community Survey. In panel research, a different sequential approach is often used; there the most expensive interview mode is used first for the recruitment interviews or first panel wave to guarantee a high response for the baseline survey. Data for subsequent waves are then collected with a less expensive mode. This design has proved to be successful for the establishment of probability-based online panels. Since there are currently no sampling frames for the population of Internet users, a probability sample is drawn using a well-established sampling frame (e.g., of street addresses or postal delivery points) and an interview survey is used for recruitment to the online panel. A prime example is the pioneering work of the Dutch online Longitudinal Internet Studies for the Social Sciences (LISS) panel, where a probability sample of Dutch households was recruited using the face-to-face mode. To reduce coverage error, the LISS-panel offered a free Internet connection and a simple PC to those who had none.

A slightly different approach was used by the GESIS-Leibnitz Institute for the establishment of the German GESIS panel. Similar to the LISS-approach, a probability-based sample was recruited using face-to-face interviews; however, those without Internet were not offered an Internet connection, but in the next waves were surveyed using postal mail surveys, while those with Internet were surveyed online. In other words, after recruitment, the GESIS panel uses a concurrent online-paper-mail approach.

Whether or not mixing modes improves response rates depends on the type of design used. Sequential mixed-mode designs do work and switching to a second, or even third mode in a sequential mixed design has proven to increase response rates in studies of the general population as well as for special populations (De Leeuw and Berzelak 2016). However, a consecutive approach does not clearly increase response rates. While offering two modes and giving the respondents a choice has an intuitive appeal – it appears respondent friendly since respondents themselves can decide what is most suitable to them – it also increases the respondent burden. When presented with a mode choice, respondents have to make two decisions instead of one: not only whether or not to respond, but also through which mode if they do decide to participate. Furthermore, the choice dilemma may distract from the researchers' carefully formulated arguments on the importance and saliency of the survey (De Leeuw and Berzelak 2016). As a result, Tourangeau (2017) advises researchers not to offer respondents a choice and to prevent them from procrastinating with carefully scheduled multiple contacts, such as reminders or a sequential mixed-mode approach. From a cost perspective it pays to start with the most cost effective method and reserve more expensive modes for the follow-up. Regarding the improvement of coverage, empirical studies are scarce. In their review, De Leeuw and Berzelak (2016) conclude that different modes do bring in different types of respondents and do improve representativity.

Mixed-mode surveys may reduce coverage and nonresponse error, but what about measurement error? There has been a long tradition of empirical mode comparisons and they all point to small but systematic differences between interviewer-administered and self-administered surveys. These differences may influence the overall measurement error in a mixed-mode design. From a Total Survey Error perspective, researchers wish to reduce all survey errors, including measurement error. There are two general approaches to designing questionnaires for mixed-mode and mixed-device surveys. The first approach is the unified or unimode design, where the goal is to produce equivalent questionnaires in each mode. An example is using a series of yes/no questions in both online and telephone interviews, instead of a yes/no format in telephone and a check-all-that-apply format online. The second approach is to try to optimize each mode independently in order to minimize overall measurement error; this approach could result in different question formats and implementation procedures for each mode. The latter approach is only desirable when one overall population estimate is needed and for factual questions only since attitudinal questions are more susceptible to question format effects (Tourangeau 2017). When the goal of the survey is the comparison of groups, researchers should try to minimize mode measurement effects by design and use equivalent questionnaires. This is extremely important in cross-national studies, where different modes are used in different countries, mixed-mode longitudinal studies, and multi-site studies (e.g., schools, hospitals). But also in cross-sectional studies subgroups are often compared and if certain subgroups are overrepresented in a certain mode or device use (e.g., younger more online and/or younger more mobile phones), nonequivalent questionnaires over mode or over device may threaten the validity of the comparisons.

Designing equivalent questionnaires does *not* mean regression to the lowest common denominator. De Leeuw and Berzelak (2016) summarize the design principles of Dillman and illustrate these with two examples. When self-administered and interview surveys are mixed, there are two mode-inherent differences: (1) availability of interviewer help and probes or not, and (2) the sequential offering of questions in an interview versus grouped questions (e.g., in a grid) in a self-administered form. De Leeuw et al. (2016) showed that it is possible to successfully emulate interviewer probes in an online survey and by doing this implement an interviewer procedure in an online self-administered questionnaire. The second example (sequential offering versus grid questions) is of importance for both mixed-mode and mixed-device studies. In online questionnaires, a set of similar questions or statements are often presented together in a matrix (grid) format. The advantages of grid questions are that the response format saves space, the questionnaire appears to be shorter, and respondent burden is relatively low because respondents do not generally have to click the next button as often. A main disadvantage is that respondents often do not pay as much attention to each question separately as they do when questions are offered sequentially and are more prone to satisficing behavior (e.g., straightlining). A new online question format, the so-called auto-advance or carrousel question, does present questions one-by-one as in an interview, but because of the auto-advance function there is no extra respondent burden. After the respondent has given an answer, the next question automatically appears on the screen, mimicking an interviewer-administered survey. Auto-advance questions have proved themselves in online and online-interview mixes. This format may also be promising for mixed-device surveys as grid questions are burdensome on mobile phones. For a detailed description and examples, see De Leeuw and Berzelak (2016).

Careful expert design of multiple mode surveys improves quality and helps prevent unwanted mode-measurement effects (e.g., more do-not-know answers or missing data and less differentiation in online surveys). Still, the data may contain mode inherent measurement effects. Consequently, researchers should always try to estimate mode differences and, if these occur, adjust for mode measurement effects in the data. Several statistical methods for estimation and adjustment have been proposed and are still under development. For an introduction and overview, see Hox et al. (2017).

Mixed-Device Surveys

Mixed-device surveys are a unique sort of concurrent mixed-mode surveys since online surveys are being completed on a range of different devices that respondents can choose at their own convenience. It is important to distinguish between mobile phone, tablet, laptop, and desktop PC devices since they differ in several dimensions such as the size of the screen, technology features (e.g., processing power, connectivity, method of navigation), user characteristics, and context of use (Couper et al. 2017). Mobile penetration rates differ greatly per country. But simply possessing a mobile device does not necessarily mean that people use their mobile device for survey completion. For example, in 2013 in the Netherlands, the majority, about three out of four people, owned a mobile phone with Internet access. Only about 11 percent used their mobile device for survey completion in the Dutch LISS Panel (2 percent mobile phone and 9 percent tablet); similar rates are found for the GESIS-panel in Germany. However, with a clear invitation for mobile phone use and a mobile-friendly (optimized) design the percentage of mobile phone completion can increase to 57 percent (Toepoel and Lugtig 2014).

Survey software is increasingly adapting to the demands presented by mobile survey responding via implementations of responsive survey designs. The software detects the device being used to access the survey and optimizes the format accordingly. Browser-oriented online surveys can either use responsive design and be optimized for mobile devices or involve no optimization and be designed for completion on computers (with only the possibility of being completed on mobile devices without optimization). Optimization for mobile devices can involve shorter question text, other types of response options (sliders, tiles), and formats (no grids). Most market research organizations have changed their format into a responsive (optimized) design. Other online surveys use apps. They need more action from the respondents since they have to be installed on the respondent's own device. The main advantage of mobile apps is that they give researchers more control over the design of the online surveys. However, separate versions of these apps must be designed for different platforms such as Android or iOS, and the respondents must be willing to download these apps.

Lynn and Kaminska (2013) propose a theoretical framework of ways in which mobile surveys may differ from computer-assisted web surveys, including issues such as multi-tasking, distraction, the presence of others, and differences inherent in the technology such as input mode (e.g., clicking on a PC versus touching on a mobile device). Empirical research on mixed-device surveys either uses a natural setting in which respondents can choose their own device for completing a survey spontaneously, or an experimental design in which respondents are assigned to use a particular device. Some find differences between mobile phone, tablet, and regular desktop PC respondents including longer survey completion times, lower unit and higher partial and item nonresponse rates, shorter open responses and different personal characteristics for mobile responses compared to the other devices, while others find no differences between devices. In general, response rates for mobile online surveys are lower than for PC and there is evidence for a higher mobile break-off rate. Furthermore, surveys take longer to complete on mobile devices both for optimized and nonoptimized mobile surveys. Positive is that there is little evidence for lower data quality in mobile surveys. For a detailed summative review of research on mixed devices, see Couper et al. (2017). Also the cognitive processing between PC-administered web surveys and mobile web surveys appears to be similar. Lugtig and Toepoel (2015) demonstrate by using consecutive waves of a panel that measurement errors do not change with a switch in device within respondents.

The main differences between mobile and PC surveys lie in the way the survey invitation can be send (text versus e-mail), survey length, question format, and the possibility of measuring without asking questions. Text is faster for mobile and designed survey length is ideally shorter for mobile phone completion. Grids or matrix questions should be eliminated since they are too difficult to render in an equivalent manner on small screens and larger screens. Tiles, in which entire areas of question text are clickable are preferable for mobile phones since they give more area to tap in comparison with traditional (online survey) radio buttons. In addition, passive data collection offers new opportunities for mobile devices.

Mobile data can be collected from respondents while they are on the go as well as passively collected data. Examples of passively collected data include user agent strings, biomarkers, and GPS coordinates. While passive data collection still requires initial permission from the respondents, they are generally collected without the respondent having to provide direct answers to survey questions (Buskirk 2015). This passive data collection not only reduces respondent burden, but can also reduce measurement error since they are collected on the spot and are less susceptible to recall and estimation bias.

Future Research

Society and technology are continuously changing and our data collection methods are changing accordingly. Online surveys were pioneered at the beginning of the twenty-first century; probability-based online panels started in 2007 and are now established in both Europe and the USA. Mixed-mode surveys and mixed-device surveys show promise to answer the challenges of single mode surveys and improve response and data quality at affordable cost. However, combining several modes or devices in one survey also has implications for questionnaire design and analysis and we have summarized the challenges and best practices previously from a Total Survey Error perspective. It is evident that more research is still necessary. As suggested by Buskirk (2015), to further understand survey errors in both mixed-mode and mixed-device surveys, we need experiments that compare question formats both within and across modes and devices to understand mode effects. Researchers should focus on disentangling effects that are associated with self-selection, question design, and mode/device inherent factors. Future research should emphasize the minimization of measurement error across modes and devices. Research on adjustment for measurement error is still under development and at present need detailed auxiliary data and complex statistics. Further research in this field is of great importance (see also, Tourangeau 2017; Hox et al. 2017).

The mobile society also has consequences for attention span, multitasking, and changing societal patterns. Respondents do not want to spend a lot of their precious free time on surveys; furthermore mobile devices are typically used for short messaging. As a consequence, the optimal survey duration might be shorter for mobile surveys. Short surveys, or if this is not possible, multiple measures using data chunking, in which a questionnaire is divided and administered in several smaller parts, may help to increase response rates for online, mixed, and mobile surveys. How this affects data quality is a matter of further investigation.

Finally, we have entered the world of big data and passive measurement (see Callegaro and Yang and Lessof and Sturgis, this volume). Sometimes respondents are aware of this, as they are requested to download specific apps. Many respondents still refuse to take part in these measurements and are, for instance, concerned about privacy issues; how to overcome their reluctance is of great importance. Often big data are harvested without the active awareness of respondents. Both forms involve privacy concerns that should be addressed. Finally, harvested big data are usually not collected with a primary research question in mind. How to address the validity of big data studies, what are the lacks in the obtained information, and how to decide and design for additional surveys are high on the research agenda.

Areas for future research:

- Experiments into optimizing question formats and reduce measurement error across modes and devices
- Disentangling (self) selection and measurement effects in mixed-mode and mixed-device studies
- Further development of adjustment method in general
- Development of adjustment methods that are applicable in daily survey practice
- Applicability and consequences of implementing short surveys, segmented surveys, and data chunking
- Investigating the use of apps and sensors (GPS, health) to reduce the number of questions being asked in a survey

References

Buskirk, T. D. (2015) “The Rise of Mobile Devices: From Smartphones to Smart Surveys.” *The Survey Statistician*, 72, 25–35. Available at http://isi-iass.org/home/wp-content/uploads/N72.pdf

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile Web Surveys: A Total Survey Error Perspective. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N.C. Tucker, & B. T. West (eds.), *Total Survey Error in Practice* (Chapter 7). New York: Wiley.

De Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21(2), 233–255. Freely available at http://www.jos.nu/Articles/abstract.asp?article=212233.

De Leeuw, E. D., Hox, J. J., & Boeve, A. (2016). Handling Do-Not-Know Answers: Exploring New Approaches in Online and Mixed-Mode Surveys. *Social Science Computer Review*, 34(1), 116–132.

De Leeuw, E. D., & Berzelak, N. (2016). Survey Mode or Survey Modes?. In C. Wolf, D. Joye, T. W. Smith, & Y.-C. Fu (eds.) *The Sage Handbook of Survey Methodology* (Chapter 11). Los Angeles: Sage.

Hox, J. J., De Leeuw, E. D., & Klausch, T. (2017). Mixed mode Research: Issues in Design and Analysis. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N.C. Tucker, & B. T. West (eds.), *Total Survey Error in Practice* (Chapter 24). New York: Wiley.

Lugtig, P., & Toepoel, V. (2015). The Use of PCs, Smartphones and Tablets in a Probability Based Panel Survey. Effects on Survey Measurement Error. *Social Science Computer Review*, 34 (1), 78–94.

Lynn, P., & Kaminska, O. (2013). The Impact of Mobile Phones on Survey Measurement Error. *Public Opinion Quarterly*, 77 (2), 586–605.

Peterson, G., Griffin, J., LaFrance, J., & Li, J. J. (2017). Smartphone Participation in Web Surveys: Choosing Between the Potential for Coverage, Nonresponse, and Measurement Error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N.C. Tucker, & B. T. West (eds.), *Total Survey Error in Practice* (Chapter 10). New York: Wiley.

Toepoel, V., & Lugtig, P. (2014). What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users. *Social Science Computer Review*, 32 (4), 544–560.

Tourangeau, R. (2017). Mixing Modes: Tradeoffs among Coverage, Nonresponse, and Measurement Error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N.C. Tucker, & B. T. West (eds.), *Total Survey Error in Practice* (Chapter 6). New York: Wiley.

Edith Desiree de Leeuw is MOA Professor of Survey Methodology at the Department of Methodology and Statistics, Utrecht University (http://edithl.home.xs4all.nl/). She was Fulbright scholar with Don Dilmann, Washington State University and visiting scholar with Jan de Leeuw (no relative) at UCLA, and was awarded the Visiting International Fellowship of the University of Surrey. Edith has over 140 scholarly publications and is co-editor of 4 internationally renowned books on survey methodology: *The International Handbook of Survey Methodology, Total Survey Error in Practice, Advances in Telephone Methodology, Survey Measurement and Process Quality*. Her recent publications focus on mixed-mode studies and online surveys, nonresponse, total survey error, and cross-national research. Edith is currently associate editor for the Journal of Official Statistics (JOS) and editor of MDA

(Methods, Data, Analyses); she is on the editorial board of international journals in the field of survey methodology, such as IJPOR, Field Methods, SMR, & BMS.

Vera Toepoel is Assistant Professor at the Department of Methods and Statistics at Utrecht University, the Netherlands. She did her PhD on online survey design at Tilburg University and is the chairwoman of the Dutch Platform for Survey Research. Her research interests include the entire survey process, from recruiting participants, designing the instrument, and correcting for possible biases. She published her work in numerous journals, for example, Public Opinion Quarterly, Sociological Methods and Research and Social Science Computer Review. Vera is the author of the book “Doing Surveys Online” published by Sage. In addition, she has worked on chapters in several handbooks of survey methodology.

9

The Use and Effects of Incentives in Surveys

Eleanor Singer

Introduction

Twenty years ago there was a consensus that incentives should not be used for surveys that were less than an hour in length. There was also great debate about whether response rates to some of the large national household surveys were declining or not. Today there is no doubt that response rates are declining even for the best-managed, costliest, most important surveys, and incentives are used in most of them. In the ANES, PSID, and GSS, which are the largest surveys funded by NSF, the largest portion of nonresponse is attributable to refusals rather than noncontacts. Monetary incentives, especially prepaid incentives, are capable of reducing nonresponse, primarily through reducing refusals. However, very little is known about the effects of incentives on nonresponse bias, which signals an important area of future research that the NSF should consider funding.

Survey research has expended considerable effort and research funds in examining the effects of incentives on a variety of outcomes. These outcomes include response rates in different types of surveys, sample composition, response quality, and response distributions. Much of this research has been conducted in the context of experiments attempting to improve

E. Singer (✉)
Survey Research Center, Institute for Social Research, Ann Arbor, MI 48106, USA
e-mail: elsinger@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_9

response rates. The findings presented in this section may not apply uniformly to all surveys. The surveys that these findings are most applicable to are

- Large, usually national surveys done for purposes related to social research
- Often longitudinal
- Typically funded by government statistical agencies or research organizations supported with government research grants
- Results are intended to be generalizable to a defined population

Market research, customer satisfaction surveys, polls with a field period of a week or less, and similar surveys are not included and the findings presented next may not apply.

One of the most consistent findings of research on survey incentives has been that prepaid incentives increase response rates; this has been demonstrated across survey modes. In the cross-sectional mail mode, a meta-analysis (Church 1993) found that prepaid incentives yielded significantly higher response rates than promised or no incentives, monetary incentives yielded higher response rates than other gifts, response rates increased with increasing amounts of money, though not necessarily linearly. Edwards and colleagues (2002) reported very similar results in a subsequent meta-analysis and, with very few exceptions, more recent experiments have yielded similar findings.

Two meta-analyses of experiments with interviewer-mediated surveys (Singer et al. 1999b; Cantor et al. 2008) found that while the results were generally similar to those in mail surveys, the effects of the incentives were generally smaller. More specifically, Cantor and his colleagues present a number of findings regarding incentives in interviewer-mediated surveys:

- Prepayment of $1–5 increased response rates from 2 to 12 percentage points over no incentives
- Larger incentives led to higher response rates, but at a decreasing rate
- The effect of incentives has not declined over time, but baseline response rates have dropped substantially
- Prepaid incentives used during refusal conversion had about the same effect as those sent at initial contact, but at a lower cost
- Promised incentives of $5 and $25 did not increase response rates; larger incentives sometimes did
- More recent experiments involving interviewer-mediated surveys, including face-to-face surveys, have found similar patterns of results

Longitudinal surveys have typically made use of incentives as part of a larger motivational package designed to both recruit and retain respondents. Similar to findings in cross-sectional surveys, incentives in longitudinal surveys increase response rates, usually by reducing refusals but occasionally by reducing non-contacts (McGrath 2006). A considerable number of studies have indicated that an initial payment may continue to motivate respondent participation in subsequent waves, meaning that an up-front investment in incentives may have greater effect for longitudinal surveys than cross-sectional ones (Singer and Kulka 2002; McGrath 2006; Creighton et al. 2007; Goldenberg et al. 2009). Other research has indicated that prepaid incentives in longitudinal surveys may increase response among those who have previously refused, but not among those who have previously cooperated, this may indicate a "ceiling effect" (Zagorsky and Rhoton 2008). Further, a study by Jäckle and Lynn (2008) found (1) incentives at multiple waves significantly reduced attrition in all waves; (2) they did so proportionately among certain subgroups and so did not reduce attrition bias; (3) the effect of the incentive decreased across waves; (4) incentives increased item nonresponse; and (5) nevertheless, there was a net gain in information.

Recently, some research has examined the effects of incentives on response quality. Typically, item nonresponse and the length of answers given to open-ended questions are used to measure response quality, but other measures would be desirable to assess accuracy and reliability. There are two alternative hypotheses about the effect of incentives on response quality. One posits that the respondent perspective is "You paid me some money and I am going to do this survey, but I am not going to work very hard at it." The second hypothesis is that respondents feel they have an obligation to answer the survey and do their best to answer correctly. Most research has found little to no support for the notion that incentives influence response quality; only one study found that incentives increased item nonresponse across waves in a panel study but decreased unit nonresponse, resulting in a net gain of information (Jäckle and Lynn 2008). Cantor and his colleagues (2008) argue that the two hypotheses need to be tested controlling for factors such as survey topic, size, and type of incentive (e.g., prepaid, promised, refusal conversion), and whether studies are cross-sectional or longitudinal. For this, a much larger pool of studies would be required and this is an area warranting future research.

1. Medway (2012) examined this question using a very large pool of measures of effort (e.g., item nonresponse, length of open-ended responses, straightlining, interview length, underreporting to filter questions, lack of

attention to question wording, use of round numbers, order effects, etc.) as well as the potential interaction of a number of demographic characteristics with receipt of an incentive. The findings of this study indicated that there were significant differences on only two effort indicators – reduced item nonresponse and less time to complete; neither was significant once cognitive ability and conscientiousness were controlled. There were also no significant interaction effects between demographics and incentives on an index of satisficing. But, because this study was implemented using a telephone survey, an important research question that remains is whether or not the same results would be found in a self-administered survey context. The study by Jäckle and Lynn (2008) found greater effects of incentives on unit and item nonresponse in mail than in phone administration of same survey, indicating that the potential interaction effect of incentives and mode of data collection on data quality needs further research. Additional aspects of quality, such as the potential effects of incentives on reliability and validity, also need study. Some further areas in need of research are discussed in the paragraphs that follow. Incentives have been shown to influence sample composition, meaning that the characteristics of people recruited are altered when incentives are used. Some of the characteristics of the sample that have demonstrated differences in response to incentives are education, political party membership, social-economic status, and civic duty. However, the majority of studies reporting these findings have done so as *ex post facto* explanations. Specific attempts to use incentives to improve response rates among certain categories of respondents who may be less disposed to respond because of their lower interest in the survey topic have received only qualified support. Importantly, no studies have looked at the effect of incentives targeted to refusals. Theoretically, one would expect such targeted incentives to be more successful in changing the composition of the sample, thereby potentially reducing nonresponse bias, so this is an area ripe for future research.

2. Another aspect of incentives that has generated some controversy is that of differential incentives. Differential incentives refer primarily to refusal conversion payments, which are typically higher than prepaid incentives. Two arguments have been made in favor of differential incentives. First, they are more economical than prepaid incentives, and second, they are more effective in reducing bias. Historically, the primary argument against using differential incentives is that they are unfair. However, economists argue that differential payments are fair; those who refuse consider the survey more burdensome and therefore need/are entitled to bigger

incentives. Respondents who are informed about differential incentives consider them unfair, but say they would respond to a new survey by same organization even when told it engages in the practice (Singer et al. 1999a). Experimental research indicates that these respondents do indeed respond to a survey purportedly by another organization a year later; there is no statistically significant difference in response by receipt of an incentive or perception of fairness. The research on differential incentives has generated two recommendations for best practice. First, survey organizations should offer small, prepaid, incentives to all sample members; this will increase sample size and help satisfy the fairness criterion. Second, they should offer differential incentives to those who refuse (or a subsample) for bias-reduction reasons, but this practice should be accompanied by research to detect whether or not refusal conversion actually reduces bias.

3. To maximize the value and return from incentives, pretesting is extremely helpful. Different people may be motivated by different appeals; research is needed to find out which are most effective for a particular study. This is true at the individual-study level and in a more general sense across survey research. Researchers should also test the effectiveness of different combinations of appeals in introductory materials, including, but not limited to, monetary incentives. For large and expensive surveys, a pretest that can yield quantitative estimates of likely response and the effectiveness of incentives, by important subgroups, may be warranted. Researchers should also take care to use pretesting to investigate respondents' and nonrespondents' perceptions of the costs and benefits of survey participation. The goal of such research is to develop empirically based efforts to improve the survey experience. Incentives are a part of this equation but the net benefits extend well beyond simply informing how to best spend incentive money.
4. More research is needed on how best to use incentives to bring about decreases in nonresponse bias for the most important dependent variables in a survey. Since all prior studies have used prepaid incentives, one recommendation is to focus research on targeted refusal conversion payments instead or in addition. Another recommendation for future research is to explore using address-based sampling rather than RDD to draw the initial sample for telephone surveys, sending letters with prepayment to a random subsample, and measuring nonresponse and nonresponse bias in both groups. A number of studies have shown that advance letters including incentives can substantially increase

response in telephone surveys (letters without incentives do not appear to have such effects). However, the percentage of RDD sample members for whom addresses can be obtained is limited, and they tend to differ from those for whom addresses cannot be obtained. As a result, this tactic results in recruiting more respondents like those who would have been recruited even without the letters (Curtin et al. 2005), thus minimally affecting nonresponse bias.

5. Another important area for future research should measure long-term effects of incentives on public willingness to participate in research going forward by adding questions about expectations for incentives to a sample of existing cross-sectional surveys (e.g., GSS, Surveys of Consumers). There is no evidence that the increasing use of incentives has had long-term effects on such willingness, but existing studies have looked at change over short intervals and with panel respondents, who may consider multiple waves as one survey.
6. Additional future research is also needed to examine changing interviewer expectations about the use of incentives and the effect of these on their response rates. It is plausible to assume that interviewers' expectations will change over the long run as a result of their experience with the increasing use of incentives. The decline in response rates over the past 15 years may in part reflect changing interviewer expectations and behavior, cultivated by reliance on monetary incentives. To shed light on whether and how motivations for survey participation are changing, it would be useful to sponsor systematic inquiry over time into reasons for responding and not responding, using experiments and open-ended questions. Do motives differ by age, gender, ethnicity, race, and income? Are altruistic motives declining?
7. There is no good evidence that monetary incentives reduce response rates, but there are indications that there may be ceiling effects (Zagorsky and Rhoton 2008; Groves et al. 2000). Why should this be? Why are incentives not simply additive with other motives for responding?
8. Research is also needed to find out if incentives are coercive. Do they have undue influence on sample members' decisions about survey participation, in the sense of inducing them to undertake risks they would not otherwise take? Research so far suggests it does not, but experiments are needed that employ a wider range of incentives and a greater variety of risks among differentially susceptible populations.
9. Finally, research is needed on the cost-effectiveness of incentives compared with other efforts to increase response rates and reduce nonresponse bias.

References and Further Reading

Cantor, D., O'Hare, B. C., & O'Connor, K. S. (2008). The Use of Monetary Incentives to Reduce Nonresponse in Random Digit Dial Telephone Surveys. In *Advances in Telephone Survey Methodology* (pp. 471–498). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9780470173404.ch22

Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly*, *57*(1), 62. http://doi.org/10.1086/269355

Creighton, K. P., King, K. E., & Martin, E. A. (2007). *The Use of Monetary Incentives in Census Bureau Longitudinal Surveys* (No. Survey Methodology - #2007-2). *Census Report Series*. Washington DC: U.S. Census Bureau.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly*, *69*(1), 87–98. http://doi.org/10.2307/3521604?ref=search-gateway:caa52ab457186bf1ab67030aea262727

Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Pratap, S., Wentz, R., & Kwan, I. (2002). Increasing response rates to postal questionnaires: systematic review. *British Medical Journal*, *324*(7347), 1183–1185.

Groves, R. M., Singer, E. & Corning, A. (2000). Leverage-salience theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64, 299–308.

Goldenberg, K. L., McGrath, D. E., & Tan, L. (2009). The Effects of Incentives on the Consumer Expenditure Interview Survey (pp. 5985–5999). Presented at the Annual Meeting of the American Association for Public Opinion Research.

Jäckle, A. E., & Lynn, P. (2008). Respondent incentives in a multi-mode panel survey: cumulative effects on nonresponse and bias.

McGrath, D. E. (2006). An Incentives Experiment in the U.S. Consumer Expenditure Quarterly Survey (pp. 3411–3418). Presented at the ASA Section on Survey Research Methods.

Medway, R. L. (2012). *Beyond Response Rates: The Effect of Prepaid Incentives on Measurement Error*.

Singer, E., & Kulka, R. A. (2002). Paying Respondents for Survey Participation. In M. Ver Ploeg, R. A. Moffitt, & C. F. Citro (Eds.), *Studies of Welfare Populations: Data Collection and Research Issues* (pp. 105–127). Washington DC: National Academies Press.

Singer, E., Groves, R. M., & Corning, A. D. (1999a). Differential Incentives: Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation. *Public Opinion Quarterly*, *63*(2), 251–260. http://doi.org/10.2307/2991257?ref=search-gateway:e350d899bd9b4af2cfe85f85b5a78443

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T. E., & McGonagle, K. A. (1999b). The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys. *Journal of Official Statistics*, *15*(2), 217–230.

Zagorsky, J. L., & Rhoton, P. (2008). The Effects of Promised Monetary Incentives on Attrition in a Long-Term Panel Survey. *Public Opinion Quarterly*, *72*(3), 502–513.

Eleanor Singer's research focused on the causes and consequences of survey non-response. She published widely on the role of incentives in stimulating response to surveys, especially among respondents for whom other motivating factors, such as interest in the topic or connection to the sponsor, are lacking. Her research also explored how concerns about confidentiality, and informed consent procedures more generally, affect survey response. The recipient of a PhD in sociology from Columbia University, at the time of writing, she was Research Professor Emerita at the Survey Research Center, Institute for Social Research, University of Michigan, a past Editor of *Public Opinion Quarterly*, a past President of the American Association for Public Opinion Research, and the author or co-author of numerous books and articles on the survey research process. **Editor's note:** shortly prior to the publication of this book, Eleanor passed away. Her loss is felt widely in the survey research community.

10

Methods for Determining Who Lives in a Household

Kathleen Targowski Ashenfelter

Many probability sample surveys begin with a series of questions intended to elicit rosters of all household members at an address so that a random selection can be made among these members. Other surveys use roster techniques as the basis for determining which household members need to answer the survey questions and which do not qualify as respondents, while still others generate rosters in order to help respondents more accurately count the people living at the sample address. This chapter generalizes across surveys with different goals, but makes the case that an accurate roster is necessary for most survey goals.

Commonly used rostering approaches, typically based on who “usually” lives or stays at a residence, are generally assumed to be effective for their purpose, but they have some problems. The rules used to determine who should be considered a household member, and who should not, are remarkably unstandardized across U.S Census Bureau surveys and most other large surveys that utilize these procedures. For example, the Census Bureau employs different rules in different surveys. Even within a single survey, different instructions for interviewers and respondents sometimes contradict one another. However, these inconsistencies have often been warranted, due to the differing goals among surveys. In some cases, differences among rules

K.T. Ashenfelter (✉)
Senior data scientist on the Cyber Analytic Services, Unisys Corporation,
Washington D.C., USA
e-mail: ktashenfelter@yahoo.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_10

for determining who should be counted as a resident in a household may be sensible, depending on the purpose of, or the constraints faced by, certain surveys. Thus, household rostering is an arena in which new conceptual and operational research is warranted and could help survey researchers to optimize a procedure that is used across a large number of surveys, including for the decennial Census.

While determining how many people are living at a sample address seems like it should be a fairly straightforward and simple task, there are a number of problems that survey designers must anticipate when formulating these rules. For example, some people may not live at a particular address all the time, such as retirees who spend summers and winters living in different states, or college students who may live at school for part or most of the year.

Most research on rostering has been based on the methodology used by four major surveys conducted by the U.S. Census Bureau: the Decennial Census, the American Community Survey (ACS), the Survey of Income and Program Participation, and the Current Population Survey (e.g. National Research Council, 2006). Although these surveys are all presumably designed to ascertain equivalent lists of all people living at each selected address (with each person being assigned to only one address, to prevent double-counting in principle or in practice), these four surveys employ different procedures and, therefore, seem likely to yield different results. More importantly, each individual survey occasionally offers instructions to respondents and interviewers that are logically inconsistent with one another, and the instructions sometimes entail the use of terms and concepts that are not sufficiently or clearly defined to allow respondents to easily and uniformly interpret the instructions and then comply with them whilst generating their responses. Research is needed that directly compares the outcomes from different approaches to generating a household roster so that we can assess whether or not different approaches aimed at achieving the same result are actually effective, or whether key differences arise in the rosters drawn under these different methods.

One of the major issues that has been identified with current practice in household rostering is the presence of inconsistencies with respect to the date or dates used as temporal reference points for determining whom should be listed as a resident. For example, for the 2010 Decennial Census, respondents who were having trouble filling out the paper form had the option to call in and provide their responses to the Census over the telephone. Some instructions for interviewers tell them to count people who were residing in the household on April 1, 2010, such as: "count the people living in this house, apartment or mobile home on April 1, 2010." But then in another portion of the same set of instructions, the language changes to become less specific, instructing the interviewer, for example to "indicate foster children if they were staying at the address on or around April 1." In the latter example, the modification to include dates around the target date is a significant departure from the original instruction. The ACS contains similar inconsistencies in its set of instructions for generating the roster. In one section, the instruction to respondents says to "Include everyone who is living or staying here for more than two months, include yourself if you are living here for more than two months, including anyone else staying here who does not have anywhere else to stay, even if they are here for less than two months." Thus, in a single section, the instructions provide contradictory directions to the respondents for which people they should include on the roster.

Another common area of roster-related ambiguity involves how a survey should enumerate college students and military personnel. For example, the 2010 Decennial Census instructions indicated that "college students and armed forces personnel should be listed where they live and sleep most of the time." But "most of the time" is never defined, leaving the respondent (and/or interviewer depending on the mode of administration) to arbitrarily determine who should be counted. Similarly, in the 2011 ACS, Interviewers using the Computer-Assisted Personal Interview mode were directed to read the following sequential list of instructions and questions to respondents:

- I am going to be asking some questions about everyone who is living or staying at this address. First let's create a list of the people, starting with you.
- Are any of these people away *now* for more than two months, like a college student or someone living in the military?
- Is there anyone else staying here even for a short time, such as a friend or relative?
- Do any of these people have some other place where they usually stay?

The aforementioned instructions give no temporal reference point or any definitions for what constitutes these rather vague concepts of time period. That is, there is no clearly defined set of time-based metrics that the interviewer or respondent can use to determine what a "short time" or "usually stay" means. These terms could mean very different things to many respondents to a particular survey, leading to differences in the final inclusion or exclusion of individuals in the resulting household roster.

Other problems exist beyond inconsistency issues. One example is that instructions to respondents about whom to count and whom to exclude are often vague. Additionally, some survey instructions are intentionally designed as a feature of the instrument that is only seen by interviewers and never shown to respondents during the survey interview. From a methodological standpoint, this asymmetry in availability of rostering information could impact data quality. From a human factors and usability standpoint, the additional context found in these interviewer instructions could be extremely helpful for respondents while they are answering the roster questions. Another common issue is that household rostering procedures are often unnecessarily complicated and include complicated branching patterns, which increases the opportunity for mistakes.

Roster complexity is an important, although often overlooked, contributor to the relative ease of use of a survey instrument and is a concept that warrants in-depth research. American households and living situations can be very complex. Rostering rules typically attempt to account for this complexity by providing instructions to interviewers and respondents for how to accurately determine whom to actually count as a member of the household. There are many living situations that increase the difficulty of building household rosters accurately according to the given set of residence rules, including the following common issues, which do not reflect a complete set of the diverse circumstances represented across American households:

Complex households

- Large households, which may or may not include members of extended families.
- Tenuous attachment (Tourangeau 1993).
- Roommates.
- Roomers and boarders (Hainer et al. 1988; McKay 1992).
- Students who attend boarding school.
- College students.
- Commuters who may or may not have a second residence to be closer to their place of work Monday–Friday.
- Babies, whom certain respondents tend to exclude from household rosters.
- Households where there children in a shared custody arrangement where the children are not at the sample residence every day, but might be considered as usually living or staying there by one or both parents.
- People temporarily living away from the sample address for a variety of reasons, either in their own second residence or a residence not owned by them.

- Concealment of household members due to respondents' fear of losing their welfare benefits if the government discovers that the additional person or people usually live or stay at the sample address. Respondents who are immigrants, especially those containing household members who are illegally residing in the United States may also fear deportation, arrest, or other serious consequences that they believe are associated with becoming linked to an identifiable address or residence (Hainer et al. 1988; McKay 1992).
- Homelessness, either temporary or permanent.
- New categories, such as "couch surfers" who find residences where they can sleep on a couch, usually for a short time period, and who usually go online and use the Internet (e.g., Web sites such as Craigslist.com), to locate amenable couch-owning residences.

One common approach to addressing these challenges to accurate household rostering has been to use equally complex systems of rules, the goal of which is to determine who should count as a member of the household. However, a major drawback to this approach, especially for researchers hoping to compare data between surveys, is that these rules are not standardized in terms of content or structure. The same lack of consistency can be found across surveys if one examines the definitions provided for important terms and concepts contained within the survey questions. For example, the concept of "usual residence" is ubiquitous in rostering questions and can seem like a relatively simple concept upon initial consideration. However, consider the wide variety of methods that are employed in the process of determining whether an address is the usual residence for a generic individual named Joe:

- Does Joe contribute money for rent, food, bills, or anything else?
- How frequently does Joe sleep here?
- Does Joe use this address to receive mail or phone messages?
- Does Joe usually eat here?
- Do you consider Joe a member of the household?
- Does Joe have another place/places where he usually stays?
- Does Joe perform chores like cleaning?
- Does Joe have a say in household rules?
- Does Joe usually live or stay here?

Compared to the large number of different ways that someone can be considered a member of the household, there is a proportionally small body of research that has examined at whether complex living situations have a significant on response tendencies and on overall data quality. Although it is possible that incorporating complex rostering rules into the design of a survey is one solution to the challenges presented by complex households, there simply has not been enough research conducted in order to draw this conclusion. Many programs of extensive empirical research are sorely needed in order to inform survey designers and researchers' approach to conducting household rostering.

Additional rostering topics that similarly want for further research include a line of experiments aimed at determining best practices for question branching and for identifying ways to reduce the cognitive and time-related burden, for both respondents and interviewers, associated with conducting or responding to interview questions that ask respondents to apply residence rules to generate some form of a roster. Additionally, more research on self-response rostering is also needed so that researchers may gain a clearer understanding about the impact that a rostering approach may have on the survey's data quality (instead of simply making assumptions about what the impact might be). Further, the opportunity to utilize a convergence of scientific evidence on which to base decisions about rostering approaches is absent from the corpus of survey methodology research. Specifically, much of the data that we do have about collecting roster-related data from hard-to-reach cases of highly complex living situations, and populations that were at high risk for Census under coverage, comes from a single survey, the Living Situation Survey, which was last conducted in 1993 (Schwede, L., 1993; Schwede, Blumberg, & Chan, 2005). Revisiting this targeted approach to understanding living situations in the present time period could provide a great deal of direction for designers of large surveys when they are determining how their surveys should approach household rostering.

Areas for future research:

- Identifying the effects of using different rules in different surveys.
  - Even within the same survey, different instructions to interviewers and respondents sometimes contradict one another. What are the effects of this inconsistency?
  - Do different approaches yield different results in terms of accuracy? Response rate? Data quality? What specific impact does rostering approach have
    - On response rates?
    - On data quality?

  - On interviewer and respondent satisfaction with the interview interaction itself?
  - On the respondents' opinion of the agency sponsoring the survey? Has it improved, declined, or stayed the same based on their experience with the interview? Does the direction or magnitude of the change seem to be related to the mode in which the survey was conducted?

- Whether the ways that surveys operationalize usual residence are consistent with respondent notions of usual residence.
- Which basic residence rules make the most sense and are easiest for respondents to understand?
- What is the optimal wording for the rostering questions so that they are easy for the interviewer to consistently read, pronounce, and annunciate?
- The literature lacks a set of strong conclusions based on large-scale, empirical investigations of *de jure* (e.g., a rule-based method for rostering meant to count a person his or her legal residence is, regardless of where they happened to be staying when the survey was conducted) versus *de facto* (e.g., location-based method based on where the person was at the time of the survey) types of residence rules.
- The impact of newly introduced technological tools and methods of analysis available online and on mobile devices for rostering purposes (overlays, pop-ups, etc.) needs to be objectively assessed.

Ultimately, there are many avenues of research that have not been conducted to date related to the concept of household rostering. We need to gather and analyze more empirical data information in order to keep working toward a more accurated and efficient approach to the design, application, and presentation of residence rules might be and whether the same set of rules that makes sense to people also helps generate a more accurate head count.

References and Further Reading

Fein, D. J., & West, K. K. (1988). Towards a theory of coverage error: An exploratory assessment of data from the 1986 Los Angeles Test Census. *Proceedings of the Fourth Annual Research Conference*, Washington, DC: U.S. Census Bureau, pp. 540–562.

Gerber, E. R. (1990). Calculating residence: A cognitive approach to household membership judgments among low income blacks. Unpublished report submitted to the Bureau of the Census.

Hainer, P., Hines, C., Martin, E., & Shapiro, G. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*, Washington, DC: U.S. Census Bureau, pp. 513–539

Martin, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly*, 63, 220–236.

McKay, R. B. (1992). Cultural factors affecting within household coverage and proxy reporting in Hispanic (Salvadoran) households. A pilot study." Paper presented at the 1992 meetings of the American Statistical Association, August, 1992, Boston, MA.

National Research Council (2006). Once, only once, and in the right place: Residence rule in the decennial census. Panel on residence rules in the decennial census. In Daniel L. Cork and Paul R. Voss (Eds.), *Committee on National Statistics, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

Schwede, L. (1993). Household composition and census coverage improvement. Paper presented at the American Anthropological Association Annual Meetings, November 1993.

Schwede, L., Blumberg, R. L., & Chan, A. L. (Eds.) (2005). *Complex Ethnic Households in America*. Lanham: Rowman & Littlefield Publishers, Inc.

Schwede, L., & Ellis, Y. (1994). Exploring associations between subjective and objective assessments of household membership.

Tourangeau, R. (1993). Final report: SIPP roster questions. U.S. Census Bureau Report.

Kathleen Targowski Ashenfelter earned her PhD in Quantitative Psychology, with a focus on Dynamical Systems Modeling, from the University of Notre Dame in 2007. She also holds a master's degree in psycholinguistics. During her tenure as the principal researcher for the Human Factors and Usability Research Group at the U.S. Census Bureau, Dr. Ashenfelter's empirical research focused on identifying potential methods for improving the efficiency and accuracy of conducting residential enumeration, reporting the standard error associated with U.S. Census Bureau statistics, and enriching the user/respondent's experience with U.S. Census Bureau Web sites, data products, and surveys. Kathy is currently a senior data scientist on the Cyber Analytic Services for Unisys Corporation. She applies advanced behavioral algorithms and dynamical systems models to large structured and unstructured datasets (from private industry as well as trusted government sources like the U.S. Census Bureau) for clients in both the private and public sector.

11

Harmonization for Cross-National Comparative Social Survey Research: A Case Study Using the "Private Household" Variable

Jürgen H. P. Hoffmeyer-Zlotnik and Uwe Warner

Introduction

Survey questions about the respondent's attitude, opinion, and behavior are often translatable from one survey language into another without problems. In comparative surveys, these questions measure country -or culture- specific differences in attitudes, opinions, and social behaviors of survey participants. Socio-demographic measures are different. Socio-demographic variables are embedded in the cultural and legal context and depend on the structure of the national states participating in the comparative study (Przeworski and Teune 1970: 42). Simple translation of the question wording does not ensure that researchers obtain equivalent measures across cultures and countries during the survey interviews (Johnson 1998). This weakness can be illustrated measuring the three central variables of the respondent's socio-economic status: education, occupation, and income.

In multi-national surveys, the educational systems are country specific (Hoffmeyer-Zlotnik and Warner 2014: 81 ff.). In some countries, the

J.H.P. Hoffmeyer-Zlotnik (✉)
University of Giessen, Giessen, Germany
e-mail: juergen.hoffmeyer-zlotnik@sowi.uni-giessen.de

U. Warner
Independent Expert, Perl, Germany
e-mail: uwe.warner@orange.lu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_11

upper secondary sector follows a simple structured lower secondary sector. The upper secondary sector can be organized differently across the countries. In some other countries, the lower secondary sector is already structured into different tracks and types of schools leading to various school leaving degrees and have different options to continue education in higher sectors.

Occupational activity depends on the opportunities offered by the national labor market (Hoffmeyer-Zlotnik and Warner 2011, 2014: 106 ff.). National labor market regulations define mandatory qualifications to exercise an occupational activity and are associated with remuneration. The national organizations of work predetermine the work relations and the occupational upward mobility.

The national tax systems, the country-specific systems of social protection and different mandatory or voluntary contributions to the social security have strong influences on the total net household income (Hoffmeyer-Zlotnik and Warner 2014: 137 ff.).

Different countries and cultures apply different definitions of "private households." In some countries, households are defined by common dwelling units, other countries use different forms of organizing the housekeeping, and finally family relations and kinship determine the household membership. These country -and culture- specific definitions have an impact on the comparability of the private household measurements. (Hoffmeyer-Zlotnik and Warner 2008: 19–21, 53–60). A translation of the interview question and the response categories from one language to another does not take into account these legal, social, and political distinctiveness that exists across different countries, or the cultural understanding and formal national organization of social life.

We use the term "*harmonized*" socio-demographic variables for measures allowing the comparison of data *between* two or more cultures or countries. Studies such as the International Social Survey Programme, the World Values Study, and the European Social Survey[1] all require such harmonization. "*Standardized*" socio-demographic variables allow the comparison of two or more data sources *within* one cultural context or country, for example, survey data from interviews with population census from registers of one country. Standardized and harmonized variables allow comparisons within and between countries and enable accurate statistics about this like the European Union Statistics on Income and Living Conditions.[2]

[1] www.issp.org, www.worldvaluessurvey.org, www.europeansocialsurvey.org

[2] http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions

Five Steps of Harmonization

We demonstrate five steps toward harmonized socio-demographic measures using the example of the questions about the "private household" (Hoffmeyer-Zlotnik 2008: 5–24; Hoffmeyer-Zlotnik and Warner 2014: 7–14).

The first step clarifies the *common measurement concept* used in the comparative survey. The main task at this step is addressing the question, what should be measured? In theories about social structure of modern societies, private households monopolize resources (social, economic, and human capital) and minimize social and economic risks. Private households contribute to the production of welfare, they provide services (health care, family support), and create products (dwelling, consumer durables). Households and their members decide how to use scarce disposable time and how to allocate tasks, responsibilities, costs, and expenses. Households distribute resources (time, income, saving, expenditures). Household members share the same socioeconomic characteristics "and often are homogeneous in belief and ideology" (Rossi 1988: 143). The researcher's selection of the measurement concept establishes the variable of "private household" which is to be used in the comparative study.

The second step toward comparative social measurement concepts analyses the *underlying structures* of the variable to be examined. Across modern societies, private households are organized by four main types of living arrangements and their mixtures (Hoffmeyer-Zlotnik and Warner 2008: 19–21):

a) Housekeeping (financial): share common budget, sharing income, sharing expenses, sharing cost of living
b) Housekeeping (organizational): common housekeeping, common living room, sharing food, sharing meals
c) Cohabitation: living together, sharing a dwelling, residing at the same address
d) Family: degree of legal relationship by blood, marriage, adoption, or guardianship

Best practice dictates that a team of experts from the participating countries evaluate the implementation of the selected common concept in the survey. This is guided by the acknowledgment that each survey respondent in each country needs to understand the survey question, that they must search and find information for the answer, that they evaluate the information for the answer, and that they map the final answer onto the provided response format.

The third task looks for the *answer categories* necessary for the researcher's analysis. In the case of the private household variable, it is possible to collect information about the household size as number of adults and number of children in the household. Measures about household composition are also important to collect information about the type of people that constitute a household.

The fourth step starts with the search for pre-existing national survey instruments that measure the focal concept. If researchers decide to use country-specific instruments, they must establish rules for transforming the national outcomes from the survey into equivalent variables for comparison. The result of this *output harmonization* is a comparable measure. If no suitable national measurement instrument is available, questions and answers must be developed specifically for comparison of the concept between countries. The new instrument must be deployable in all participating cultures or countries, must measure in valid and reliable ways, and must be comparable, measuring the same fact everywhere. This is the *input harmonization*.[3]

Our review of the questions about private household from national population censuses shows that nearly all observed countries use their own national definition (Hoffmeyer-Zlotnik and Warner 2008: 19–21; http://unstats.un.org/unsd/demographic/sources/census/censusquest.htm). For comparison of countries, such national measurements cannot be used because they do not measure the same social fact across countries. To establish comparative measures about households in comparative social surveys it is advisable to develop a new survey instrument that has been input-harmonized. If researchers decide to use output harmonization, a fifth step follows. The data collected with the best available survey questionnaire in each country are transposed to a *common classification*. This comparable classification is predetermined by the already given common concept of the measurement.

[3] *Target harmonization* is often used by the official statistics. A target measure is predefined. The national statistical offices collect the necessary information by the best national way to generate the inquired variables and to generate the common indicator.

Using the "Private Household" Concept as an Example for the Five Steps of Harmonization

In Germany, using in-depth interviews, 46 students, 25 researchers, and 118 CATI interviewers were asked about their understanding of "private household." All three of these groups of respondents used eight different elements to describe private household (Hoffmeyer-Zlotnik and Warner 2008: 38–43):

1. The dwelling unit: living under one roof, having an entrance door, and/or a rental agreement
2. Dwelling-shared with common housekeeping, described in terms of "living together with common housekeeping"
3. The family: "being related to each other" and "living together in one house"
4. Some respondents stress affective ties which are also described using the words "being very close"
5. Common activities: (a) common housekeeping, (b) working together with the emphasis of "sharing housework", (c) common living arrangements: eating, sleeping, etc.
6. Financial dependence: common financial budget, sharing of the costs of living, etc.
7. Common planning or live planning, taking care of each other
8. The same address

Coast et al. (2016) examine the census documents of England/Wales and France from 1960 to 2012; and they carried out interviews with data experts on household data production and users of household statistics. Political and institutional country differences effect country-specific interpretation of the international and comparative household concept given by the Statistical Division of the United Nations (2008). They conclude "…the term 'household' may mean different things in different contexts and is not strictly comparable."

Because of these differences between researchers, interviewers, and interviewees in the use of the term "private household", it is necessary to add definitions to clarify the survey question text to all respondents. An experiment across countries shows that the household differs together with the different household definitions applied to for the same group of persons by people in different countries. The consequences are that also household-related social information, such as socio-economic status or the total net household income and the poverty line, varies according to the different understandings and definitions of household membership (Hoffmeyer-Zlotnik and Warner 2008: 53–60).

For comparative social surveys and in accordance with the UN definition of household, we recommend using the housekeeping concept as the basis for the comparative definition. The household is therefore defined to consist of a single person or a group of two and more persons living together, not necessarily within the same housing unit, and taking common provisions for common life. Kinship and family ties are not part of this concept. The common housekeeping emphasizes the organization of everyday life and is not limited to the financial or economic dimensions. This more sociological definition focuses on the reciprocal division of labor and responsibilities among the household members. It is essential to communicate this concept during the interview.

In a vignette study, Gerber et al. (1996) modify the rules for including/excluding household members and asked for the respondents' understanding about the household membership. This study shows the need to present a list of "typical" household persons (a) included in the household and (b) excluded from the household to the respondent because belonging to the household is not self-explanatory. Summing up the included minus the excluded persons gives the household size; aggregating the types of household members produces the household typology (Hoffmeyer-Zlotnik and Warner 2014: 226–227). The final task is to harmonize the language for housekeeping for all cultures and countries under study to ensure that the same concept is measured identically and remains comparable.

Lessons Learned About Harmonization and a Recommendation

The proposed input harmonized instrument for comparative social surveys combines two dimensions of the private household background measure: (a) living together and (b) the common housekeeping as the shared organization of life. Both elements together constitute a unique household concept across cultures and countries measuring the common living arrangements for comparison. In some countries, the concept differs from the common perception of the respondents. This may happen for other socio-demographic background variables too. Therefore, it is important to communicate the applied measurement concept to the interview participants; to the interviewer during their training and to the interviewee during the survey question. The national-specific survey teams provide the guidelines to the fieldwork agencies. They make sure that the country-specific survey instruments obtain comparative measures. Finally, they provide the data users with country- or culture-specific documents. These documents allow the quality controls about the measurement qualities in the participating countries or cultures.

The international or multi-cultural coordinators develop the common measurement concepts. Driven by social science theories, they define the comparable concepts. They supervise the harmonization steps and instruct the national survey teams about the intended measurement and comparability. They report their decision to the scientific community so that the data users with various national or cultural backgrounds do not misinterpret the measures of socio-demographic explanatory variables during the comparative research.

We recommend improving the transparency of the harmonization process. Together with the data sets, the actors implementing surveys across countries or cultures publish reports about the creation of the data collection, the potential comparability of the socio-demographic background measures, and the quality of the explanatory variables.

References and Further Reading

Coast, E., Fanghanel, A., Lelièvre, E., & Randall, S. (2016). Counting the Population or Describing Society? A Comparison of England & Wales and French Censuses. In *European Journal of Population*.

European Social Survey. (2002). Project Instructions (PAPI). Round 1, 2002. ESS Document Date: July 15, 2002. http://www.europeansocialsurvey.org/docs/round1/fieldwork/source/ESS1_source_project_instructions.pdf [accessed March 25, 2016].

Gerber, E. R., Wellens, T. R., & Keeley, C. (1996). "Who Lives Here?" The Use of Vignettes in Household Roster Research. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp. 962–967. http://www.census.gov/srd/papers/pdf/erg9601.pdf. [accessed March 25, 2016].

Hoffmeyer-Zlotnik, J. H. P. (2008). Harmonisation of Demographic and Socio-Economic Variables in Cross-National Survey Research; Bulletin de Methodologie Sociologique N.98, April 2008, pp. 5–24.

Hoffmeyer-Zlotnik, J. H. P., & Warner, U. (2008). *Private Household Concepts and Their Operationalisation in National and International Social Surveys*. Mannheim: GESIS-ZUMA. http://www.gesis.org/uploads/media/SM1_Gesamt.pdf [accessed March 25, 2016].

Hoffmeyer-Zlotnik, J. H. P., & Warner, U. (2011). *Measuring Occupation and Labour Status in Cross-National Comparative Surveys*. Mannheim: GESIS Series Volume 7.

Hoffmeyer-Zlotnik, J. H. P., & Warner, U. (2014). *Harmonising Demographic and Socio-Economic Variables for Cross-National Comparative Survey Research*. Dordrecht: Springer Science + Business Media.

Johnson, T. P. (1998). Approaches to Equivalence in Cross Cultural and Cross-National Survey Research. In ZUMA-Nachrichten Spezial 3. Mannheim: ZUMA. pp. 1–40.

Przeworski, A., & Teune, H. (1970). *The Logic of Comparative Social Inquiry*. New York: John Wiley.

Rossi, P. H. (1988). On Sociological Data. In N. J. Smelser (ed.), *Handbook of Sociology*. Newbury Park et al: Sage Publications, Inc. pp. 131–154.

United Nations (2008). Principles and Recommendations for Population and Housing Censuses Revision 2. Department of Economic and Social Affairs Statistics Division, Statistical papers Series M No. 67/Rev.2. New York.

Prof. Dr. Jürgen H.P. Hoffmeyer-Zlotnik is Associate Professor at the Institute for Political Science of the University of Giessen, Germany. The main focus of his research is the standardization and the harmonization of demographic and socio-economic variables in national and cross-national comparison.

Dr. Uwe Warner was Senior Researcher at CEPS/INSTEAD, Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques in Esch sur Alzette, Luxembourg. His main research topic is on cross-national comparative survey research. In metodološki zvezki – Advances in Methodology and Statistics they published articles on "income" (2006 and 2015), "education" (2007), "private household" (2009), and on "ethnicity" (2010) as socio-demographic variables in cross-national surveys.

12

Answering for Someone Else: Proxy Reports in Survey Research

Curtiss Cobb

Not all answers to survey questions are provided by the sampled respondent, often the person selected to be the respondent in a survey is unavailable when the interviewer is at the home or on the phone. In these cases some surveys will allow another person to respond on behalf of the target this person is a proxy for the target. This proxy may be another member of the household such as spouse or child, or a friend or co-worker. Proxy reports, then, are the answers to survey questions about the respondent that are provided by someone other than the target respondent.

The practical appeal of using proxy reporting in survey research seems obvious: using proxy respondents can make obtaining information faster and less expensive. Proxies can increase contact and cooperation rates when the targets themselves would be difficult to contact or are reluctant to be interviewed. Many surveys collect proxy reports for topics like political participation, immigration status, social stratification, employment status or changes, and health and illness. Thousands of research articles in the social sciences have been written based on data that include proxy reports, often without the express knowledge or acknowledgment of the authors. Important surveys that collect data from proxies include the U.S. Census, the Current Population Survey (CPS), General

C. Cobb (✉)
Menlo Park, CA, USA
e-mail: curtisscobb@gmail.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_12

Social Survey (GSS), and the National Longitudinal Study of Adolescent to Adult Health (Add Health).

The use of proxy reports has been estimated to save up 17 percent of survey costs for the CPS. The monetary savings come at a variable cost to data quality – laboratory tests of CPS measures indicate agreement between target reports and proxy reports as low as 67 percent for questions related to the number of hours worked in the previous week by the target and as high as 92 percent for questions related to union membership (Boehm 1989). Similar results have been found for items on voting turnout, household expenditures, labor participation and other important variables (e.g., Borjas, Freeman & Katz 1996; Highton 2005; Mathiowetz 2010; Oliver 1996; Rubenstein et al. 1984; Weeden 2002). Thus, future research may need to focus on the types of questions that are accurately provided by proxies and the types of questions that are not.

The potential risk of proxy reports seems obvious as well: survey respondents may be less accurate when describing other people than when describing themselves. Moreover, proxy motivation may be different from target motivation, leading to differences in effort and data quality. There may also be important differences in perspective when observing the same phenomenon. Proxies may observe or process information streams differently than targets. Thus, the proxy report would constitute the proxy's perception of the target's attribute, whereas the self-report would constitute a self-perception, both valid but different.

On the other hand, it is also possible that proxy reports are sometimes more accurate than self-reports, such as when the proxy has access to more information than the target or in situations where social desirability effects may impact self-disclosure. The accuracy of proxy reports may vary depending upon the construct being measured, the relationship of the proxy to the target, and access that the proxy and target have to information with which to answer the question. It important to expand our understanding of how accurate proxy reports truly are, what might cause inaccuracy in proxy reports, and the conditions under which such inaccuracies are most likely to occur.

A large body of research spanning more than fifty years has sought to compare the accuracy of proxy reports and self-reports in surveys. A meta-analysis of 93 studies on the topic found that the design of many studies makes it difficult to draw generalizable conclusions with any confidence because they lack the basic features required to assess accuracy (Cobb, Krosnick & Pearson, forthcoming). Following is a list of the necessary features a study needs to understand the accuracy of proxy reports, along with how many out of the 93 studies reviewed include that feature in parentheses:

- Both targets and proxies should be interviewed (17)
- Targets and proxies should constitute representative samples of the same population (25)
- Questions asked of targets and proxies should be identical (21)
- Independent, external measure of the attribute being assessed should be used to measure accuracy (76)

With these features, there are two research strategies that could be employed. In the first, sampled individuals are randomly assigned to provide answers for themselves or a proxy. Then the results are aggregated and compared, because of random assignment the results should be very similar, and these sets of results can then be compared against the external benchmark validation values. The second approach is to obtain measurements from matched targets and proxies, this approach allows researchers to evaluate the role of non-response and assess the association between reporting errors made by targets and reporting errors made by proxies.

In the meta-analysis of 93 studies, Cobb, Krosnick and Pearson (forthcoming) found that only six studies had the necessary features to evaluate proxy accuracy, specifically. It is important to note that accuracy is defined as a measure of validity and is different from level of agreement between the target and proxy. This finding indicates that much of the existing literature is inadequate for its intended purpose and considerably more research is warranted given the relatively large number of proxy studies that already exist.

Of the six studies with the necessary features to assess proxy accuracy, all involved reports on medical events. In four of the six studies, proxies were similarly equally accurate at reporting health information about targets as target respondents were reporting about themselves (Cobb et al. 1956; Thompson & Tauber 1957; Andersen et al 1973; Balamuth 1965). One study found proxies to be less accurate than self-reports when reporting on daily activities (Magaziner et al. 1997), while another study found proxies as more accurate at reporting about doctor visits when the targets were minor children (Cannell & Fowler 1963). These are promising findings for the validity of using proxy reports; at least in health-related studies proxies can be relied on to be accurate at reporting information as targets. These findings need further research, particularly since five of the six studies are over 40 years old.

Despite not being adequately designed to assess the accuracy of proxy reports, many of the other studies in the meta-analysis were still informative in other ways. A number of the relevant and important findings are summarized as follows:

- Cognitive studies reveal that memories about others are less elaborate, less experientially based, and less concerned with self-presentation (Moore 1988).
- Proxies anchor answers based on their own behaviors and attitudes (Sudman, Bickart, Blair & Menon 1994).
- Proxies estimate more versus recall (Bickart et al. 1990; Schwarz & Wellens 1997).
- Time together increases agreement and the similarity of cognitive strategies used to arrive at responses (Amato & Ochiltree 1987; Bahrick, Bahrick & Wittlinger 1975; Cohen & Orum 1972; Lien, Friestad & Klepp 2001).
- Proxies are more likely to under-report behavior and events, except those that involve care-taking activities (McPherson & Addington-Hall 2003; Hrisos et al. 2009; Miller, Massagli & Clarridge 1986).
- Knowledge of and exposure to the question topic increases agreement (Magaziner et al. 1988; Grootendorst, Feeny & Furlong 1997).
- Stable traits and characteristics lead to more target/proxy agreement than changing activities
- Observable information is easier for proxies to report on than unobservable information

While these findings suggest a number of best practices for using proxy reports in surveys, there are still a number of important areas for future research. First, there is a need for more correctly designed studies on the accuracy of proxy reports across a variety of topical domains. Second, more research is needed on identifying question designs that increase proxy accuracy, this involves identifying appropriate reference periods, question formats, and features that may increase proxy motivation. Third, future research needs to explore how the characteristics of proxies impact the accuracy of their reports, for example, are household members more or less accurate than non-household members as proxies? Lastly, researchers need to identify optimal strategies for implementing best practices when designing new questions that may be answered by either targets or proxies.

Areas for future research:

- Investigating the implications of increasing levels of proxy reporting
- Identifying questions that are and are not appropriate for proxy reports to be collected on

  - How accurate are proxy reports on different types of questions?
  - What causes proxy inaccuracy?
  - Under what conditions are inaccuracies most likely to occur?

- How can questions be optimally designed to maximize accuracy for both targets and proxies?
- What are the impacts of question design features on proxy report accuracy?
- What impact do proxy characteristics have on the accuracy of their reports?

References and Further Reading

Amato, P. R., & Ochiltree, G. (1987). Interviewing children about their families: A note on data quality. *Journal of Marriage and the Family*, 49(3), 669–675.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of experimental psychology: General*, 104(1), 54.

Balamuth, E. (1965). Health interview responses compared with medical records.

Bickart, B. A., Blair, J., Menon, G., & Sudman, S. (1990). Cognitive Aspects of Proxy Reporting of Behvior. NA-Advances in Consumer Research Volume 17.

Boehm, L. M. (1989). Reliability of proxy response in the current population survey. In Proceedings of the Survey Research Methods Section, American Statistical Association.

Borjas, G. J., Freeman, R. B., & Katz, L. F. (1996). Searching for the Effect of Immigration on the Labor Market (No. w5454). National Bureau of Economic Research.

Cannell, C. F., & Fowler, F. J. (1963). Comparison of a self-enumerative procedure and a personal interview: A validity study. *Public Opinion Quarterly*, 27(2), 250–264.

Cobb, C., Krosnick, J. A., & Pearson, J. (forthcoming). The accuracy of self-reports and proxy reports in surveys.

Cobb, S., Thompson, D. J., Rosenbaum, J., Warren, J. E., & Merchant, W. R. (1956). On the measurement of prevalence of arthritis and rheumatism from interview data. *Journal of Chronic Diseases*, 3(2), 134–139.

Cohen, R. S., & Orum, A. M. (1972). Parent-child consensus on socioeconomic data obtained from sample surveys. *The Public Opinion Quarterly*, 36(1), 95–98.

Grootendorst, P. V., Feeny, D. H., & Furlong, W. (1997). Does it matter whom and how you ask? Inter-and intra-rater agreement in the Ontario Health Survey. *Journal of Clinical Epidemiology*, 50(2), 127–135.

Highton, B. (2005). Self-reported versus proxy-reported voter turnout in the current population survey. *Public Opinion Quarterly*, 69(1), 113–123.

Hrisos, S., Eccles, M. P., Francis, J. J., Dickinson, H. O., Kaner, E. F., Beyer, F., & Johnston, M. (2009). Are there valid proxy measures of clinical behaviour? A systematic review. *Implementation Science*, 4(1), 37.

Lien, N., Friestad, C., & Klepp, K. I. (2001). Adolescents' proxy reports of parents' socioeconomic status: How valid are they? *Journal of Epidemiology and Community Health*, 55(10), 731–737.

Magaziner, J., Simonsick, E. M., Kashner, T. M., & Hebel, J. R. (1988). Patient-proxy response comparability on measures of patient health and functional status. *Journal of Clinical Epidemiology*, 41(11), 1065–1074.

Magaziner, J., Zimmerman, S. I., Gruber-Baldini, A. L., Hebel, J. R., & Fox, K. M. (1997). Proxy reporting in five areas of functional status comparison with self-reports and observations of performance. *American Journal of Epidemiology*, 146(5), 418–428.

Mathiowetz, N. (2010, December). Self and proxy reporting in the consumer expenditure survey program. In Bureau of Labor Statistics Consumer Expenditure Survey Methods Workshop (Vol. 8).

McPherson, C. J., & Addington-Hall, J. M. (2003). Judging the quality of care at the end of life: can proxies provide reliable information? *Social Science & Medicine*, 56(1), 95–109.

Miller, R. E., Massagli, M. P., & Clarridge, B. R. (1986). Quality of proxy vs. self reports: evidence from a health survey with repeated measures. In American Statistical Association: Proceedings of the Section on Survey Research Methods (pp. 546–51).

Moore, J. C. (1988). Miscellanea, self/proxy response status and survey response quality, a review of the literature. *Journal of Official Statistics*, 4(2), 155.

Oliver, J. E. (1996). The effects of eligibility restrictions and party activity on absentee voting and overall turnout. *American Journal of Political Science*, 498–513.

Rubenstein, L. Z., Schairer, C., Wieland, G. D., & Kane, R. (1984). Systematic biases in functional status assessment of elderly adults: Effects of different data sources. *Journal of Gerontology*, 39(6), 686–691.

Schwarz, N., & Wellens, T. (1997). Cognitive dynamics of proxy responding: The diverging perspectives of actors and observers. *Journal of Official Statistics*, 13(2), 159.

Sudman, S., Bickart, B., Blair, J., & Menon, G. (1994). The effect of participation level on reports of behavior and attitudes by proxy reporters. In *Autobiographical memory and the validity of retrospective reports* (pp. 251–265). Springer: New York.

Thompson, D. J., & Tauber, J. (1957). Household survey, individual interview, and clinical examination to determine prevalance of heart disease. *American Journal of Public Health and the Nations Health*, 47(9), 1131–1140.

Weeden, K. A. (2002). Why do some occupations pay more than others? Social Closure and Earnings Inequality in the United States 1. *American Journal of Sociology*, 108(1), 55–101.

Curtiss Cobb leads the Population and Survey Sciences Team at Facebook. His research focuses on cross-cultural survey methods, web surveys, technology adoption patterns, and evolving attitudinal trends related to people's online "presence." Prior to Facebook, Curtiss was Senior Director of Survey Methodology at GfK and consulted on survey studies for clients such as the Associated Press, Pew Research Center, CDC, U.S. State Department and numerous academic studies. Curtiss received his BA from the University of Southern California and has an MA in Quantitative Methods for Social Sciences from Columbia University. He holds an MA and PhD in Sociology from Stanford University.

13

Improving Question Design to Maximize Reliability and Validity

Jon A. Krosnick

There are three primary goals that researchers should keep in mind when evaluating questions and trying to identify the best ways to ask questions. The first is to minimize administration difficulty. That is, use questions that can be asked and answered as quickly as possible. Second, survey designers would like respondents to make as few completion errors possible. So if a respondent is asked to pick a point on a rating scale on a paper questionnaire survey designers don't want them circling two or three points saying, "I am somewhere in this range, but I don't know where." Lastly, all other things equal, researchers would like respondents to enjoy answering the question and not be very frustrated by it. But, all else is not equal. However, at the end of the day, researchers should be willing to use longer questionnaires and have respondents be frustrated if that's what it takes to maximize the reliability and validity of the measurements. Fortunately, the literature suggests that what goes quickly and easily for respondents also produces the most accurate data.

An important perspective on questionnaire design approaches the issue with a goal of understanding the cognitive steps in question answering. Many books have been written on this topic but in general the process is fairly well described from a theoretical perspective. When a respondent is asked a

J.A. Krosnick (✉)
Department of Communication, Stanford University, CA, USA
e-mail: krosnick@stanford.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_13

question they need to engage in four steps to provide the response that the researcher is seeking:

1. Understand the intent of the question – that is, what is meant by the question as it may differ from the literal interpretation of the words.
2. Search their memory for relevant information
3. Integrate the relevant information into a summary judgment
4. Translate the judgment into the required format for the response alternatives

When respondents engage each of these steps before providing a response it is known as "optimizing" the survey response. Unfortunately, a large and influential body of work suggests that people often do not perform all four steps before providing their response, instead they satisfice, they settle for shortcuts (Krosnick, 1991, 1999; Vannette & Krosnick, 2014). There are two ways that this may happen, one is to superficially engage the two middle stages of searching and integrating information rather than doing so effortfully, this is what researchers describe as "weak satisficing." Alternatively, if the respondent has entirely given up on providing good responses they will skip the middle two steps entirely and simply understand the question and then provide a response, this is called "strong satisficing." In this case respondents may look to the question and situation for cues pointing to apparently plausible answers that would be easy to justify without thinking. There are a number of satisficing strategies that may be employed by respondents including

- Selecting the first reasonable response
- Agreeing with assertions
- Non-differentiation in ratings
- Saying "don't know"
- Mental coin-flipping

Three primary causes of satisficing have been implicated by the existing research and they include (1) respondent ability, (2) respondent motivation, and (3) task difficulty. The existing research indicates that to the extent that researchers can make the survey task motivating and simple respondents will be less likely to satisfice.

Another important perspective on questionnaire design considers the conversational norms and conventions that survey interviews share or violate with regard to normal conversations. Survey questionnaires may be considered scripts for somewhat unique conversations and yet respondents often do not realize that the rules of this conversation are different than the rules of normal conversation, particularly if there is an interviewer involved. Respondents often assume that the same rules apply and yet, if questions violate those rules, respondents can be misled or confused because they misinterpret the context. This becomes an important concern because survey questions routinely violate the rules of everyday conversation. For example, in a normal conversation, if a person asked you, "How are you doing today?" You said, "Good." It would be a violation of conversational norms for the other person to ask, "How are you doing today?" after that. And yet surveys routinely use multiple questions to measure the same construct, particularly with long batteries of questions. Given the frustration that this situation can cause for respondents it is important that researchers not needlessly subject them to this kind of treatment. Grice (1975) provides a series of maxims for how to adhere to conversational norms and survey designers ignore these at their own peril.

Open questions have a number of distinguishing features:

- Only questions are standardized
- No response alternatives are suggested
- Verbatim transcription is required
- Interviewers may probe the respondent to say more about a topic
- Interviewers must be well-educated
- Interviewers must be trained extensively
- Results are more subjective due to elaborate coding schemes that must be developed to classify responses
- Analysis is expensive and time-consuming
- Answers may be provided freely and without bias

Similarly, closed questions have a number of distinguishing features:

- Both questions and answers are standardized
- Respondents code their own answers
- Interviewer training is simple
- Administration is fast and cheap
- Data are easy to analyze
- Results are objective: no bias from questioner (in probing) or coder

Fortunately, decades of research have indicated a number of best practices for how to evaluate the costs and benefits of open and closed questions on surveys. First, in studies of reliability, open questions prove to be more reliable than closed questions and in lots of different studies of validity, open questions prove to be superior to close questions across the board using these various different methods of assessing validity.

A second concern was that open questions might be particularly susceptible to salience effects. For example, if a survey asks "What is the most important problem facing the country?" and the respondent happened to have seen a news story about crime on television the previous night maybe that enhances the likelihood that the respondent would retrieve crime as a potential problem to answer with whereas, with a closed question on a list, that salience effect may be minimized. Empirical evidence seems to indicate no support for this notion. In fact, salience appears to affect open and closed questions equally.

Lastly, there has been concern about frame of reference effects. For example, if a survey asks "What is the most important problem facing the country?" the respondent needs to understand what counts as a problem and what counts as an acceptable answer? That can be ambiguous with open questions whereas with closed questions if you offer a set of choices, what is an acceptable choice is made explicit. Indeed, there is evidence that in some cases open-ended questions are ambiguous enough that the frame of reference is not established, but that is not necessarily an inherent problem with open questions, but it an inherent problem with some open questions that might better be solved in other ways.

With regard to closed questions, it turns out that a series of concerns have been articulated, all of which do have empirical support. One concern is non-attitudes, that by offering people options, people select choices without actually having any substantive attitudes or behavior behind them. Secondly, with regard to the numeric response options, if a survey offer ranges of 0–5 hours, 6–10 hours, and so on, the way I choose those ranges sends a signal to a respondent about what an acceptable and normal answer would be. The answers in the middle of the range are what people assume to be what a normal person would pick so there is gravitation toward the middle. As a result, it may be better not to offer them. Lastly, there is the notion that if a survey asks "What is the most important problem facing the country?" "Is it the federal budget deficit, crime, inflation, unemployment or something else." The "something else" category will cover the range of responses that respondents would otherwise provide unaided. However, this assumption turns out to be a serious problem. In fact, research dating back as far as 70 years or more has consistently demonstrated that offering a "other" or "something else" option does almost nothing. People almost never select it and they think what the survey is asking “which of the following is the most important problem facing the country?” If the respondent insists on picking something else they can, but the survey seems to prefer that they don’t, rendering this approach ineffective. In conclusion, this literature suggests that open-ended questions should be used more frequently. This is because survey designers can’t be sure of the universe of possible answers to a categorical question and the “other” response option does not work.

Similarly, a large amount of research has been conducted with regard to the ideal number of points to include on a rating scale. There is considerable variation in what is done in practice, including even within a single survey. For example, the American National Elections Studies use everything from a two point scale, “do you approve or disapprove of the president’s job performance” up to 101 point scales.

Theoretically, there are a variety of principles that can guide the decision regarding the best length of a rating scale. In order to understand as much as possible about respondents and to make the process of mapping their feelings on to a ratings scale easier, maybe more points is better. However, if too many points are offered on a rating scale, respondents might get confused. What is difference between 75 and 79, for example? So while it is theoretically possible that increasing the length of response scales may improve the precision of ratings, this is not the case. In fact, research indicates that long scales become ambiguous, and thus researchers end up seeking more refined opinions than people actually have to offer and reducing data quality. The empirical evidence is quite clear that completion errors increase with longer scales. On long response scales, respondents are more likely to perceive their response as falling within some relatively accurate range but the reliability of these responses on the particular scale point is very low, indicating that the response scale is too long.

Best practices for scale design indicate that, to maximize discrimination without sacrificing reliability, bipolar survey questions should use seven-point scales and unipolar questions should use five-point scales. Branching bipolar dimensions can also be helpful, for example, the ANES asks “Generally speaking, do you consider yourself to be a Republican, a Democrat, an Independent, or what?” after which Democrats and Republicans are asked if they’re strong or not very strong partisans and Independents are asked if they lean toward the Democratic or Republican party. Using this approach a seven-point scale of Strong Democrat to Strong Republican can be formed in a more valid and reliable manner than simply asking respondents to place themselves on the seven-point scale.

With regard to verbal labels of scale points, some surveys present scales with numbers on all points and words only on the ends, others put words and numbers on all of the points, and others still get rid of the numbers and just have words on each of the points. As survey designers consider selecting those labels, a series of goals are worth pursuing. One is that respondents should find it easy to interpret the meaning of all the scale points. After they're done interpreting them, the meanings of each scale point should be clear. Third, all respondents should interpret the meanings of the scale points identically. That is, we don't want different people interpreting the scale points differently from each other. Fourth, the scale point labels should differentiate respondents as much and as validly as possible. And lastly, the resulting scale should include points that correspond to all points on the underlying continuum.

Previous research on labeling scale points indicates that numbers alone seem intentionally ambiguous and longer scales seem potentially more ambiguous. There has been concern in the literature that labeling only the end points might attract people to those end points if the labels clarify the meanings of those points more so than other points. But, if survey designers pick vague labels they might cause problems and if you pick labels that are overly specific, respondents may be unable to find the place on the scale where they belong. So some optimal degree of vagueness might be desirable. However, in terms of evaluating the quality of data, the literature is quite clear: respondents prefer scales with more verbal labels, reliability is higher for scales with more verbal labels, and validity is higher in various ways for scales with more verbal labels.

Question wording is another important area of questionnaire design that has generated considerable prior research and a number of best practices can be extracted from the literature. In general the conventional wisdom regarding question wording is

- Simple, direct, comprehensible
- No jargon
- Be specific
- Avoid ambiguous words
- Avoid double-barreled questions
- Avoid negations
- Avoid leading questions
- Include filter questions
- Be sure questions read smoothly aloud
- Avoid emotionally charged words

- Avoid prestige names
- Allow for all possible responses

However, there are a number of important challenges regarding question wording that there is not nearly enough empirical guidance on to generate best practices. For example, slight changes in question wordings that researchers believe to represent the same or similar underlying constructs such as “support” versus “favor” sometimes generate surprisingly different results. While survey designers often try to follow a number of common sense guidelines for question wording, more research is needed to confirm whether or not these approaches are providing the anticipated results or if there is an expanded set of best practices that can be empirically confirmed. Determining the optimal language to use in survey questions is an important area for future.

Areas for future research:

- Getting the most out of open questions
- Optimal language for question wording
- Dissemination of empirically confirmed best practices to the broad array of survey users across disciplines

References and Further Reading

Grice, H.P. (1975). “Logic and Conversation,” Syntax and Semantics, vol.3 edited by P. Cole and J. Morgan, Academic Press. Reprinted as ch.2 of Grice 1989, 22–40.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567.

Vannette, D. L., & Krosnick, J. A. (2014). Answering Questions. In A. Ie, C. T. Ngnoumen, & E. J. Langer (Eds.), *The Wiley Blackwell Handbook of Mindfulness* (First Edition, Vol. 1, pp. 312–327). John Wiley & Sons.

Jon A. Krosnick is the Frederic O. Glover Professor in Humanities and Social Sciences at Stanford University, Stanford, CA, USA, and a University Fellow at Resources for the Future. This work was supported by National Science Foundation Award [1256359].

14

Cognitive Interviewing in Survey Design: State of the Science and Future Directions

Gordon Willis

In the context of total survey error, response error— as a form of measurement error— is a type that researchers can control through questionnaire design. The assertion is that because this is a serious, yet controllable type of error, it is worthy of attention and continued research, due to the fact that small changes in question wording and questionnaire design and format can make a substantial difference in the answers that respondents provide to questions. For example, simply asking respondents how much time they spend on a common daily activity often results in over-reports of that activity when compared to a questionnaire that first asks if the respondents engage at all in the activity, and then following up to request the amount of time only respondents who report the behavior.

Cognitive testing is an applied approach to identifying problems in survey questionnaires and related materials, with the goal of reducing the associated response errors. Typically, a preliminary version of the questionnaire is developed, members of the targeted population are recruited and paid for their time, and then one-on-one interviews are conducted, usually in a face-to-face context. The cognitive interview is conducted using verbal probing techniques, as well as "think-aloud," to elicit thinking about each question. These probes take a number of forms such as

G. Willis (✉)
National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
e-mail: willisg@mail.nih.gov

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_14

- Comprehension probe: “What does the term ‘dental sealant’ mean to you?”
- Paraphrasing: “Can you repeat the question in your own words?”
- Confidence judgment: “How sure are you that your health insurance covers…”
- Recall probe: “How do you know that you went to the doctor 3 times…?”
- Specific probe: “Why do you think that breast cancer is the most serious health problem?”
- “Back-pocket” probe: “Can you tell me more about that?”

The goal of using these probes is to note apparent problems related to question wording, ordering, and format, and then to suggest modifications that address the problems. Best practices suggest doing this as an iterative process consisting of multiple testing rounds.

There are a number of questions that have been raised about the use of cognitive interviewing, some of which have been addressed by research, and others that still require future research to be conducted. For example, despite the widespread use of cognitive pretesting to the evaluate questionnaires – particularly in government survey labs – it is unclear whether or not independent researchers testing the same questionnaire would reach the same conclusions. Preliminary research has provided promising results about the reliability of the cognitive pretesting findings, but existing research has been limited and incomplete. This indicates a promising area for future research on the cognitive pretesting method. A key question needing to be addressed is: Under what conditions are cognitive interviewing results stable and reliable, and what can researchers do to enhance those conditions?

Additional research is also needed on best practices for designing cognitive pretesting studies themselves. For example, because cognitive interviewing is a qualitative research endeavor, it is often unclear what sample sizes are necessary. Identifying effective practices with regard to cognitive interview sample size is important for two reasons. First, it is necessary to know how many interviews will be enough to identify a problem, and then how many more will be necessary to assess the seriousness or impact of the problem. One existing study has examined this issue and found that additional interviews continued to produce observations of new problems, although the rate of new problems per interview decreased (Blair and Conrad 2011; POQ, p. 654). This finding needs further study and replication. Developing future research on these issues is important so that researchers can make the most of the resources invested in cognitive interviewing.

It is also important for future research to focus on identifying the utility of cognitive interviewing for mixed-mode surveys and novel administration methods, as survey research moves into the future. Much of cognitive interviewing research to date has focused on differences between administration modes, because the cognitive issues that appear in self-administered modes are somewhat different from those that are interviewer-administered. Increasingly, the focus has shifted to identifying cognitive issues surrounding web usability and Internet administered surveys, but this is very new and requires significant future research. Other new areas of research have looked at pushing cognitive interviewing itself to different modes such as Skype, or other Internet-based approaches to soliciting feedback from participants. For example, research has been done on administering probes to web-survey respondents after each evaluated question, using an open text box for them to provide their responses. This practice enables many more cognitive interviews to be performed for the same cost, but it is unclear what is lost in terms of information that an interviewer otherwise may have been able to obtain. There is also the option of conducting some traditional in-person interviews in tandem with Internet-based approaches, to try to maximize the value of both. However, essentially no research has examined this, and it is necessary to adapt cognitive interviewing to the future of survey research.

Finally, more research is needed on applications of cognitive interviewing techniques for addressing issues surrounding cross-cultural comparability within and between surveys. Although cross-cultural differences have been widely recognized by survey researchers, with careful steps taken in sampling, language of administration, and weighting, relatively little has been done with cognitive interviewing to test the differences in cognitive problems that different cultural groups may have with a questionnaire. Further, researchers have not established whether current cognitive interviewing techniques are applicable across cultures, so additional research is needed in this area. Once appropriate cognitive interviewing techniques are identified, they can be applied to ensure that surveys exhibit cross-cultural measurement comparability. A related issue arises from linguistic and translational issues in cross-cultural surveys, which cognitive interviewing should theoretically be able identify. Even basic translations can go badly if good evaluation and pretesting practices are ignored. In short, cognitive interviewing hold great promise for increasing the ecological validity of survey research in increasingly diverse research contexts, but considerable research is needed to maximize the value of the method.

Areas for future research:

- Under what conditions are cognitive interviewing results stable and reliable?
  - What steps can researchers take to enhance those conditions?
- How many cognitive interviews are necessary to
  - a) Identify a problem (number of interviews before problem X occurs)
  - b) Validate a problem (of X interviews, problem occurs in at least Y cases)
- Identifying the utility of cognitive interviewing for mixed-mode surveys
- Identifying and testing novel administration methods for cognitive interviews
- Identifying the applicability of cognitive interviewing methods across cultures
- Identifying best practices for using cognitive interviewing to increase cross-cultural comparability

References

Blair, J., & Conrad, F. G. (2011). Sample Size for Cognitive Interview Pretesting. *Public Opinion Quarterly*, *75*(4), 636–658. http://doi.org/10.1093/poq/nfr035

Futher Readings

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, http://doi.org/10.2307/4500375?ref=search-gateway:87fd162ce18ea7452c5ae91bc5a46d55

Campanelli, P. (1997). Testing survey questions: New directions in cognitive interviewing. *Bulletin De Méthodologie Sociologique*, (55), 5–17. http://doi.org/10.2307/24359674?ref=search-gateway:87fd162ce18ea7452c5ae91bc5a46d55

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, 12(3), 229–238. http://doi.org/10.1023/A:1023254226592

Collins, D. (2014). *Cognitive Interviewing Practice*. London: SAGE.

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly*. http://doi.org/10.1093/poq/nfp013

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology*, 24, 73–104. http://doi.org/10.2307/270979?ref=search-gateway:f342e4e1e263663ce4fc9be75f881cec

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Gordon Willis is Cognitive Psychologist at the National Cancer Institute, National Institutes of Health. He has previously worked at Research Triangle Institute and at the National Center for Health Statistics, Centers for Disease Control and Prevention, to develop methods for developing and evaluating survey questions. Dr. Willis attended Oberlin College and Northwestern University. He has co-authored the Questionnaire Appraisal System for designing survey items, and the Cognitive Interview Reporting Format for organizing study results; and has written two books: Cognitive Interviewing: A Tool for Improving Survey Questions; and Analysis of the Cognitive Interview in Questionnaire Design. Dr. Willis also teaches questionnaire design and pretesting for the Joint Program in Survey Methodology, and at the Odum Institute, University of North Carolina. His work involves the development of surveys on health topics such as cancer risk factors, and focuses on cross-cultural issues in questionnaire design and pretesting.

15

Survey Interviewing: Departures from the Script

Nora Cate Schaeffer

Standardized survey interviews are typically scripted so that the interviewer follows a predefined path through the instrument. The interviewer is trained to read questions exactly as they are worded and to avoid deviating from the questions as written. Interviewers mediate between the organization or researcher and the respondent.

Looking at the changes in major research studies that have occurred in the last decade, it is possible to guess the following about future studies: In addition to the sorts of opinion studies or other studies currently being conducted, there will be a class of research studies that will be very complex and demanding for both respondents and interviewers. As the cost of reaching sample members increases, from the researcher's point of view it is economically sensible to ask face-to-face interviewers to do many complex tasks, once the interviewer has persuaded the sample member to become a respondent. For these complex interviews to be successful, we need to understand more about how measurement is accomplished within interaction and how to motivate respondents.

There are many factors that influence the behavior of interviewers. One model of interviewer behavior is an interactional model of the process of recruiting survey respondents, which focuses on the actions of the parties and the sequence of actions during a part of the interaction that is not scripted

N.C. Schaeffer (✉)
University of Wisconsin-Madison, Madison, USA
e-mail: schaeffe@ssc.wisc.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_15

(Schaeffer et al. 2013). A preliminary model of the interaction between the interviewer and respondent during the interview itself is presented in Schaeffer and Dykema (2011b).

The model of interaction during the interview proposes that the main influences on the behavior of interviewers – and their deviations from the script – are

- training in the interviewing practices of standardization;
- technology, which has both a direct effect on the interviewer's behavior and an indirect effect through the way technology shapes and limits the characteristics of the survey question;
- the characteristics of the survey question (see Schaeffer and Dykema 2011a), which affects the interviewer's behavior directly as she reads the question and indirectly by the way the question affects the cognitive processing of the respondent and the respondent's subsequent behavior;
- the behavior of the respondent, which may require that the interviewer respond in ways that challenge her compliance with standardization;
- interactional and conversational practices, some of which may be made more or less relevant by characteristics of the question or the behavior of the respondent.

Technology – whether paper or some electronic technology – presents the script to the interviewer in a way that is often incomplete, so that the interviewer must improvise using the principles of standardization. For example, a paper grid may allow the interviewer to have an overview of the structure of the task and also allow her to enter information that the respondent provides before the interviewer requests it (e.g., the ages of other members of the household). A Computer-Assisted Personal Interviewing (CAPI) instrument, on the other hand, may require that each piece of information be entered on a separate screen, and these constraints may in turn motivate the interviewer to reinforce that standardized order with the respondent.

Interactional and conversational practices are sometimes in tension with standardization. Some of these tensions may have only minor consequences for data quality. For example, interviewers routinely use "okay" both as a receipt for an answer and to announce an upcoming return to the script of the next question. Other tensions may be more consequential. For example, when cognitive assessments or knowledge questions are administered to respondents, the respondent's perception that they are being tested may cause discomfort that leads to laughter, self-deprecating remarks by the respondent, or reassurance by the interviewer (Gathman et al. 2008). These social tensions are not provided for in the rules of standardization, and the improvisations by the interviewer may affect the respondent's willingness to engage in further disclosures or their level of motivation to work hard at answering. Interviewing practices and the training that interviewers receive in how to behave in standardized ways also shape the interviewer's behavior. Future research on how conversational practices enter into the interaction between interviewer and respondent will help researchers understand better which behaviors might affect the quality of measurement and how to adapt standardized interviewing practices to changing technology and to maintain the motivation of the respondent.

One occasion on which interviewers deviate from the script occurs when respondents volunteer a lot of information all at once. This conversational practice presumably is occasioned by the respondent's inference about what the interviewer may ask next. So the respondent may tell the interviewer, "Everybody who lives in this household is white," for example. When that happens, if the interviewer is in a situation where the interview schedule instructs her to ask the respondent the race of each household member, the interviewer must quickly determine how to manage the situation. The interviewer still must follow the rules of standardization, but the interviewer now knows the answers to upcoming questions – or at least what the respondent thinks are the answers. When a respondent gives information to the interviewer "prematurely," the interviewer must balance interactional practices that require that she show she has heard and understood what the respondent just said with the practices of standardization. "Verification" or "confirmation" is a label for interviewing practices that some interviewing shops deploy in this situation, for example: "The next question is 'How old were you on your last birthday?' I think you said you were 65. Is that correct?" In situations like this the interviewer may not ask all the questions they were supposed to ask, or the interviewer may not follow the rules of standardization because they are trying to manage the information that the respondent has supplied.

There are number of different sites at which interviewers deviate from the survey script or the practices of standardization. Examples include:

- During the initial reading of the question.
- During follow-up behaviors, for which there are principles of standardized interviewing, but not an actual script. Follow-up actions include:
  - Providing definitions authorized by the instrument or project training

  - Feedback or acknowledgments
  - Other follow-up behaviors

- When respondents provide information in variable form, such as during an event history calendar, timeline, or grid.
- When respondents ask questions or make comments.

There are surprisingly few studies that have examined deviations from the survey script at these sites using a strong criterion and taking into account that respondents are nested within interviewers, and this is an area that would benefit from future research. (See review in Schaeffer and Dykema 2011b.) The few studies that have been conducted seem to indicate that changes in question wording make little difference for reliability (Groves and Magilavy 1986; Hess et al. 1999). Record-check studies that compared answers to records have found that, for most questions examined, substantive changes in question wording by the interviewer had no effect. But there are a few instances in which changes increased or decreased response accuracy (Dykema and Schaeffer 2005; Dykema et al. 1997). No explanation has been identified for these observed differences in the impact of question reading on accuracy, and this warrants future research.

Behaviors of the interviewer other than question reading are also important. "Probing" is a difficult behavior to identify reliably. Some studies refer to "probing" and others to "follow-up." When probing or follow-up occurs, it is almost always associated with lower-quality data, regardless of the adequacy of the interviewer's follow-up or their adherence to standardized practices of follow-up. This is presumably because the interviewer's follow-up was occasioned by the inadequacy of the respondent's answer – an inadequacy that the interviewer might not be able to remedy with her follow-up techniques.

Providing definitions can improve the respondent's understanding of complex concepts when the respondent's situation requires that definition. Sometimes standardized interviews include a definition as part of the wording of a question or as part of an instruction to an interviewer to be used "as needed." "Conversational interviewing" (Schober and Conrad 1997) augmented the requirement to read a question as worded by authorizing the interviewer to offer definitions ad-lib when the interviewer thought they were needed. However, there have not been studies that compare the *ad lib* method of providing definitions with other methods of providing definitions (many of which are discussed by Schober and Conrad 1997 in their paper on this topic). It would be helpful if future research on interviewing practices expanded the comparison, not just to providing *ad lib* definitions versus no definitions, but to examining other ways of providing definitions, and doing it in a study design that allowed both variable errors and bias to be assessed simultaneously. There is also a need for the right statistical design and analyses to be applied in this sort of research; in particular, future research should ensure that there are sufficient numbers of interviewers and respondents per interviewer. Analyses should model the structure of the data including the hierarchical structure of respondents nested within interviewers.

Deviations can be thought of as initiated by either respondent behavior (that is, the deviation by the interviewer is in response to the actions of the respondent) or by the interviewer. In the case of deviations initiated by the behavior of respondents, research suggests that these classes of behavior are associated with tensions in standardization (see Schaeffer and Dykema 2011b; Schaeffer and Maynard 2008). Some deviations begin with informative contributions by the respondent. These informative contributions include:

- Relevant information in place of properly formatted answer ("reports")

    - Substantive responses that suggest that an assumption in the question or the response categories does not fit the respondent's situation ("just reading glasses")
    - Synonyms for response categories ("probably")
    - Uncertainty markers ("never thought about it")

- Respondents may also initiate deviations from the script by additions to a properly formatted answer. These include:

    - Information beyond that requested by the initial question but sought by subsequent questions (e.g., "yes, I talked with my husband")
    - Qualifiers (e.g., "I think") and mitigators (e.g., "about")
    - Considerations (e.g., "yes, I have a job, but I'm on maternity leave right now")
    - Other deviations begin with interruptions by the respondent during the reading of the question or the reading of the response categories. When the respondent interrupts — even with an otherwise codable answer — the interviewer must still complete the reading of the question and response categories, and the interviewer must engage in training of the respondent.

These behaviors of respondents can occur because of state uncertainty (meaning that the respondent is unsure of their answer) or because of task uncertainty (the respondent is unsure how to fit their answer into the structure of the task) (Schaeffer and Thomson 1992).

Interviewers may depart from the rules of standardization during different types of actions:

- During the initial reading of the question
  - By offering definitions or response categories in questions where they were not scripted
  - By tailoring questions in a series or battery of questions to reduce repetition
- During follow-up actions that confirm or code respondents' answers
  - By engaging in tuning, that is, working to get a more precise response from the respondent by only repeating response categories that appear to be in the vicinity of the respondent's answer
  - By verification of an answer that was provided in response to a previous question
- During follow-up to an answer that is not adequate by doing some combination of the following:
  - Providing or applying definitions (e.g., "A 'weekday' would not include Saturday or Sunday")
  - Reducing or simplifying the task (e.g, "you can just give us your best guess")
  - Asking follow-up questions that target an ambiguity (e.g., "So would you say they 'understood you completely'?")
  - Repeating all or part of the question (e.g., "the question says 'understood you completely'") or response categories
- When giving feedback or acknowledgments after the respondent has answered, for example, by engaging in one of these actions:
  - Receipting or acknowledging the answer
- With the token "Okay"

- By confirming or repeating the respondent's answer
  - Announcing a return to agenda (e.g., "And the next question asks…")
  - Reinforcing and motivating ("Thank you. That's the kind of information we are looking for.")

Some topics or tasks in the survey instrument are probably more common sites for departures from standardization. These include complex topics such as batteries of evaluation questions that use the same response categories, household listings or rosters to determine the structure of a household and event history calendars or complex tasks such as physical measurements, cognitive assessments, and obtaining permission for records linkage. These are just some examples of common areas for which the quality of measurement could be improved by attention to the design of the instrument and the development of appropriate interviewing practices.

Interviewing practices are a complex array of intersecting, and occasionally colliding, demands that interviewers must navigate. It is easy for researchers to focus on ways that interviewers deviate from prescribed behaviors and condemn the negative influence of these behaviors on data quality, but all too often the role of the instrument is ignored. In fact, when interviewers are well trained and monitored, most deviations from the script are probably due to problems in the wording or structure of the questions or to ways in which the questions do not fit the situation of the respondent. Observers may assume that there is a good reason for every protocol and approach in a survey instrument, and it is the duty of the interviewer to "simply" follow the script and collect the data. In reality, as exhibited in the discussion of rosters, no instrument can provide a complete script, and some elements of the instrument design may be frustrating for both respondents and interviewers. These frustrating aspects of the interview may have neutral or negative (if they increase interviewer variability) effects on data quality. On rare occasions, features of the interview that the respondent finds frustrating may motivate the respondent to break off and not complete the interview. Although such breakoffs are not common in interviewer-administered interviews, motivation to participate in subsequent interviews could suffer. Any negative consequences of the experience of the interview for respondents add to the complications interviewers face in responding to demands that they achieve high response rates.

The role of the instrument in occasioning deviations from the script implies that researchers bear some responsibility for fielding instruments that minimize the occasions on which interviewers deviate from good interviewing practices. Interviewing is a constrained and specialized interaction because of the needs of measurement, but it is still an interaction between people, and instrument designers need to bear this in mind. By observing interviewers and respondents in action, researchers can see the problems that interviewers and respondents face and the ways that they solve them; these observations can and should be used to inform instrument design and interviewer training. Lastly, changing and novel interviewing technologies and the varieties of types of information being collected by surveys may require innovations in interviewing practices. Survey researchers should remain keenly aware of the demands that these new challenges place on interviewers and the possible consequences for the quality of measurement.

Studies of interviewing practices require research designs that can allow us to draw conclusions because they preserve features of large-scale production surveys, include and document the methods for training and monitoring interviewers, include manipulation checks, and assess reliability or validity or both.

Areas for future research include the following:

- How the interaction between the interviewer and respondent affects both the motivation of the respondent and the quality of the resulting measurement in a way that considers interviewer effects.
- How conversational practices enter into the interaction between interviewer and respondent, their impact on the motivation of the respondent, and the quality of resulting measurement.
- How the interviewing practices that the interviewer uses to manage information that the respondent supplies before it is requested (e.g., "verification") affects the motivation of the respondent and the quality of measurement, and what practices interviewers should use to deal with this information.
- Under what conditions do changes that the interviewer makes in the wording of a question when it is originally read lead to an increase or decrease in the accuracy of responses?
- How to improve practices that interviewers use to follow up answers that express uncertainty so that the quality of the resulting data is improved.
- How effective are different methods for providing respondents with definitions for complex target objects considering that the assessment must

  – Consider both variable errors and bias simultaneously
  – Include methods suitable for long production surveys with many interviewers

- Reassessing principles of questionnaire design to find methods that reduce the burden on both interviewers and respondents.
- Building "smarter" survey systems that integrate and display information previously recorded and allow interviewers to enter answers in the order that the respondent provides them so that interviewers do not need to ask for redundant information.
- How real interviewers and real respondents interact with the survey technology and questionnaires in the real world to devise improved question designs and rules for interviewing.
- What interviewing practices do we need for complex interviews of the future that include such complex tasks as physical measurement, cognitive assessments, and so forth.

References and Further Reading

Belli, R. F., & Lepkowski, J. M. (1996). "Behavior of Survey Actors and the Accuracy of Response." Pp. 69–74 In *Health Survey Research Methods Conference Proceedings, DHHS Publication No. (PHS) 96–1013*, edited by R. Warneke. Hyattsville, MD: Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics.

Conrad, F. G., & Schober, M. F. (2000). "Clarifying Question Meaning in a Household Telephone Survey." *Public Opinion Quarterly* 64: 1–28.

Dijkstra, W. (1987). "Interviewing Style and Respondent Behavior: An Experimental Study of the Survey Interview." *Sociological Methods and Research* 16: 309–334.

Dykema, J., Lepkowski, J. M., & Blixt, S. (1997). "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study." Pp. 287–310 In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin. New York: Wiley-Interscience.

Dykema, J., & Schaeffer, N. C. (2005). "An Investigation of the Impact of Departures from Standardized Interviewing on Response Errors in Self-Reports About Child Support and Other Family-Related Variables." Paper presented at the annual meeting of the American Association for Public Opinion Research, May, Miami Beach, FL.

Fuchs, M. (2000). "Screen Design and Question Order in a CAI Instrument: Results from a Usability Field Experiment." *Survey Methodology* 26: 199–207.

Fuchs, M., Couper, M. P., & Hansen, S. E. (2000). "Technology Effects: Do CAPI or PAPI Interviews Take Longer?" *Journal of Official Statistics* 16(3): 273–286.

Fuchs, M. (2002). “The Impact of Technology on Interaction in Computer-Assisted Interviews.” Pp. 471–491 In *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D. W. Maynard, H. Houtkoop-Steenstra, J. Van Der Zouwen, & N. C. Schaeffer. New York: Wiley.

Gathman, E., Cabell, H., Maynard, D. W., & Schaeffer, N. C. (2008). “The Respondents are All Above Average: Compliment Sequences in a Survey Interview. *Research on Language and Social Interaction* 41: 271–301.

Groves, R. M., & Magilavy, L. J. (1981). “Increasing Response Rates to Telephone Surveys: A Door in the Face for Foot in the Door?.” *Public Opinion Quarterly* 45: 346–358.

Groves, Robert M. and Lou J. Magilavy. (1986). “Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys.” *Public Opinion Quarterly* 50(2): 251–66.

Hak, T. (2002). “How Interviewers Make Coding Decisions.” Pp. 449–470 In *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D. W. Maynard, H. Houtkoop-Steenstra, J. Van Der Zouwen, & N. C. Schaeffer. New York: Wiley.

Hess, J., Singer, E., & Bushery, J. M. (1999). “Predicting Test-Retest Reliability from Behavior Coding.” *International Journal of Public Opinion Research* 11: 346–360.

Mangione Jr., T. W., Fowler, F. J., & Louis, T. A. (1992). “Question Characteristics and Interviewer Effects.” *Journal of Official Statistics* 8: 293–307.

Moore, R. J., & Maynard, D. W. (2002). “Achieving Understanding in the Standardized Survey Interview: Repair Sequences.” Pp. 281–312 In *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. Van Der Zouwen. New York: Wiley.

O’Muircheartaigh, C., & Campanelli, P. (1998). “The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision.” *Journal of the Royal Statistical Society, Series A* 161: 63–77.

Schaeffer, N. C., & Dykema, J. (2011a). “Questions for Surveys: Current Trends and Future Directions.” *Public Opinion Quarterly* 75: 909–961.

Schaeffer, N. C., & Dykema, J. (2011b). “Response 1 to Fowler’s Chapter: Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions.” Pp. 23–39 In *Question Evaluation Methods: Contributing to the Science of Data Quality*, edited by J. Madans, K. Miller, A. Maitland, & G. Willis. Hoboken, NJ: John Wiley & Sons, Inc.

Schaeffer, N. C., & Maynard, D. W. (2008). “The Contemporary Standardized Survey Interview for Social Research.” Pp. 31–57 In *Envisioning the Survey Interview of the Future*, edited by F. G. Conrad & M. F. Schober. Hoboken, NJ: Wiley.

Schaeffer, N. C., & Thomson, E. (1992). “The Discovery of Grounded Uncertainty: Developing Standardized Questions about Strength of Fertility

Motivation." Pp. 37–82 In *Sociological Methodology 1992*, Vol. 22, edited by P. V. Marsden. Oxford: Basil Blackwell.

Schaeffer, Nora Cate, Dana Garbarski, Jeremy Freese, and Douglas W. Maynard. (2013). "An Interactional Model of the Call for Participation in the Survey Interview: Actions and Reactions in the Survey Recruitment Call." *Public Opinion Quarterly* 77(1): 323–51.

Schnell, R., & Kreuter, F. (2005). "Separating Interviewer and Sampling-Point Effects." *Journal of Official Statistics* 21: 389–410.

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

Suchman, L., & Jordan, B. (1990). "Interactional Troubles in Face-to-Face Survey Interviews." *Journal of the American Statistical Association* 85: 232–253.

Nora Cate Schaeffer is Sewell Bascom Professor of Sociology and the Faculty Director of the University of Wisconsin Survey Center at the University of Wisconsin-Madison. Her current research focuses on interaction when the sample member is recruited and during the interview and on instrument design issues.

16

How to Improve Coding for Open-Ended Survey Data: Lessons from the ANES

Arthur Lupia

In surveys, responses to open-ended questions are often released to the public in coded, categorized forms. The typical rationale for converting open-ended responses to coded categories is to protect respondent privacy. Specifically, survey participants sometimes express their responses in ways so unique that others could use that information to identify them. Many leading surveys have concluded that the public release of such information is not worth the risk to respondent privacy and produce codes instead. Many survey researchers, moreover, like the apparent convenience associated with using coded data rather than having to work on their own to translate words into statistically analyzable elements.

Unfortunately, methodological approaches to open-ended coding in survey research have been anything but consistent and credible. The lack of widely accepted best practices might be partially responsible for the decline in the use of the open-ended response format. This is unfortunate because poorly defined methods and best practices should not prevent the use of open-ended responses that can often provide a uniquely rich source of data to survey researchers.

The fundamental question that researchers seeking to code open-ended data must address is "What is the correct inference for a user to draw from a coded open-ended response to a survey question?" The answer to that

A. Lupia (✉)
University of Michigan, Ann Arbor, USA
e-mail: lupia@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_16

question is going to depend on what question the survey asked, what the respondent says, and then very importantly, on processing decisions that are made after the interview is conducted about converting respondent answers into numbers. A key point is that researchers creating and using these data need to pay greater attention to possible problems caused by inattention to these details.

For this discussion, examples will be drawn heavily from the 2008 ANES, so it is worth outlining the key features relating to open-ended questions on the ANES. There are four general types of questions that the ANES asked in the open-ended format: (1) "most important problem," (2) candidate "likes-dislikes," (3) party "likes-dislikes," and (4) political knowledge.

The "political knowledge" questions are noteworthy because they are among the most frequently used variables that the ANES produces. These questions solicit answers to fact-based quiz questions about politics such as

> Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers and the like.... What about...William Rehnquist – What job or political office does he NOW hold?

This question is administered in an open-ended format, allowing respondents to answer in their own words. In most cases, the participant's complete response has been recorded verbatim. Later, all answers have been coded as either "correct" or "incorrect." For decades, the publicly available versions of these codes have been treated as valid measures of respondents' knowledge of the question. However, questions have been raised about the accuracy and relevance of this data.

Generally, ANES data users expect the study investigators to convert open-ended responses into numbers indicating a response as correct or incorrect. A critical point about this process is that the users base their inferences on beliefs about what each of these numbers means. Many users believe that open-ended coding is easy to do, that it generates valid measures, and that it's performed well by survey organizations. The evidence tells a different story. For the 30 years that the ANES has been asking open-ended recall questions, there is little to no record of users asking critical questions such as: "How did you make decisions about what answers were correct or incorrect?" There is a similarly sparse record of requests for reliability statistics. Yet, the questions are widely used under the assumption that the data are valid and reliable.

Moreover, after examining the ANES's open-ended question data under more intense scrutiny, researchers found that the reality of the state of open-ended coding is much worse than many users assumed. Coding practices for other major surveys were examined they were found to be similarly disappointing. So it is important to evaluate where things have gone wrong.

Consider, for example, problems and controversies associated with the ANES question about the position that William Rehnquist held. Many studies have examined this and similar items and drawn such conclusions as

> "Close to a third of Americans can be categorized as 'know-nothings' who are almost completely ignorant of relevant political information, which is not, by any means, to suggest that the other two-thirds are well informed." (Bennett 1998)
>
> "The verdict is stunningly, depressingly clear: most people know very little about politics." (Luskin 2002)

Indeed, a study by Gibson and Caldiera (2009) found that only 12 percent of respondents provided a "correct" response to the Rehnquist question. At first glance, this statistic appears to support the critics' conclusions. However, a closer examination of the coding of the responses to the question found that, for the 2004 ANES, responses were marked as "correct" only if respondents specifically identified that Rehnquist was "Chief Justice" and on the "Supreme Court" – meaning that the addition 30 percent of respondents that identified him as a Supreme Court justice were marked "incorrect" due to not specifying "Chief Justice." When Gibson and Caldiera (2009) asked respondents to state whether William Rehnquist, Lewis F. Powell, or Byron R. White was Chief Justice, 71 percent correctly identified Rehnquist. Similarly, on the 2000 ANES, 400 of the 1,555 respondents said that Rehnquist was a judge or said that he was on the Supreme Court but were coded as having answered incorrectly. Other "incorrect" answers from the 2000 ANES included

- "Supreme Court justice. The main one."
- "He's the senior judge on the Supreme Court."
- "He is the Supreme Court justice in charge."
- "He's the head of the Supreme Court."
- "He's top man in the Supreme Court."
- "Supreme Court justice, head."
- "Supreme Court justice. The head guy."

- "Head of Supreme Court."
- "Supreme Court justice head honcho."

Along similar lines, another political knowledge question asked "…Tony Blair, What job or political office does he NOW hold?" For this item there was a serious error in the coding instructions. ANES coders were told that in order to be marked as "correct," "the response must be specifically to 'Great Britain' or 'England' – United Kingdom is *NOT* acceptable (Blair was not the head of Ireland), nor was reference to any other political/geographic unit (e.g. British Isles, Europe, etc.) correct." In fact, Blair was "Prime Minister of the United Kingdom." While many scholars and writers used this variable to characterize public ignorance, there is no record that any such user questioned how it was produced prior to subsequent ANES PIs discovering this error.

With this example in hand, we return to questions about how these errors occur and what can be done to improve open-ended coding practice. The answer to the first question is that researchers don't really know due to poor recordkeeping prior to 2008. Typically for the ANES, interviewers would transcribe the respondent answers and staff would then implement a coding scheme, often with cases coded by a single person. There are few or no records of instructions given to staff regarding coding and there is no documentation of coding reliability analyses. We later discovered that such recordkeeping inadequacies are not particular to the ANES and are common across major surveys.

In response to these discoveries, the ANES moved to make redacted transcripts available whenever possible, conducted a conference to discover and develop best practices, and formed expert committees to develop mutually exclusive and collectively exhaustive coding schemes. The key outcomes included establishing transparent and replicable processes for turning text into numerical data.

The expert discussion about "partial knowledge" was particularly interesting. Recall that the question begins, "What's the job or political office" that a particular person holds. For example, when this question was asked about Dick Cheney in 2004, a lot of people provided the correct response of "Vice-President." Others, however, would say "anti-Christ" or "chief puppeteer." How should these responses be coded? Some experts wanted to count such responses as representing partial knowledge. A breakthrough occurred when the experts returned to the question wording and determined to focus the coding on whether or not a correct answer was given to the question that the ANES actually asked. So if a respondent says that Dick Cheney shot his friend in a hunting accident, they have decided not to answer the question. They may be providing knowledge about Dick Cheney, but still not answering the question as it was asked. So the whole concept of "partial knowledge" was changed by asking, "is it partial knowledge with respect to this question?"

The ANES's new coding framework for open-ended questions about political office place emphasis on the following factors:

- Did the respondent say something about the political office?
- Did they identify any part of the title of this person's political office correctly?
- Some people have multiple offices; for example, the Vice-President of the United States is also President of the Senate. The person who is the Speaker of the House is also a congressperson.

The first factor on which the new coding scheme focuses is "did the respondent say anything correct or partially correct about the political office?" Since the question asks about the person's job, a correct answer to the question that was asked would also constitute descriptions of what this person does, of what their job is, so things like making legislation or organizing a political party, would also constitute a correct or partially correct response. So for each of the questions the ANES asked, a long list of jobs that a particular person has was identified. So, unlike before where the ANES produced a single "correct" or "incorrect" code for each question, now they have a political office code: "Does the respondent give a correct answer regarding the political office?" The ANES also has a code to indicate if the respondent provided a complete description of a job or an incomplete description? And then finally there is an "other" code. And "other" is anything that the person says that is not pertaining to the job or political office of that person. With respect to the "other" responses, the ANES coding system does not make any judgments about them. In sum, the codes reflect the question, "Did the respondent name the political office, did they name the job, or have they said anything else?"

In terms of procedural transparency, the ANES has decided that it is important for users to be able to see how the code frame was implemented so they can figure out whether there is something about the coding practices that skew the numbers. Written documentation of all decisions, written documentation of all conversations that the ANES had with coding vendors, particularly regarding instructions that confused coders, are now available on the ANES website. The goal of making such information available is for people who are using the data, or who are developing their own coding schemes, to be able see what the ANES did. With such information in hand scholars who obtain different results than the ANES have better ways of determining "Why?"

Several attributes of the new ANES coding scheme provide some insight into what may be best practices for other surveys or researchers to consider when coding open-ended responses. First, the scheme is theoretically defensible, second, it has demonstrated high inter-coder reliability, third, it is mutually exclusive and collectively exhaustive, and fourth, it is accessible to other scholars who may want to use public data to compare other code frames. More general steps that have been taken include:

- Increased documentation at all stages
- Evaluation at many stages
- Increased procedural transparency
- High inter-coder reliability

In summary, if researchers believe that the scientific credibility of survey-based research depends on transparency and replicability of analysis, then it is imperative that data be manufactured and distributed in ways that facilitate transparency and replicability. The kinds of practices needed to help survey producers and analysts make more effective decisions about how to code open-ended data and interpret the resulting variables are not what they could be. Future research should identify ways to improve practice (and data quality) in ways that need not require extra resources to implement. Doing so should give a wide range of scholars the direction and motivation needed to increase documentation of past coding schemes and engage in more systematic thinking about how best to develop future coding algorithms.

Areas for future research:

- Effective approaches to increasing coding transparency.
- Identifying ways to develop shared coding schemes and practices between survey organizations that are asking the same questions.
- Specifying best practices regarding the production and dissemination of documentation.

References and Further Reading

Bennett, Stephen. (1998). "Know-Nothings Revisited: The Meaning of Political Ignorance Today." *Social Science Quarterly* 69: 476–490.

Gibson, J. L., & Caldeira, G. A. (2009). Knowing the Supreme Court? A Reconsideration of Public Ignorance of the High Court. *The Journal of Politics*, *71*(2), 429–441.

Luskin, R. C. (2002). From Denial to Extenuation (and Finally Beyond): Political Sophistication and Citizen Performance. In J. H. Kuklinski (Ed.), *Thinking about Political Psychology*. Cambridge, UK: Cambridge University Press.

Arthur Lupia is the Hal R. Varian Professor of Political Science at the University of Michigan and research professor at its Institute for Social Research. He examines how people learn about politics and policy and how to improve science communication. His books include *Uninformed: Why Citizens Know So Little About Politics and What We Can Do About It*.

He has been a Guggenheim fellow, a Carnegie Fellow, is an American Association for the Advancement of Science fellow, and is an elected member of the American Academy of Arts and Sciences. His awards include the National Academy of Sciences Award for Initiatives in Research and the American Association for Public Opinion's Innovators Award. He is Chair of the National Academy of Sciences Roundtable on the Application of the Social and Behavioral Science and is Chairman of the Board of Directors for the Center for Open Science.

17

Applying Human Language Technology in Survey Research

Mark Liberman

Human Language Technology (HLT) is a broad class of approaches to computer processing of speech and text. The term HLT comes from work sponsored over the past three decades by the Defense Advanced Research Projects Agency in its Speech and Language program. HLT includes methods such as Natural Language Processing and Computational Linguistics, and covers many applications.

When the input is text, HLT refers to tasks that include

- Document retrieval ("Googling")
- Document classification
- Document understanding
- Information extraction from text
- Summarization
- Question answering
- Sentiment analysis (opinion mining)
- Machine translation

When the input is speech, HLT tasks include

M. Liberman (✉)
University of Pennsylvania, Philadelphia, USA
e-mail: myl@cis.upenn.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_17

- Speech recognition (“speech to text”)
- Speech synthesis (“text to speech”)
- Speaker identification/classification
- Language identification
- Diarization (“who spoke when”)
- Spoken document retrieval
- Information extraction from speech
- Question answering
- Human/computer interaction via speech
- Speech-to-speech translation

As a result of the past decades of research, HLT methods are now good enough that survey researchers are increasingly looking to them as a means of quickly coding open-ended responses to survey questions without needing to hire and train human coders. HLT researchers have been aware of the potential to apply these methods in similar applications for over 15 years, yet they have not seen widespread use in survey research.

There are a number of potential reasons for the slow adoption of HLT in survey research. First, HLT applications might not work well enough to replace human coders in this context, because the error rate that researchers are willing to accept is relatively low and the variability in results between different HLT methods can be quite high. Second, early attempts at using HLT for survey applications were seen as unsuccessful, and as a result many researchers write off the approach even though the technology has improved considerably. Third, generic out-of-the-box HLT solutions will not always work to solve particular survey research problems, and there has not been enough demand for survey-specific research to generate methods targeted at this application. This is further complicated by the fact that the process of figuring out whether or not any given pre-existing approach will work in a particular survey context may be somewhat protracted and difficult.

One reason that HLT applications in the survey context can be particularly difficult is the fact that human annotation of text, in the absence of careful definition and training, is extremely inconsistent. If you give annotators a few examples, or a simple definition, and turn them loose, the resulting level of inter-annotator agreement is generally quite poor. This is true even in the identification of apparently well-defined entities, which are things like people, places, organizations, or, in a biomedical domain, genes, gene products, proteins, disease states, organisms, and so on. It is even harder to annotate relationships among entities in a consistent way.

For the sake of understanding the complexity of the annotation task in identifying entities consider an example: If you were to take a large group of Ph.D. political scientists, give them scientific papers that are about areas in their specialization, and ask them if they can determine when politicians are mentioned in those papers, they'll look at you like you're an idiot because the task seems so simple. However, if you were to take two of them, put them in separate rooms, and ask them to do the same task, and then look at how well they agree on the output, you would be very fortunate if they agree 50 percent of the time. It gets even worse for what are called "normalized entities." That is, instead of simply asking, "is this referring to a politician" you want to know which individual is referred to.

For example, here are some of the issues that the researchers would be likely to encounter: Which categories count as "politicians" (judges, attorneys general, political pundits, appointed officials, protest leaders, etc.)? Do references to groups count (e.g., "the members of the Warren court")? What about names used as modifiers ("Stalinist techniques")? What about specific but as yet unknown individuals ("the Republican nominee for president in 2016")?

These problems arise because human generalization from examples to principles is variable, and human application of principles to new examples is equally variable. As a result, the "gold standard," is not naturally very golden. The resulting learning metrics are noisy, and an F-score (the harmonic mean of precision and recall) of 0.3 or 0.5 is not a very attractive goal. If researchers tell people that they can write a program that will agree with their intuitions 30 percent of the time, they're not very impressed, even if their intuitions and their neighbor's intuitions only agree 30 percent of the time.

For research on information extraction from text, HLT researchers have developed an iterative approach that can achieve relatively high rates of agreement among human annotators. This process is analogous to the development of common law: a relatively simple statute says what the rules are, and then a long list of particular cases provides precedents that can be applied to new instances. (Of course, survey researchers face analogous problems in deciding how to classify the answers to open-ended questions.)

The resulting guidelines for HLT annotation tasks can sometimes be hundreds of pages long. These approaches are slow to develop and hard to learn, but in the end they can produce inter-annotator agreement rates of 90 percent or better. And with adequate amounts of training data of this type, there are standard ways to create automated systems whose agreement with human annotation is nearly as good.

So, while HLT holds great promise for automating the coding of open responses to survey questions, there is still a considerable amount of research that is needed before it will see widespread application in practice.

This is an important area for ongoing research for at least three reasons. First, it will reduce the costs of analyzing open responses to surveys and enable the open response format to be more widely used. As was mentioned in the question design section before, this format is an important source of high-quality survey data. Second, a considerable amount of archival open response data could be analyzed in a more consistent manner than that current patchwork approach of having different teams of human coders conduct coding at different points of time.

While it seems like a daunting task to reconcile the current state of HLT with respect to the needs of survey research, it is important to continue pushing research forward into unknown and untested domains. Allan McCutcheon drove this point home when he stated:

> We can't plan for the technology of today. We've got to start thinking about the technology of tomorrow. Right? And the technology – I mean if you told people ten years ago that you'd be talking **to** your cell phone, they would have said, "Well, yeah, to my mother." Right? But, "No, no, no, you'll be talking **TO** your cell phone." Right? They'd look at you as if you just had your head screwed on wrong. But today, people are doing it, and they're saying, "Well, it doesn't do it perfectly. It doesn't understand me as well as my mother does." Wait for ten years.

Areas for future research include:

- Developing survey-specific HLT applications for coding open response data
  - Identifying the unique features of survey data that will need to be addressed by the statistical algorithms in HLT
  - Identifying similar projects in other research domains that could benefit from the development of HLT in the survey domain to maximize intellectual cross-fertilization and share development costs
- Identifying additional external sources of text that can be analyzed and coded by HLT and linked with survey data (e.g., government or medical records)

- Applications of sentiment analysis to survey responses (including even yes/no responses).
- Identifying ways for HLT and survey research to simultaneously break new ground via collaborative research projects that will benefit both field

Mark Liberman is an American linguist. He has a dual appointment at the University of Pennsylvania, as Christopher H. Browne Distinguished Professor of Linguistics, and as a professor in the Department of Computer and Information Sciences. He is the founder and director of the Linguistic Data Consortium. Liberman's main research interests lie in phonetics, prosody, and other aspects of speech communication. His early research established the linguistic subfield of metrical phonology. Much of his current research is conducted through computational analyses of linguistic corpora.

18

Maintaining Respondent Trust and Protecting Their Data

Roger Tourangeau

An important concern for all researchers collecting survey data is respondent confidentiality. Even surveys that do not collect sensitive data need to be concerned about confidentiality since it affects willingness to participate in surveys and thus may influence response rates and nonresponse bias.

Privacy concerns can be thought of involving a respondent's willingness (or unwillingness) to reveal information at all; respondent feelings or beliefs that "it's none of your business" in response to a request or potential request for information, reflects concerns about privacy. Even respondents who are willing to divulge information to the researchers may not want that information shared with anyone else; concerns about the latter are confidentiality concerns. And there are at least two different classes of third parties that respondents might be concerned about. The respondent may fear that somebody in the respondent's household might overhear what he or she said during an interview and might learn something that the respondent would rather they didn't know. Or he or she may fear that some third party outside the household will get hold of the data, maybe a criminal or another federal agency.

Many survey interviews are not done in private. Some research on the American National Election Studies (Anderson et al. 1988) suggests that up

R. Tourangeau (✉)
Westat, Rockville, MD, USA
e-mail: RogerTourangeau@westat.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_18

to 40–50 percent of the interviews are done with some other household member present. Other evidence comes from the World Mental Health Surveys (Mneimneh et al. 2015), which are conducted in many countries around the world. The findings indicate that there is wide variability across participating countries in interview privacy conditions. For example, in Japan, about 13 percent of the interviews were done with somebody else present, whereas in India 70 percent are done with another person present. These findings are more than a little troubling because a lack of privacy may influence respondent willingness to report certain attitudes or behaviors.

There are three major groups of factors that affect whether someone beside the interviewer or respondent is likely to be present during a survey interview. The first set involves the household's characteristics – its size, whether the respondent is married, and, if so, whether the spouse is employed. Second are cultural norms regarding privacy. In some countries, respecting people's privacy is a value; in other countries (e.g., those with collectivist cultures), sharing with other people is more important and privacy is not an important value. A third set of variables involves the amount of effort that interviewers make to provide a private interview context. There seems to be much variation across interviewers in how much they understand that the interview is supposed to be done in private and in how much effort they make to provide those conditions for the respondent.

The consequences of a lack of privacy also vary depending on several factors. First, it depends on whether the other person present during the interview already knows the information being requested of the respondent and, if not, whether there are likely to be repercussions if he or she finds out the respondent's answer. Perhaps as a result, there are typically lower levels of reporting of sensitive behaviors when the respondent's parents are present but fewer effects when the respondent's spouse is present. More generally, there is evidence of increased social desirability bias when interviews are not done in private.

One approach to dealing with the issues surrounding privacy and confidentiality concerns is to offer anonymity to the respondents. This is often hard to do convincingly, especially in face-to-face surveys where the interviewer clearly knows the household address and may also ask for the respondent's signature on a consent form. However, some studies attempt to collect data anonymously despite these challenges. Monitoring the Future, for example, is a study of high school seniors about drug use, and its questionnaires are sent to schools, where they are distributed in classrooms. There is no identifying information on the questionnaires. Even in the cases where this can be accomplished, there are concerns about the potentially negative effects of anonymity on respondent accountability and data quality (Lelkes et al. 2012).

The three items discussed thus far – privacy, confidentiality, and anonymity – are typically discussed in the context of another issue, which is asking sensitive questions. There are at least three distinct meanings for "sensitive question":

- Intrusiveness: The question is inherently offensive (thus, it is the question rather than the answer that is sensitive);
- The risking of disclosure to third parties (various types of third parties), including
  - Other family members or persons,
  - Other agencies, or
  - Analysts or hackers (disclosure avoidance methods are designed to reduce likelihood that this will happen); and
- Social desirability (socially approved vs. socially disapproved answers).

Social desirability is the focus of much of the attention given to confidentiality, privacy, and anonymity in survey research. In a classic description of the problem, Sudman and Bradburn (1974) say:

> Some questions call for the respondent to provide information on topics that have highly desirable answers … If the respondent has a socially undesirable attitude or if he has engaged in socially undesirable behavior, he may … desire to appear to the interviewer to be in the socially desirable category. It is frequently assumed that most respondents resolve this conflict in favor of biasing their answer in the direction of social desirability. (pp. 9–10)

There are three primary concerns about the consequences of question sensitivity. The first is unit nonresponse, meaning that people may fail to participate at all if they think they will be asked sensitive questions. The second is missing data, meaning that some respondents may skip offensive or embarrassing questions. The third and perhaps greatest concern is reporting errors. This refers to respondents overreporting desirable attitudes or behaviors and underreporting undesirable ones, in either case, providing false information to the researcher.

Fortunately, researchers have come up with a number of techniques for addressing these concerns about reporting errors. First, self-administration seems to help because respondents no longer need to worry about self-presentation to the interviewer. Some surveys are primarily interviewer-administered but the sensitive questions are asked in a separate self-administered section. Second, open-ended responses have been demonstrated to provide better data than closed items. Finally, the randomized response technique (RRT) and bogus pipeline approaches (both described in more detail later) have shown promise for reducing inaccuracy in reporting. Two key early papers on these topics are Locander et al. (1976) and Blair et al. (1977).

A popular approach to minimizing social desirability bias has been the RRT. This involves estimating the prevalence of some characteristic without knowing what question any specific respondent received. The response reflects one of two statements. For example, statement A might be, "I am for legalized abortion on demand," and statement B is, "I'm against legalized abortion on demand." Respondents get one or the other of these items with some known probability. A common randomization approach is a coin flip. RRT often seems to work, in that researchers get a higher estimate of various sensitive characteristics under RRT than from a direct question. However, no production survey uses this method because it is difficult to implement in the field and because it increases the variance of the estimates. Because of the impracticality for real application to large production surveys, RRT may not be high priority for future work.

Other clever methods have also been developed, including the item count technique (ICT) and the bogus pipeline, but it's not clear whether they add much in terms of being widely applicable as approaches to reducing measurement error due to sensitive questions. These approaches (RRT, ICT, bogus pipeline, etc.) may have promise for certain very specific applications but none of them is sufficient to address the real scope of the problem posed by sensitive questions. RRT and ICT have the additional drawback that they do not produce values for individual respondents and only aggregate statistics can be formed, which reduces their utility further.

Researchers often worry about privacy and confidentiality, but many surveys are still done in the presence of other people; this implies that perhaps more training is needed to impress on interviewers that privacy is an important condition to achieve for conducting interviews. Second, measurement error can be a very large problem on surveys that ask sensitive questions, often swamping other sources of error, at least at high levels of aggregation. This suggests that more studies should be done to identify the specific, most worrisome sources of error for each survey because there is wide variability in how different types of error affect different surveys.

Certainly, surveys in general should continue focusing resources on issues like sampling bias, but for some surveys it is likely that measurement error due to sensitive questions is a larger component of the total survey error. Thus, reducing measurement errors should a goal to which more resources are devoted. Third, self-administration seems to help reduce reporting error but it is not sufficient to eliminate it; there is still considerable room for improvement.

In terms of future research, we need to devise new methods for collecting accurate information on sensitive topic. And, in thinking a little bit more about recommendations to the National Science Foundation, the research that, in my view, ought to be funded falls under three headings – causes, consequences, and fixes.

Under the "causes" heading: First, most researchers focus on "lying" by respondents; but that term may be too strong for what's really going on. My colleagues and I have used the phrase "motivated misreporting," but we just don't really understand the processes that lead to misreporting very well. What is it that people are thinking, and what are they doing? These are key areas for future research. It may be a semiconscious process that influences these misreports. It's possible that people are so adept at dodging embarrassing questions in everyday life, that they do it unthinkingly in surveys. It may be a kind of conversational skill carried over into the interview setting. Many researchers seem to think that once they invoke "social desirability," they are done, and that is the explanation for the measurement errors. The second area on the causal side where more work is needed is on what topics people regard as embarrassing. The presumption is that these potentially embarrassing topics are the ones they'd lie about, but that may not be true either. Future research is needed to develop a firm understanding of what topics are really sensitive and for whom. We need research to understand both the processes by which people modify their answers and the determinants of sensitivity that lead them to do this.

In terms of "consequences": first, more studies are needed to evaluate the relative magnitude of the different sources of survey error. These studies are tough to do but they will help researchers to avoid simply shooting in the dark. It would be good to have a body of a hundred studies, not just two or three, that look at this issue. Then researchers could say, "These hundred studies have looked at the relative magnitudes of the different sources of error, and they suggest that we really ought to be worrying about X, at least when the topic is Y." Second, researchers need to be cleverer about record check studies and other forms of validating information provided by respondents.

Regarding “fixes” for these problems, it is very unclear what the right strategies are given the current state of research. Researchers have spent a lot of time over the years on things like RRTs. But moving forward, further variations on RRT are not part of the solution. This research needs to go in some new directions and break away from techniques that are either ineffective or not widely applicable.

Areas for future research:

- Identifying best practices for ensuring that surveys are conducted in private settings, without other people present
- Mapping the cognitive processes underlying inaccurate answers to sensitive questions
- Evaluating the relative magnitude of different sources of survey error for individual surveys
- More and better approaches to validating respondent answers to get a better sense for actual levels of misreporting on particular items
- New practical methods for collecting sensitive information

References and further reading

Anderson, B. A., Silver, B. D., & Abrahamson, P. R. (1988). The effects of the race of the interviewer on measures of electoral participation by blacks in SRC National Election Studies, *Public Opinion Quarterly*, *52*, 53–88.

Blair, E., Sudman, S., Bradburn, N., & Stocking, C. (1977). How to ask questions about drinking and sex: Response effects in measuring consumer behavior, *Journal of Marketing Research*, *14*, 316–321.

Lelkes, Y., Krosnick, J. A., Marx, D. M., Judd, C. M., & Park, B. (2012). Complete anonymity compromises the accuracy of self-reports, *Journal of Experimental Social Psychology*, *48*, 1291–1299.

Locander, W. B., Sudman, S., & Bradburn, N. M. (1976). An investigation of interview method, threat, and response distortion, *Journal of the American Statistical Association*, *71*, 269–275.

Mneimneh, Z. M., Tourangeau, R., Pennell, B.-E., Heeringa, S. E., & Elliott, M. R. (2015). “Cultural Variations in the Effect of Interview Privacy and the Need for Social Conformity on Reporting Sensitive Information.” *Journal of Official Statistics*, *31*, 673–697.

Sudman, S., & Bradburn, N. (1974). *Response effects in surveys: A review and synthesis*. Chicago: Aldine.

Roger Tourangeau is a Vice President in the Statistics Group and co-director of Westat's Survey Methods Group. Tourangeau is known for his research on survey methods, especially on different modes of data collection and on the cognitive processes underlying survey responses. He is the lead author of *The Psychology of Survey Response*, which received the 2006 AAPOR Book Award, and he was given the 2002 Helen Dinerman Award, the highest honor of the World Association for Public Opinion Research for his work on cognitive aspects of survey methodology. He is also the lead author of *The Science of Web Surveys*. Before coming to Westat, Dr. Tourangeau worked at NORC, the Gallup Organization, and the University of Michigan; while at the University of Michigan, he directed the Joint Program in Survey Methodology for nine years. He has a PhD from Yale University and is Fellow of the American Statistical Association.

19

Tackling Panel Attrition

Peter Lynn

Introduction

Longitudinal panel surveys involve repeatedly collecting data from the same respondents over time (Lynn 2009), typically for the purpose of studying dynamic processes. They can vary greatly in scale, complexity, objectives, and resources, but what they have in common is that to achieve their objectives and provide valuable data they must retain sample members in the survey and successfully collect data from them repeatedly over a period of time. The phenomenon of failing to retain all sample members is known as panel attrition (Binder 1998). The term reflects the idea that the sample tends to gradually erode, as members are lost for various reasons. Though the essay by Olsen in this volume proposes a shift of attention from respondent attrition to data attrition, retaining respondents remains a prerequisite to collecting data, so the avoidance of panel attrition remains essential even from Olsen's viewpoint. The potential reasons for attrition are many, and some will be specific to the design or context of a particular survey, but they can be sorted conceptually into three main categories (Lepkowski and Couper 2002): non-location, non-contact, and refusal to cooperate.

Non-location is a failure to locate a sample member at a particular wave of a panel survey. This generally only happens when the sample member's

P. Lynn (✉)
University of Essex, Colchester, UK
e-mail: plynn@essex.ac.uk

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_19

location details (i.e., residential address, phone number, etc.) have changed since the previous wave at which they were successfully located (Couper and Ofstedal 2009). The extent to which failure to locate sample members contributes to panel attrition is therefore dependent on the prevalence of location change between waves, which is a function of the extent of mobility in the study population and the between-wave interval.

The process of making contact with a sample member takes rather different forms for different data collection modes. In a face-to-face at-home interview survey, contact requires the sample member to answer the door when the interviewer visits. In a telephone survey, contact requires the sample member to answer the telephone when the interviewer calls. In both cases, the likelihood of successful contact is a function of the interaction between when the respondent is at home and when the interviewer attempts to make contact. In a mail or web self-completion survey, contact requires the sample member to receive and pay attention to the (e)mailing inviting them to take part in the survey.

Once contact has been made, the sample member may or may not agree to cooperate. The good news for the second and subsequent waves of panel surveys is that people who have already cooperated on at least one previous occasion are relatively likely to agree to do so again, partly because of the tendency toward consistency when it comes to compliance behavior (Groves et al. 1992). Nevertheless, some sample members will not always agree to cooperate. As for any survey, this decision will depend partly on survey-specific factors and partly on external and situational factors, but a feature unique to panel surveys is that sample members already have direct experience of what it is like to participate. As a consequence, the respondent's perceptions of the time taken to participate, the cognitive burden, the sensitivity of the questions, and so on, are likely to have a direct impact on the likelihood of continued participation. Panel surveys that are particular burdensome to respondents, that are uninteresting, or that cause embarrassment or anxiety, are therefore at increased risk of suffering from attrition due to non-cooperation.

Why Is Panel Attrition Important?

Panel attrition is an important problem for researchers for two main reasons. First, a high *rate* of attrition will cause the initial sample to rapidly dwindle in size and it may soon become too small to provide useful estimates due to low precision. For example, 20 percent attrition at each wave would result in less than one-third of the initial sample remaining after five waves. Due to the need for repeated measures, simply replacing the lost sample members with new samples is not a solution as it is rarely if ever possible to collect retrospectively the data that would have been collected at the earlier waves. The size of the sample that has provided data for *all* waves determines the precision of estimates.

Second, regardless of the *rate* of attrition, the *nature* of attrition can introduce bias to survey estimates (Fitzgerald et al. 1998). If the tendency for attrition depends on relevant characteristics of sample members, the sample will become skewed in terms of those characteristics, potentially introducing bias. The three main categories of reasons for attrition (non-location, non-contact, and refusal) can have distinctive impacts on nonresponse bias. For example, sample members who cannot be located will tend to be those who have recently moved, and may also tend to be disproportionately those who have changed jobs, or have ended or started a relationship, and so on. A survey aiming to study phenomena that are related to one or more of these types of life events could suffer particularly severe nonresponse bias unless considerable efforts are made to minimize the extent, and therefore the impact, of attrition due to a failure to locate sample members. Also, the extent and nature of nonresponse bias can change over time as more waves of data are collected. This may happen if the balance of reasons for attrition changes over waves and if the reasons depend on relevant characteristics of sample members. For example, it could happen that attrition at the early waves is dominated by refusals related to the survey content, while later attrition is dominated by failure to locate, and that the types of people who refuse to continue participation due to their views on the survey content are different from the types of people who move and cannot subsequently be located for future waves.

Ways to Reduce Attrition

High precision and low bias are fundamental to survey objectives. As panel attrition is a serious threat to both, considerable resources are often committed to tackling attrition. Panel surveys are designed and implemented in ways that prioritize sample retention. The methods used to minimize attrition generally aim to tackle one or more of the three categories of causes (non-location, non-contact, and refusal). The extent to which these methods reduce attrition, and the level of resources required to implement them, depend partly on characteristics of the study population, such as mobility rates (Lepkowski and Couper 2002; Behr et al. 2005) and socio-demographics. The efforts made to reduce attrition will depend on the survey objectives but broadly all the features outlined later could in principle be applied to any kind of panel survey.

A survey design feature with major implications both for resource requirements and for the control of attrition is the choice of data collection mode(s). Face-to-face data collection using field interviewers offers the greatest chance to locate sample members (Couper and Ofstedal 2009) and also tends to produce the highest cooperation rates (De Leeuw 2005). But it is also the most expensive of data collection modes, as interviewers must be reimbursed for their time and expenses traveling to sample members' homes, in addition to the time spent carrying out interviews. Self-completion data collection methods (e.g., web or mail) are much less expensive, but offer much more limited opportunities for locating a sample member who has moved or for persuading a reluctant sample member to continue participating. Increasingly, panel surveys are seeking a balance between the cost and attrition implications of different modes by employing mixed-mode designs in which personal home visits are utilized only when other cheaper modes have been unsuccessful in securing participation (Lynn 2013; Jäckle et al. 2015).

Features that can reduce attrition due to non-location include the design of mailings requesting address updates (Fumagalli et al. 2013; McGonagle et al. 2013), the use of multiple search methods to trace movers (Groves and Hansen 1996; Laurie et al. 1999), and the between-wave interval (Duncan and Kalton 1987; Taylor and Lynn 1997). As is the case for cross-sectional surveys, non-contact rates, and consequent attrition, can be reduced by increasing the number, and diversity, of contact attempts. However, panel surveys additionally offer the opportunity for better-informed choice regarding the timing of contact attempts by using paradata from previous waves to tailor the attempts (Lagorio 2016) or to prioritize attempts with sample units predicted to be hard-to-contact (Calderwood et al. 2012).

In addition to the choice of data collection mode(s), features that can reduce attrition due to refusals include the use, and level, of respondent incentives (Laurie and Lynn 2009), the size of interviewer workloads (Nicoletti and Buck 2004), interviewer continuity between waves (Lynn et al. 2014), and questionnaire length (Zabel 1998; but see also Lynn 2014a). Undoubtedly, the sample member's perception of the enjoyment, difficulty, burden, etc., associated with previous participation also plays an important role in the decision regarding whether to take part again (Kalton et al. 1990; Hill and Willis 2001; Olsen 2005). Future research could usefully focus on identifying which survey design features under the researcher's control affect the extent to which panel sample members enjoy participation or find it easy and hence affect refusal-related attrition.

However, most of the scientific literature concerned with testing the effects of survey design features on attrition is focused on mean effects, in other words the sample-level effect of applying the survey design feature uniformly to the whole sample. This may be because researchers generally try to standardize survey procedures, so that all sample members for a particular survey are treated in exactly the same way. Each person is sent the same letter, offered the same incentive, called using the same call scheduling algorithm, and so on. This orthodoxy has been challenged from time to time but, with the exception of interviewer introductions (Groves et al. 1992; Morton-Williams 1993), most survey design features remain standardized. However, many survey design features have been shown to have effects that are heterogeneous across subgroups of sample members (e.g., the form and value of incentives (Singer 2002; VanGeest et al. 2007), the length of the invitation letter (Kaplowitz et al. 2012), and interviewer calling patterns (Bennett and Steel 2000; Campanelli et al. 1997)).

Pushed by the need to make efficiency gains due to budget cuts in recent years, survey researchers have begun to exploit this heterogeneity of effects by moving away from standardized survey designs to designs in which some features are targeted to different subgroups of the sample (Lynn 2014b). The idea is simple: if a particular design features with an associated cost (e.g., extra callbacks on Sundays) is only effective for a particular sample subgroup, then it should only be applied to that subgroup. Or if different versions of a feature (e.g., motivational statement, type of incentive, reminders) are optimum for different subgroups, then each subgroup should be applied the version that is optimum for them. Targeting can be applied to more than one design feature on the same survey, possibly using different target groups (e.g., Luiten and Schouten 2013). Furthermore, it should be noted that targeting can be used to tackle either or both of the precision and bias issues mentioned earlier. To minimize the *level* of attrition, each sample subgroup can be assigned the design features that should maximize participation rates. To minimize attrition *bias*, costly but effective design features can be restricted to subgroups that would otherwise suffer from higher attrition rates, thereby improving sample balance.

To target design features requires knowledge about membership of relevant subgroups and about the relative effectiveness of features across subgroups. Panel surveys are in a strong position to meet these requirements, given the wealth of relevant information collected about sample members at previous waves, including substantive measures, paradata, and participation behavior. In future, rather than questioning *whether* design features should be targeted, researchers may ask themselves *which* features of their survey should be targeted. In failing to target design features, we may be making it harder to optimize sample retention within budget constraints.

Ways to Reduce the Statistical Impact of Attrition

Data collection is not the end of the story regarding the effect of attrition on survey estimates. Adjustment methods, notably through sample weighting, can be used to reduce the negative impacts of attrition on estimation, particularly in terms of bias. It is therefore the combination of data collection and adjustment that determine the bias of estimates. The essay by Brick in this volume points out that relatively little is known about links between data collection procedures and statistical adjustments. Brick suggests that certain data collection procedures may reduce the need for later adjustment. But arguably it might be more effective to identify (relatively inexpensive) adjustment procedures that could reduce the need for relatively expensive data collection procedures. Panel survey data may help to shed light on these questions using rich auxiliary data from previous waves. Some recent research suggests that attempts to improve sample balance at the data collection stage improves survey estimates over and above the improvement that can be achieved through weighting alone (Schouten et al. 2016; see also Tourangeau et al. 2017), though these findings are to some extent context-specific. The interplay between data collection and adjustment in the context of panel attrition is certainly a ripe area for further research.

A few points about weighting adjustments for panel attrition are worth noting as some issues are rather distinct from those relevant to the context of cross-sectional surveys. First, adjustment often involves dealing with hierarchical phases of nonresponse, where the possibility of participating at each step in the process depends on having participated at each previous step. Examples of this include surveys, of which there are many, that only attempt to collect data at wave t from sample members who responded at the wave $t - 1$. This data collection policy produces data with a monotone attrition pattern, meaning that unit nonresponse is a one-way process: once data is missing for a wave it will never be present at subsequent waves. In this situation, there is a choice to be made regarding the number of steps in the weighting process. If the weights are calculated in a single step, only frame and external data can be used as auxiliary variables, limiting the extent to which weighting is likely to reduce bias. But if each successive wave is treated as a separate step, so that covariates can be used from the prior wave, there may be increased random variation in the weights (as the coefficients of each weighting model are only sample-based estimates, subject to the usual sampling error). Intermediary solutions are also possible. There is little or no research-based guidance on how best to make these choices.

If missing data patterns are not restricted to monotone attrition, then after n waves the number of wave combinations that could potentially be of interest as an analysis base is $2^n - 1$. With even a modest number of waves, it becomes impractical for data providers to provide users with such a large number of weighting variables. Instead, it is necessary to identify a manageable subset of wave combinations, recognizing that for some analyses only suboptimal weighting will be available. How best to identify the appropriate subset is an open question in need of additional research (Lynn and Kaminska 2010).

Keeping track of the eligibility status of each sample member can be challenging or even impossible in the presence of attrition, especially when a panel survey runs for many years. Once contact with a sample member has been lost, there are typically very limited means available to ascertain whether they are still alive and still resident in the country/region of the study, for example. If some transitions to ineligibility are not identified, weighting adjustments will tend to over-represent remaining sample members who share covariate characteristics with those who have become ineligible (e.g., died) but whose ineligibility is not known to the researchers. Estimates can become biased as a result, so it is important both to make efforts to update eligibility status for all sample members, including those who have dropped out of the survey (e.g., through regularly checking death records), and to adapt weighting methods to deal with the likely under-identification of ineligibility (e.g. by estimating the extent of under-identification within subgroups and applying a subgroup-level adjustment to the weights; Sadig 2014).

Summary and Conclusion

In summary, researchers need to employ a broad range of techniques to minimize the risk to panel survey estimates posed by attrition. These techniques encompass both data collection and statistical adjustment. Many of the techniques are already well known, and researchers are using them in increasingly sophisticated ways, reflecting the complexity of panel attrition. The use of targeted design features and other adaptive design methods exemplifies this recent sophistication.

However some techniques, particularly in the weighting domain, are under-researched. Much remains to be done to identify the most effective ways of implementing nonresponse adjustment weighting for panel surveys and to better understand the interplay between data collection and weighting.

Areas for future research:

- More effective use of micro-level paradata from previous waves to target participation-enhancement methods at the current wave
- Identifying how survey design features under the researcher's control affect the extent to which panel sample members enjoy participation or find it easy and hence affect subsequent refusal-related attrition
- Identifying more effective ways to target design features to sample subgroups with high risks of attrition
- Identifying adjustment procedures that could reduce the need for relatively expensive data collection procedures
- Better understanding the implications of alternative weighting adjustment methods to deal with monotone attrition patterns
- Identifying the most appropriate subsets of wave combinations for which to produce adjustment weights and understanding the consequences for analysis based on other subsets
- Adapting weighting and estimation methods to deal with increasing uncertainty over time (waves) regarding the continuing eligibility of sample members

References and Further Reading

Behr, A., Bellgardt, E., & Rendtel, U. 2005. "Extent and determinants of panel attrition in the European Community Household Panel." *European Sociological Review* 21(5): 489–512.

Bennett, D. J., & Steel, D. 2000. "An Evaluation of a Large-Scale CATI Household Survey using Random Digit Dialling." *Australian and New Zealand Journal of Statistics* 42(3): 255–270.

Binder, D. 1998. "Longitudinal surveys: why are these surveys different from all other surveys?." *Survey Methodology* 24: 101–108.

Calderwood, L., Cleary, A., Flore, G., & Wiggins, R. D. 2012. "Using response propensity models to inform fieldwork practice on the fifth wave of the Millenium Cohort Study." Paper presented at the *International Panel Survey Methods Workshop*, Melbourne, Australia.

Campanelli, P., Sturgis, P., & Purdon, S. 1997. *Can You Hear Me Knocking? An Investigation into the Impact of Interviewers on Survey Response Rates*. The Survey Methods Centre at SCPR, London, GB. Available at http://eprints.soton.ac.uk/80198/.

Couper, M. P., & Ofstedal, M. B. 2009. "Keeping in Contact with Mobile Sample Members." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 183–203. Chichester, UK: Wiley.

De Leeuw, E. D. 2005. "To mix or not to mix data collection modes in surveys." *Journal of Official Statistics* 21(2): 233–255.

Duncan, G. J., & Kalton, G. 1987. "Issues of design and analysis of surveys across time." *International Statistical Review* 55: 97–117.

Fitzgerald, J., Gottschalk, P., & Moffitt, P. 1998. "The impact of attrition in the Panel Study of Income Dynamics on intergenerational analysis." *Journal of Human Resources* 33(2): 300–344.

Fumagalli, L., Laurie, H., & Lynn, P. 2013. "Experiments with Methods to Reduce Attrition in Longitudinal Surveys." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 176(2): 499–519.

Groves, R. M., Cialdini, R. B., & Couper, M. P. 1992. "Understanding the decision to participate in a survey." *Public Opinion Quarterly* 56(4): 475–495.

Groves, R. M., & Hansen, S. E., 1996. *Survey Design Features to Maximise Respondent Retention in Longitudinal Surveys*. Report to the National Center for Health Statistics. Ann Arbor: University of Michigan.

Hill, D. H., & Willis, R. J. 2001. "Reducing panel attrition: a search for effective policy instruments." *Journal of Human Resources* 36(3): 416–438.

Jäckle, A., Lynn, P., & Burton, J. 2015. "Going online with a face-to-face household panel: effects of a mixed mode design on item and unit nonresponse." *Survey Research Methods* 9(1): 57–70.

Kalton, G., Lepkowski, J., Montanari, G. E., & Maligalig, D., 1990. "Characteristics of second wave nonrespondents in a panel survey." In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 462–467. Washington, DC: American Statistical Association.

Kaplowitz, M. D., Lupi, F., Couper, M. P., & Thorp, L. 2012. "The effect of invitation design on web survey response rates." *Social Science Computer Review* 30(3): 339–349.

Lagorio, C., 2016. "Call and response: Modelling longitudinal contact and cooperation using Wave 1 call records data." *Understanding Society Working Paper* 2016–01. Colchester: University of Essex. https://www.understandingsociety.ac.uk/research/publications/working-papers.

Laurie, H., Smith, H., & Scott, L. 1999. "Strategies for reducing nonresponse in a longitudinal panel study." *Journal of Official Statistics* 15(2): 269–282.

Laurie, H., & Lynn, P. 2009. "The use of respondent incentives on longitudinal surveys." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 205–233. Chichester, UK: Wiley.

Lepkowski, J. M., & Couper, M. P. 2002. "Nonresponse in the second wave of longitudinal household surveys." In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little, 259–272. New York: Wiley.

Luiten, A., & Schouten, B. 2013. "Tailored fieldwork design to increase representative household survey response: An experiment in the survey of consumer satisfaction." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 176(1): 169–189.

Lynn, P. 2009. "Methods for longitudinal surveys." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 1–19. Chichester UK: Wiley.

Lynn, P. 2013. "Alternative sequential mixed mode designs: effects on attrition rates, attrition bias and costs." *Journal of Survey Statistics and Methodology* 1(2): 183–205.

Lynn, P., Kaminska, O., & Goldstein, H. 2014. "Panel attrition: how important is interviewer continuity?." *Journal of Official Statistics* 30(3): 443–457.

Lynn, P. 2014a. "Longer interviews may not affect subsequent survey participation propensity." *Public Opinion Quarterly* 78(2): 500–509.

Lynn, P. 2014b. "Targeted response inducement strategies on longitudinal surveys." In *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis, 322–338. Abingdon UK: Psychology Press.

Lynn, P., & Kaminska, O. 2010. "Criteria for developing non-response weight adjustments for secondary users of complex longitudinal surveys." Paper presented at the *International Workshop on Household Survey Nonresponse*, Nürnberg, Germany.

McGonagle, K. A., Schoeni, R. F., & Couper, M. P. 2013. "The effects of a between-wave incentive experiment on contact update and production outcomes in a panel study." *Journal of Official Statistics* 29(2): 261–276.

Morton-Williams, J. 1993. *Interviewer Approaches*. Aldershot: Dartmouth.

Nicoletti, C., & Buck, N. H. 2004. "Explaining interviewee contact and co-operation in the British and German household panels.." In *Harmonisation of Panel Surveys and Data Quality*, edited by M. Ehling & U. Rendtel, 143–166. Wiesbaden, Germany: Statistiches Bundesamt.

Olsen, R. J. 2005. "The problem of respondent attrition: survey methodology is key." *Monthly Labour Review* 128: 63–70.

Sadig, H., 2014. "Unknown eligibility whilst weighting for non-response: The puzzle of who has died and who is still alive?" ISER Working Paper 2014-35. Colchester: University of Essex. https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2014-35

Schouten, B., Cobben, F., Lundqvist, P., & Wagner, J. 2016. "Does more balanced survey response imply less non-response bias?." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 179(3): 727–748.

Singer, E. 2002. “The use of incentives to reduce nonresponse in household surveys.” In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little, 163–177. New York: Wiley.

Taylor, S., & Lynn, P., 1997. “The effect of time between contacts, questionnaire length, personalisation and other factors on response to the Youth Cohort Study,” *Department for Education and Employment Research Series*, no.8.

Tourangeau, R., Brick, J. M., Lohr, S., & Li, J. 2017. “Adaptive and responsive survey designs: a review and assessment.” *Journal of the Royal Statistical Society Series A (Statistics in Society)* 180(1): 203–223.

VanGeest, J. B., Johnson, T. P., & Welch, V. L. 2007. “Methodologies for improving response ates in surveys of physicians: A systematic review.” *Evaluation and the Health Professions* 30(4): 303–321.

Zabel, J. E. 1998. “An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labour market behavior.” *Journal of Human Resources* 33(2): 479–506.

Peter Lynn has a Chair in Survey Methodology at the University of Essex, UK. He has been co-Investigator of Understanding Society: the UK Household Longitudinal Study since its inception in 2008 and specializes in all aspects of longitudinal survey methods, including sampling, measurement, nonresponse, weighting, and inference, topics on which he has published more than 50 journal articles, books and chapters. He is Editor of the 2009 Wiley book, “Methodology of Longitudinal Surveys” and the 2014 Routledge book, “Improving Survey Methods”. Peter was one of the founders of the European Survey Research Association (ESRA), and was founding Editor of the ESRA open access journal, Survey Research Methods. He previously served as Editor of the Journal of the Royal Statistical Society Series A. In 2003 the (UK) Royal Statistical Society awarded him the Guy Medal for services to survey methods. Peter’s full CV is at www.iser.essex.ac.uk/people/plynn.

20

Respondent Attrition Versus Data Attrition and Their Reduction

Randall J. Olsen

For the past 25 years, longitudinal surveys have been an important and valued source of data collection. Of the major NSF-funded surveys, the PSID is primarily longitudinal in nature and the ANES and GSS both also have components that are longitudinal. Panel surveys have particular value for collecting time-varying information within the same population of respondents so these data can provide evidence of change in measures over time that are less prone to recall errors and biases. Another benefit of a panel study is that, over several waves, the cost of data collection may be lower than an equal number of similarly sized and equally lengthy cross-sectional surveys.

Researchers are also increasingly taking steps to make cross-sectional surveys more longitudinal. For example, the Integrated Public Use Microdata Series project at the University of Minnesota has linked records of the Decennial Census to form a longitudinal record and the American Community Survey has been used as a screener for other studies, suggesting that it might also become an originating source of longitudinal data collection. Furthermore, researchers have periodically linked the groups in the rotating panel design of the Current Population Survey to form a panel structure. These efforts to incorporate elements of panel designs into cross-sectional studies provide an indication of the value of longitudinal data to researchers.

R.J. Olsen (✉)
Ohio State University, Columbus, USA
e-mail: olsen.6@osu.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_20

However, longitudinal survey data collection efforts are commonly plagued by respondent attrition from the panel. For example, the long-standing PSID and the newer but similarly structured British Household Panel Survey (BHPS) both display a similar steep decline in response between initial recruitment and the subsequent few waves. In both surveys, the panel attrition hazard rate was above 10 percent in the first waves before settling between 1 and 2 percent attrition per wave for the PSID (Schoeni, Stafford, McGonagle, & Andreski 2003) and 3–4 percent attrition per wave for the BHPS (Uhreig 2008). Similar results have been observed across many longitudinal panels over time and the patterns of attrition seem to be consistent with population heterogeneity in innate respondent cooperativeness.

The Educational Longitudinal Survey (ELS), which is a survey of students enrolled in primary school, provides an important perspective on how different approaches to conceptualizing attrition can influence data completeness. The initial wave of the ELS achieved a completion rate of 88 percent of sampled students. Then, in wave 2, the ELS completed interviews with 91 percent of the students that responded in the first wave, representing an attrition rate of 9 percent. Wave 3 completed surveys with 87 percent of wave 1 respondents, which was only a 4 percent increase in total attrition. However, the decline in the attrition rate in wave 3 was partially due to an effort by the study to re-contact wave 2 non-respondents and recover data that was initially missed during wave 2.

The approach that the ELS applied of attempting to recover data from wave 2 respondents during the wave 3 interviews enabled them to recover 96 percent of the wave 2 data. This reveals an important consideration when evaluating how to deal with attrition in longitudinal surveys. Rather than being concerned about respondent attrition from the panel in each wave, the investigators may be better served by focusing on reducing overall data attrition by re-contacting non-respondents from prior waves in an attempt to fill in data that should have been collected in those prior waves.

The National Longitudinal Surveys (NLS), which is a study that collects data from six cohorts of respondents recruited between 1968 and 1997, has similarly experienced significant attrition over time. However, the study has reduced the impact of this attrition by continually attempting to re-contact non-respondents to prior waves and implementing event history calendar methods to fill in the missing data. These efforts have reduced the incidence of missing data substantially and the success of this approach used by the NLS suggests that a focus simply on respondent attrition may be misguided. Rather, the focus of researchers should be on reducing data attrition by any reasonable means available.

Targeted monetary incentives are one effective approach implemented by the NLS to reduce panel attrition. The differential incentives successfully reduced respondent attrition from the panel, particularly when the largest incentives were targeted at the respondents estimated to have the greatest risk of attrition. This finding supports prior research on targeted incentives as effective tools for increasing response rates but applies it in the context of a panel design (Olsen 2005).

Interestingly, in the course of experimentally demonstrating the effectiveness of targeted incentives, the NLS also found that overall field costs fell due to fewer interviewer resources being expended to gain respondent cooperation. The decrease in field costs was nearly the same as the additional cost incurred by applying the incentives but nonresponse was decreased among the panel members at the highest risk for attrition, which had significant value since the respondents had been members of the panel for 18 waves. These results were replicated in the NLS panel multiple times, confirming the opportunity for significantly reduced panel attrition over time at minimal additional cost when using targeted incentives.

Taken together, the evidence from the ELS and NLS suggests that some of the concerns about panel attrition may be misplaced, and furthermore that it may be possible to mitigate some the effects of attrition at minimal cost. More fundamentally, these findings indicate that researchers should be more concerned about data attrition than panel attrition, meaning that continued efforts to re-contact non-respondents from prior waves should be standard practice in order to maximize the potential to fill in the data from missed waves.

Areas for future research:

- Identifying ways to increase the value of panel surveys to researchers accustomed to using cross-sectional data
- Maximizing the value of having data on people who attrite from panels as there may be valuable insights for non-response to cross-sectional surveys
- Evaluation of the problem of the lack of accretion in panels to account for shifting population demographics due to immigration

References and Further Reading

Olsen, R. 2005. “The Problem of Survey Attrition: Survey Methodology is Key”, *Monthly Labor Review*, 128, 63–70.

Schoeni, R. F., Stafford, F., McGonagle, K. A., & Andreski, P. 2003. “Response Rates in National Panel Surveys”, *The Annals of the American Academy of Political and Social Science 645*(1): 60–87.

Uhreig, N., 2008. The Nature and Causes of Attrition in the British Household Panel Survey, Essex.

Randall J. Olsen is Professor Emeritus of Economics at Ohio State University and Senior Research Scientist. After graduating from the University of Chicago, he was a postdoctoral researcher at the University of Minnesota and then joined the Department of Economics at Yale. From there he moved to Ohio State University. He served as Director of the Center for Human Resource Research where he guided the National Longitudinal Surveys of Labor Market Experience for nearly 30 years. He also helped found the initiative in Population Research at Ohio State. He responsible for many innovations in longitudinal surveys relating to survey content, methods of data collection and data dissemination. He also designed a reformulation of how respondents are engaged and encouraged to cooperate in long-running longitudinal surveys. He has served of several panels for the National Academies related to survey work. He also chaired the Federal Advisory Committee for the National Children's Study.

21

Best Practices for Creating Survey Weights

Matthew DeBell

When generalizing to a population, weighting survey data can be one of the most important steps prior to analyzing the results. The decisions surrounding how weights are calculated and applied can have enormous statistical and substantive implications for the results that are derived from the data. Weighting is often one of the final steps in survey data generation and because of this it is all too often an afterthought for survey researchers. The development and application of survey weights has been the subject of a large body of research within survey methodology and a number of best practices have been developed over the years. However, there are still a number of key areas where more research is needed to fully understand best practices for weighting and disseminating these best practices to the broad community of survey data users.

Survey weights, at their most basic level, determine how many people each respondent represents. Weights are necessary because, even when sampling is done correctly, there are often unequal probabilities of inclusion among sample members and some demographic groups end up under-represented and others over-represented in the data. Survey weights balance these inequalities in the data and enable researchers to transform the observed data such that it is more representative of the target population. In a nutshell, survey weights are applied for the following reasons:

M. DeBell (✉)
Stanford, USA
e-mail: debell@stanford.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_21

- Determining how many people each respondent represents
- Fixing random error (sampling error)
- Adjusting for unequal selection probabilities
- Adjusting for nonresponse
- Fixing non-random error (within limits)

When applied correctly, weights can reduce bias and allow relatively small samples of data to be used to generate inferences about population values. There are a number of approaches to weighting that have been developed and applied widely but in general form the first steps typically include weighting in steps based on selection probabilities, first for household and then for person, then nonresponse in observable categories. After this researchers may post-stratify on key factors (generally demographic) or employ propensity scores.

Weighting is not without its limitations despite having become widely accepted as a statistically and substantively sound approach to correcting data. For example, weights are not useful for fixing non-coverage error because weights cannot be applied to values of zero. Neither can weights fix extreme nonresponse bias or nonresponse bias for factors that are uncorrelated with the weighting factors. Furthermore, weights cannot be used to correct errors on factors with unknown population benchmarks (e.g., party identification). Lastly, weights do not come without a cost. While they are able to reduce bias this comes at the cost of increased variance, which thereby increases standard errors and reduces the precision of estimates.

The literature on survey statistics covers the statistical theory of weighting in excellent detail, but this literature is not accessible to most data users or producers who are not statisticians. For the typical data analyst, there is little guidance for best practices when weighting data. This means that survey statisticians are often needed to apply weights correctly and to evaluate their effects. One area for future research may be to develop a set of procedures that the average data user can implement when applying weights, including a set of standards for when it may be necessary to involve the expertise of a survey statistician.

The result is that many data users are not using survey weights consistently or appropriately. This is largely due to the average survey researcher being unaware of how to apply weights in an appropriate fashion. Even survey statisticians often develop weights in an *ad hoc* manner. This means that the results from analyses are not always transparent, replicable, comparable, or optimal. The clear implication is that more work is needed to identify and disseminate best practices in ways that are accessible to a broader audience than survey statisticians. Thus, there are four key areas for future work:

- More accessible guidance
- Increased transparency of methods
- Clearly replicable methods
- Standardized/comparable methods rather than *ad hoc* approaches

It is impossible to distill weighting to an extremely simple set of rules and procedures that will be appropriate for all surveys. Identifying "standard" practices does not mean closing off alternatives. Rather, it means setting a starting point or frame of reference. There is always more than one way to compute legitimate and effective weights, and these alternative methods each have value and some may be more contextually appropriate than others. However, the need for flexibility in approaches should not stand in the way of building a set of best practices that can guide the average user toward more appropriate treatment of survey weights.

The American National Election Studies (ANES) provides a good example of how researchers can start to approach this task of identifying and disseminating best practices for weighting data. Leading up to the 2008 ANES, the ANES investigators assembled an expert committee of statisticians and survey methodologists to generate a set of guidelines for best practices for weighting with regard to the particular design features of the ANES. From this set of guidelines the ANES investigators and staff developed and published a set of procedures codifying these best practices (DeBell & Krosnick 2009). The goal of these procedures was to describe a general approach to weighting that all users of ANES data or similar data could use; this helped to take the guesswork out of weighting for the average user. The particular procedure recommended by the ANES was a raking algorithm with specific benchmarks and raking factors that were explicitly spelled out in detail. This approach allows the procedure to be standardized for the average user without limiting the flexibility in weighting that might be more appropriate for a specialized user.

But the ANES investigators didn't stop there; they went a step further and assisted in the development of a weighting tool that could easily be used by any researcher to apply the recommended weighting procedure. This tool, "anesrake" is a free package for implementation in the free and open-source "R" statistical program (Pasek, DeBell, & Krosnick, 2014). It is a practical and automated approach to weighting that enables anyone to apply a generic set of best practices for weighting without needing a course in survey statistics. This is a valuable tool to make it easier for more people to create sound survey weights.

However, there are a number of areas where more work is needed. Funding agencies like the NSF can play a key role in moving researchers in a positive direction. For example, by requiring a plan for weighting and a plan for the eventual disclosure of weighting methods to be submitted with grant applications, funding agencies can aid in creating a set of normative standards around the development of transparent and replicable weighting methods. One key point is that future research is needed to identify a set of scientific principles that will guide researchers as they develop and apply weights, thereby moving researchers away from the current practice of researchers using disparate and *ad hoc* methods that are not always fully reported.

Areas for future research:

- Identifying and disseminating a general set of best practices and resources for non-statisticians to use when weighting data
- Improving transparency and replicability when weights are used in practice
- Moving large surveys toward following the ANES model for making weighting accessible to average users

References and Further Reading

DeBell, M., & Krosnick, J. A. (2009). Computing weights for American National Election Study survey data. ANES Technical Report series No. nes012427.

Pasek, J., DeBell, M., & Krosnick, J. A. (2014). Standardizing and Democratizing Survey Weights: The ANES Weighting System and anesrake. http://surveyinsights.org/wp-content/uploads/2014/07/Full-anesrake-paper.pdf

Section 2

Opportunities to Expand Data Collection

22

New Kinds of Survey Measurements

Carli Lessof and Patrick Sturgis

The archetypal survey gathers data by asking respondents questions about their attitudes, opinions, and behaviors. It has long been recognized, however, that questionnaires cannot adequately capture many aspects of people's lives; self-report data can be severely constrained by limitations of human memory and various kinds of self-presentational bias (Tourangeau et al. 2000). Survey methodologists have a long tradition of innovating in light of socio-technological change and the rapid emergence of digital and mobile technologies over the past decade has created exciting opportunities for collecting data in new ways. These developments offer the promise of substantially improving measurement quality and, potentially, of transforming the very nature of survey research. For example, smartphones can reduce reliance on respondent recall by delivering activity diaries "on the go" (Scagnelli et al. 2012) and can be used to validate reported behavior using photographs taken by respondents (Angle 2015). They can enable "in the moment" surveys in the form of very short questionnaires at different times of day to measure subjective well-being (MacKerron 2012) and location-based surveys which are triggered when a study participant enters or leaves a GPS-defined geofence (Kaal et al. 2015).

C. Lessof (✉) · P. Sturgis
National Centre for Research Methods, University of Southampton, England, United Kingdom
e-mail: cl19g15@soton.ac.uk; P.Sturgis@soton.ac.uk

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_22

New measurement approaches can also increase the volume and complexity of information that it is possible to collect in a survey, for example, by asking respondents to scan barcodes of purchased groceries which can then be linked to dietary databases (Carroll et al. 2015). Accelerometers have been an important way of collecting measurement of physical activity for many years (Laporte et al. 1985) and miniaturized sensors can now be used to detect ambient features of the environment, such as temperature or humidity and exposure to pollutants (Nieuwenhuijsen et al. 2015).

New technologies also promise reductions in respondent burden, for example, by using GPS data to track journeys rather than asking respondents for this information by self-report (Feng and Timmermans 2014) and measuring political orientations and social networks by linking to social media profiles. Certain technologies, such as wearable cameras, may be unsuitable for large general population samples but can be used in parallel with other data collection approaches to assess the validity and reliability of a travel diary or the measurement of physical activity using an accelerometer (Kerr et al. 2013; Kelly et al. 2014). The "Internet of Things" also presents opportunities for new kinds of measurement, for example, using Wi-Fi-enabled digital weighing scales in respondents' homes (Mulder 2013) and smart-meters for measuring energy consumption over time (Firth et al. 2008).

These examples are selective but demonstrate the breadth and scope of possibilities for new kinds of measurement that are emerging as the expanding capabilities and falling costs of digital, miniaturized, and connected devices makes their integration within sample surveys increasingly feasible. Indeed, we could cast our net wider still to include innovative biomeasures, mobile health measurements, facial recognition, and eye tracking. By its nature, this is a dynamic and rapidly evolving area and clear-cut definitions of what constitutes "new kinds of measurement" are likely to prove elusive. What perhaps unites them is that they do not in general elicit data by asking respondents questions about their attitudes and behavior.

It is easy to fall prey to hubris when considering the likely impact of new technologies in any context but, we contend, novel forms of measurement have the clear potential to transform survey practice in the future. Much of the existing work in this area has relied on self-selecting samples of highly motivated individuals. If it proves possible to extend their use to large-scale surveys of the general population, the standard survey could change from the conventional notion of a "conversation" between interviewer and respondent (Bradburn et al. 1979), into a flexible vehicle for delivering a range of measurement instruments alongside more standard self-report items. However, significant challenges will need to be overcome for the promise of new technologies to be translated into tangible cost and quality benefits for mainstream surveys. We focus here on three key inter-related challenges to achieving this goal: ethical research practice, respondent burden, and nonresponse.

Ethics of Research

The adoption of novel measures in surveys introduces new complexities for the ethical conduct of research, notably relating to informed consent and privacy. These complexities are likely to vary cross-nationally, in terms of privacy laws, ethical frameworks, and acceptability with regard to social norms. These variations may impose different constraints and, indeed, may well hinder cross-national developments and solutions in this area.

Because surveys provide established mechanisms for formally requesting consent from respondents, many of the ethical difficulties of using new kinds of measurements can be addressed through existing procedures. Respondents can be asked, for example, if they agree to their social media or location data being accessed from their smartphones. However, as the opportunities for gathering sensitive digital data increase, so does the importance of strong ethical scrutiny, often to a level that exceeds standard consent requests. Moving beyond "tick-box" approaches to gaining consent is particularly important in this context because some respondents do not appear to scrutinize consent requests in any detail and are willing to agree to almost unlimited data sharing.

Additional scrutiny is also necessary when measurement instruments collect data about the behavior of people who are not themselves research participants and so have not given consent. Examples include the use of wearable cameras which photograph passers-by (Kelly et al. 2013), or linking to respondents' social media profiles where posts from people in the respondents' network are visible.

Linking surveys with other data sources – geographical, health, financial, and so on – can significantly increase the risk of identity disclosure (Shlomo 2014). As a consequence, studies that combine data in this way often need to specify highly restrictive access conditions. This can seriously constrain researcher access to the data, thereby limiting its value and preventing the replication of research based upon it, an increasingly important feature of modern research practice.

Where new technologies are used for data collection, researchers may try to minimize project costs by using respondents' own devices rather than providing them centrally. However the security standards of commercial products vary and may fail to meet minimum research standards (Hilts et al. 2016). These kinds of challenges are not insurmountable but it is clear that, alongside the mooted benefits of new measurement approaches, some complex ethical and legal implications will need to be addressed by the survey research community.

Respondent Burden

A key assumed benefit of many new measurement approaches is that they will reduce respondent burden. This is most likely to be true when the measurement is passive for the respondent, such as when they are asked to carry a small sensor or install an app on their smartphone. Even in these cases, however, survey designers need to carefully evaluate whether this kind of measurement will indeed be less burdensome than existing procedures. For example, asking a respondent to wear an accelerometer for a week removes the need for him or her to answer a set of questions about exercise and sedentary behavior. This will normally deliver more and better quality data to the researcher compared to self-report, but it imposes a different set of burdens on the respondent such as the need to follow a protocol, to wear an accelerometer throughout the day (and sometimes the night) and to return the device by post at the end of the study period (Yang et al. 2012).

On the other hand, measurements that rely on respondent-initiated data collection will almost certainly increase respondent burden. For instance, asking respondents to enter data and take photographs with a smartphone over an extended period removes the need (and associated cognitive effort) for them to recall past actions but may also involve downloading and installing an app and remembering to take photos each time an event occurs.

Crucially, both passive and active measurement approaches will change the nature of survey participation, from a discrete event of relatively short duration, where the respondent choses what to reveal, to a sequence of tasks carried out over a considerably longer period. This change may, in some contexts, breech existing norms by introducing a greater degree of observation than is the case in the "snapshot" survey interview. For example, environmental researchers can now use in-home monitors and sensors of various kinds to produce fine-grained energy consumption and household temperature measurements. While this may not involve a great deal of burden for the respondent, the insertion of sensors in the home may be perceived by some respondents as an unacceptable intrusion into their privacy. Moreover, the actual degree of respondent burden is unlikely to be zero as the instillation (and maintenance) of equipment would likely require some degree of input from the household.

To be sure, some respondents may be motivated to carry out these kinds of tasks, particularly if they receive feedback about their behavior, or monetary incentives. However, for many respondents, these types of activities will represent an unacceptable imposition on their time and an intrusion into their daily routines. And, moreover, such respondents may be rather different from the general population on the characteristics the survey seeks to measure. The key point here is that while some new technologies will reduce response burden relative to obtaining the same information via self-report, this is not the yardstick by which most respondents are likely to assess the reasonableness of a request. Rather, they will consider the acceptability of the activity on its own merits and, for many, this is likely to result in refusal. Survey practitioners need to carefully consider how novel measurement approaches affect how burden is experienced by respondents if new technologies of this kind are to be successfully integrated into survey data collection (Jäckle 2016; Lessof et al. 2016).

Nonresponse

We have already noted that ethical and privacy concerns will discourage some respondents from taking part in surveys using new kinds of measurement and that new measures may actually be more burdensome than conventional self-reports. It is therefore unsurprising that nonresponse can be an even greater problem for new measurement approaches, with concomitant concerns about the accuracy of estimates. In some research contexts, this matters less but for many academic and government surveys, high rates of nonresponse are likely to prove unacceptable for funders and analysts.

In the UK Millennium Cohort Study, for example, 94 percent of parents agreed to their children wearing an accelerometer, but only half returned usable data. Social disadvantage was associated with nonreturn and reliable data was less likely to be collected, for example, from boys, overweight or obese, and sedentary children (Rich et al. 2013). While the data collected still represents a significant improvement on self-reports for those who were observed (Griffiths et al. 2013), the differential nonresponse limits its more general utility, a problem that will be compounded in future waves, when repeat measures will be available for a relatively small sample.

Existing studies which combine survey responses with linked data from social media accounts have drawn nonrandom samples from existing panels where participants have already consented, or have actively recruited from Twitter or Facebook advertisements (Gibson and Southern 2016; Kim et al. 2016). The challenge for random probability surveys is that only a subset of a general population sample will have social media accounts and, of these, only some will be active users and, of these, not all will consent to linkage. The analytical sample this set of stages produces may end up being rather small and skewed relative to the general population.

A slightly different problem occurs when studies rely on respondents' own devices and equipment such as a smartphone or tablet. Using this approach, respondents who do not own a device will be excluded, while others are likely to be discouraged from participating by the requirement to download software on their own equipment. Some studies, such as the German Internet Panel and the Longitudinal Internet Studies for the Social Sciences (LISS) panel in the Netherlands, have addressed this by providing the necessary hardware and/or Internet access to offline respondents. However, this strategy will only ever be partially successful because a significant minority of the public is unable or does not wish to be online, rendering this type of incentive ineffective. In sum, while new forms of measurement hold great promise for improving data quality and enabling entirely new kinds of measurements to be made, they also open up new forms of under-coverage and nonresponse error that need to be better understood if their net impact on survey error is to be positive (Biemer 2010).

Conclusion

The possibility of new kinds of measurements transforming our conception of the social survey in the years ahead is exciting. There is a growing set of examples from commercial, government, and academic research demonstrating the benefits of innovation in new types of measurement within survey research. On closer examination, however, there are substantial challenges to successful implementation of new measurement approaches at the scale of the general population. These relate primarily to the inter-connected issues of ethics, respondent burden, coverage, and nonresponse. For new measurement approaches to realize their transformative potential, the survey research community must acknowledge and better understand the ethical and data challenges they pose. Assumptions relating to burden and survey cost must be properly assessed and each new measurement approach carefully scrutinized to evaluate whether and how it can be scaled up from small-scale convenience samples to large random samples of the general population.

References and Further Reading

Angle, H. (2015). Moving Beyond Claimed Behaviour: Using Technology and Big Data to Provide Evidence of Behaviour Change. Methodology in Context 2015. London, Market Research Society.

Biemer, P. P. (2010). “Total Survey Error: Design, Implementation, and Evaluation.” Public Opinion Quarterly 74(5): 817–848.

Bradburn, N. M., Sudman, S., & Blair, E. (1979). Improving Interview Method and Questionnaire Design, Jossy-Bass.

Feng, T., & Timmermans, H. J. P. (2014). “Extracting Activity-travel Diaries from GPS Data: Towards Integrated Semi-automatic Imputation.” Procedia Environmental Sciences 22: 178–185.

Firth, S., Lomas, K., Wright, A., & Wall, R. (2008). “Identifying Trends in the Use of Domestic Appliances from Household Electricity Consumption Measurements.” Energy and Buildings 40(5): 926–936.

Gibson, R., & Southern, R., (2016). “2015 and Social Media: New iBES Data Released!.” Retrieved 7 April 2016.

Griffiths, L. J., Cortina-Borja, M., Sera, F., Pouliou, T., Geraci, M., Rich, C., Cole, T. J., Law, C., Joshi, H., Ness, A. R., Jebb, S. A., & Dezateux, C. (2013). “How Active Are Our Children? Findings from the Millennium Cohort Study.” BMJ Open 3(8): e002893.

Hilts, A., Parsons, C., & Knockel, J., (2016). Every Step You Fake: A Comparative Analysis of Fitness Tracker Privacy and Security. Open Effect Report: 76.

Jäckle, A. (2016). Examining the Quality of Data Collected with New Technologies: A Total Error Framework and Plans for Testing. Workshop on Improving the Measurement of Household Finances. Institute for Social and Economic Research, University of Essex.

Kaal, M., Van Den Berg, J., & M.De Gier, M., (2015). Geo-triggered surveys. Presentation to Municipality of Amsterdam. Amsterdam, TNS Nipo unpublished presentation.

Kelly, P., Marshall, S. J., Badland, H., Kerr, J., Oliver, M., Doherty, A. R., & Foster, C. (2013). “An Ethical Framework for Automated, Wearable Cameras in Health Behavior Research.” American Journal of Preventive Medicine 44(3): 314–319.

Kelly, P., Doherty, A., Mizdrak, A., Marshall, S., Kerr, J., Legge, A., Godbole, S., Badland, H., Oliver, M., & Foster, C. (2014). "High Group Level Validity But High Random Error of a Self-Report Travel Diary, as Assessed by Wearable Cameras." Journal of Transport & Health 1: 190–201.

Kerr, J., Marshall, S. J., Godbole, S., Chen, J., Legge, A., Doherty, A. R., Kelly, P., Oliver, M., Badland, H. M., & Foster, C. (2013). "Using the SenseCam to Improve Classifications of Sedentary Behavior in Free-Living Settings." American Journal of Preventive Medicine 44(3): 290–296.

Kim, A., Guillory, J., Bradfield, B., Ruddle, P., Hsieh, Y. P., & Murphy, J., (2016). Information Exposure and Sharing Behavior of e-Cigarette Users: Do Survey Responses Correlate with Actual Tweeting Behavior? AAPOR 71st Annual Conference: Reshaping the Research Landscape: Public Opinion and Data Science, Austen, Texas.

Laporte, R. E., Montoye, H. J., & Caspersen, C. J. (1985). "Assessment of Physical Activity in Epidemiologic Research: Problems and Prospects." Public Health Reports 100(2): 131–146.

Carroll, C., Crossley, T.F., & Sabelhaus, J. (eds.). (2015). *Improving the Measurement of Consumer Expenditures. Studies in Income and Wealth*, Volume 74, Chicago: University of Chicago Press.

Lessof, C., Jäckle, A., Crossley, T., Burton, J.,. P., Fisher, P., Couper, M., Brewer, M., & O'Dea, C., (2016). Understanding household finance through better measurement: Background paper for the 5th Panel Survey Methods Workshop 2016. Panel Survey Methods Workshop 2016, Berlin, Germany.

MacKerron, G. (2012). Happiness and environmental quality, London School of Economics and Political Science.

Mulder, J. (2013). Weighing project August 2012 Questionnaire administered to the LISS panel. Codebook. Centerdata.

Nieuwenhuijsen, M. J., Donaire-Gonzalez, D., Rivas, I., De Castro, M., Cirach, M., Sunyer, J., Hoek, G., Seto, E., & Jerrett, M. (2015). "Variability in and Agreement Between Modeled and Personal Continuously Measured Black Carbon Levels Using Novel Smartphone and Sensor Technologies." Environmental Science and Technology 49(5): 2977–2982.

Rich, C., Cortina-Borja, M., Dezateux, C., Geraci, M., Sera, F., Calderwood, L., Joshi, H., & Griffiths, H. J. (2013). "Predictors of Non-Response in a UK-Wide Cohort Study of Children's Accelerometer-Determined Physical Activity Using Postal Methods." BMJ Open 3(3): 1, 6, 7.

Scagnelli, J., Bailey, J., Link, M., Moakowska, H., & Benezra, K., (2012). On the Run: Using Smartphones to Track Millennial's Purchase Behavior. 67th Annual Conference of the American Association for Public Opinion Research (AAPOR) Orlando, FL.

Shlomo, N. (2014). Probabilistic Record Linkage for Disclosure Risk Assessment. Privacy in Statistical Databases 2014. Ibiza, Spain. Proceedings.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The Psychology of Survey Response, Cambridge: Cambridge University Press.

Yang, L., Griffin, S., Chapman, C., & Ogilvie, D. (2012). "The Feasibility of Rapid Baseline Objective Physical Activity Measurement in a Natural Experimental Study of a Commuting Population." BMC Public Health 12(1): 841.

Carli Lessof is completing her PhD with the National Centre for Research Methods at the University of Southampton in the UK. Before returning to study she helped develop and deliver a number of complex social and biomedical surveys such as the English Longitudinal Study of Ageing, the British birth cohort studies, Understanding Society and Whitehall II. Carli has worked in Government, the not-for-profit and private sectors; as head of Longitudinal Studies and Director of Innovation and Development at NatCen Social Research and Director of Development at TNS BMRB (now Kantar Public). Throughout, she has pursued her interest in innovation in data collection methods.

Patrick Sturgis is Professor of Research Methodology in the Department of Social Statistics and Demography at the University of Southampton and Director of the ESRC National Centre for Research Methods (NCRM). He is past-President of the European Survey Research Association (2011–2015), chaired the Methodological Advisory Committee of the UK Household Longitudinal Survey (2011–2016), and is vice-Chair of the Methodological Advisory Committee of the ESS. He chaired the British Polling Council Inquiry into the failure of the 2015 UK election polls and has published widely in the areas of survey and statistical methods, public opinion and political behavior, social cohesion and trust, social mobility, and attitudes to science and technology.

23

The Role of Surveys in the Era of "Big Data"

Mario Callegaro and Yongwei Yang

Introduction: The Changing Definition of Big Data

The definition of "Big Data" is complex and constantly changing. For example, Dutcher (2014) asked 40 different thought leaders to define Big Data and obtained nearly 40 different definitions. However, there is some consensus in the literature on the main characteristics of Big Data as described by a widely cited Gartner report (Beyer and Laney 2012).

> In terms of ***Volume***, Big Data are those data that cannot be handled by traditional analytics tools.
> In terms of ***Velocity***, Big Data refers data that are coming in (almost) real-time.
> In terms of ***Variety***, Big Data are complex datasets and include very different sources of context such as unstructured text, media content such as images and videos, logs, and other data sources.

Adding to these three key characteristics of Big Data other authors have cited *variability* (how the data can be inconsistent across time), *veracity*

M. Callegaro (✉)
London, UK
e-mail: callegaro@google.com

Y. Yang
Boulder, CO, USA
e-mail: yongwei@google.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_23

(accuracy and data quality) and *complexity* (how to link multiple databases properly). In practice, what is often called “Big Data” may not possess all six of these characteristics (e.g., you can have very large data of great complexity that may not come in with high velocity). We refer the reader to Baker (2016) for an extensive definition of Big Data in the context of survey research.

For the purpose of this chapter, we contrast Big Data with survey data. In this framework, a helpful definition of Big Data was proposed by Groves (2011) who, as a contrast to designed data (survey data), calls it organic data and described it as the data produced by “systems that automatically track transactions of all sorts” (p. 868).

In the survey literature, we find *Big Data thinking* in the emerging term of “Small Big Data” where the authors use multiple survey datasets to enable richer data analyses (Warshaw 2016; Gray et al. 2015). Small Big Data are more and more a reality thanks to the availability of social science data archives (e.g., the UK Data Service or the Roper Center for Public Opinion Research). Although we do not strictly classify them as Big Data per the aforementioned description, they are worth mentioning in this chapter.

Another way to contrast Big Data with survey data is to look at potential sources of Big Data that can answer research questions. Depending on the nature of the research questions, the answer will lie in a continuum of sources from Big Data on one side and survey data on the other side and the combination of the two in the middle – our thesis of this chapter. We identify the following main sources and subclasses. This list is not meant to be highly detailed and comprehensive, and some sets of data cannot be uniquely classified in one or another class:

- *Internet data: Online text, videos, and sound data*. It encompasses all online content relevant to a research question. Using such data is commonly referred to as Internet research methods (Hewson et al. 2016).

  - *Social media data*. Social media data are a subset of Internet data and include text, photos, and videos which are publicly available by mining social media networks such as Twitter and Facebook. Social media data are probably the first and most studied Big Data for public opinion measurement (Schober et al. 2016).
  - *Website metadata, logs, cookies, transactions, and website analytics*. These are data produced by websites and analytics tools (think about Google Analytics or Adobe Analytics) and used heavily in online advertisement, shopping analytics, and website analytics.

- *The Internet of Things*. Internet of Things (IOT) (Gershenfeld et al. 2004) refers to any device that can communicate with another using the Internet as the common transmission protocol. As more and more devices become connected via the Internet, more data are generated and can be used to answer research questions.
  - *Behavioral data* are a subset of the IOT. Behavioral data come from devices such as smartphones, wearable technology, and smart watches carried by subjects and passively recording data such as locations, physical activities, and health status (e.g., Swan 2013). Behavioral data can also be manually recorded by the users.
- *Transaction data*. In the business world, transaction data have been around since before electronic data formats existed. They are records of orders, shipments, payments, returns, billing, and credit card activities, for examples (Ferguson 2014). Transaction data are nowadays part of customer relationship management tools where the attempt is to capture every interaction a customer has with a company or product. The area is also called business intelligence (Hsinchun et al. 2012). The same applies to government and public sector where more and more user interactions are stored digitally.
- *Administrative data*. Administrative data and registers are a form of Big Data collected by public offices such as national health, tax, school, benefits, and pensions, or driver licenses databases. Administrative data have a long tradition of being used for statistical purposes (Wallgren and Wallgren 2014). Survey data can be linked to administrative data as shown by Sakshaug in this volume. Health data in some countries are collected and stored by private companies but, although they are of the same nature of public health data, they are usually not discussed as administrative data in the academic literature.
- *Commercially available databases*. More and more companies are collecting, curating, and storing data about consumers. By using publicly available records, purchasing records from companies, matching techniques (Pasek, this volume), and other algorithms such as imputations from other sources (e.g., census data), these companies create a profile for each individual in their database. They combine data from the previously mentioned sources just described. Examples are Acxiom, Epsilon, Experian Marketing Services, or, in the political domain, Catalist, Aristotle, and NationBuilder. These companies are often referred to as *data brokers* (Committee on Commerce, Science and Transportation 2013).

Finally, and related to survey data, we define *paradata* (Kreuter, this volume) as a source of Big Data. Paradata is data about the process of answering the survey itself (Callegaro 2013), including data collected by systems and third parties (e.g., interviewers) before, during, and after the administration of a questionnaire. Paradata often come in real time (think about collecting answer time per question on a web survey) and are in complex formats (e.g., user agent strings, time latency, mouse movements, interviewer observations).

The Perspectives About Error and Data Quality

Big Data does not necessarily mean good quality or without any error. Often Big Data comes with *Big Noise* (Waldherr et al. 2016). Within the survey research tradition, the concept of survey errors was developed in the early 1940s (Deming 1944) and has since evolved into the Total Survey Error (TSE) framework (Biemer 2010). Applying the concept of survey error to Big Data is a healthy data quality approach where cross-fertilization among the two disciplines is at its best. TSE "refers to the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data" (Biemer 2010, 817). More specifically *specification errors* occure when some concepts that we want to measure in a survey are actually measured differently. *Measurement errors* occur from the interviewers, the respondents, and the questionnaire itself including the data collection methods used to administered the questionnaire. *Frame errors* are errors related to the quality of the sampling frame. The frame can have missing units, duplications, units that are not supposed to be in the frame, and the records themselves can contain mistakes or be outdated. *Nonresponse errors* arise when some respondents (unit nonresponse) do not answer the questionnaire altogether and when some questions (item nonresponse) are not answered (e.g., income question). Finally, there are data *processing errors* stemming from the processes of tabulation, coding, data entry, and producing survey weights.

When applying the general framework of TSE to Big Data we obtain the Big Data Total Error (BDTE) (Japec et al. 2015). Errors in Big Data arise mostly during three steps used to create a dataset from Big Data (Biemer 2014, 2016):

1. *Data generation*. It is specifically the data generation process that differentiates Big Data from surveys, censuses, and experiments (Kreuter and Peng 2014). Data generation in Big Data is sometimes a "black box" and not

always well documented. Errors can take the form of missing data, self-selection, coverage, non-representativeness, and low signal to noise ratio.

2. *Extract, Transform, and Load (ETL)*. This is the process when the data are brought together in the same computing environment with the process of extraction (data accessed, parsed and stored from multiple sources), transformation (e.g., coding, recoding, editing) and loading (integration and storage). Errors in ETL can take the form of matching, coding, editing, and data cleaning errors.
3. *Analysis and visualization*. Here the errors can be of sampling, selectivity, modeling and estimation. Finally, errors might arise in the data visualization step.

It is important to note that the BDTE concept is relatively new, and outside the survey community (e.g., Biemer 2016) "very little effort has been devoted to enumerating the error sources and the error generating processes in Big Data" (Japec et al. 2015, 854). For an exception see Edjlali and Friedman (2011). We hope this chapter will provide a good starting point to conduct more research on how Big Data and surveys can safely and validly integrate.

Challenges and New Skills Needed for the Survey Researcher Working with Big Data

Gathering, analyzing, and interpreting Big Data requires technical expertise not traditionally gained from survey or social science research training. These may include database skills (NoSQL, relational DBMS), programming skills for mass data processing (e.g., MapReduce), data visualization expertise, as well as analytical techniques not commonly taught to students dealing with survey data (e.g., random forests). Foster et al. (2016) provides a timely discussion about this topic. Even among those who are proficient with Big Data applications, there might exist differing interests and strengths, such as the type A (analysis) versus type B (pipeline building) distinction of data scientists discussed by Chang (2015). Importantly, it will not be feasible to become proficient in all new tools and skills. Instead, a winning strategy is to collaborate with others who have different expertise and strengths.

Technical skills aside, when looking at Big Data as potentially providing substantive answers to what have been studied with surveys, two classes come to mind: Google Trends and social media listening tools. Google Trends (Stephens-Davidowitz and Varian 2015) provides an index of search activities by keywords or categories as well as of interest on these keywords or categories over time. There are numerous examples of using Google Trends to forecast and approximate trends estimated from surveys. Choi and Varian (2012), for instance, show how Google Trends matches the survey-based Australian Consumer Confidence index and Scott and Varian (2015) reproduce the same results for the University of Michigan Consumer Confidence Index time series. Chamberlin (2010) explores Google trends correlations with U.K. Office of National Statistics official data on retail sales, property transactions, car registrations, and foreign trips. At the same time Google Trends does not answer more specific survey questions, such as demographic analysis (e.g., are female consumers more worried about the economy than male consumers?) or modeling questions (what are the drivers of the consumer sentiment in a particular country?).

Social media listening and monitoring tools perform two main tasks: locate social media content from a variety of social media sources and perform automated analysis (content analysis) of the text collected. These tools vary in the depth, range, and historical reach of the content aggregated. When it comes to content analysis, the most common classification of text is as positive, neutral, and negative. In order to do so, social media listening tools use different dictionaries to classify text (see González-Bailón and Paltoglou 2015). Another common usage of social media in the context of surveys is to use it as a a supplement to or replacement of pre-election polls. For example, the percent share of Twitter traffic messages mentioning the six political parties in the 2009 German election was very close to the actual election results (Tumasjan et al. 2010). Using social media tools to replace pre-election polls is not always successful as discussed in Jungherr et al. (2016). There are still many challenges from a methodological and technical point of view to be taken into account and researched.

Changes in the Survey Landscape

All the aforementioned tools and new types of data are making some wonder if surveys are eventually going to disappear because they will be replaced by Big Data. This is true, to some extent. Examples include Censuses in countries such as Finland and other Nordics countries (Statistics Finland 2004) being replaced by administrative data. Other countries go beyond the Census and use administrative data for other social statistics data collection, for example, the Netherlands (Bakker et al. 2014).

Two proponents of the rapid disappearance of surveys, Ray Poynter (2014) and Reginald Baker (2016), use ESOMAR Global Market Research (e.g., ESOMAR. 2015) reports over time to show a decline in the percent of budget spent by market research companies on surveys. For example, in comparison to 2013, the combined percent of online, telephone, face-to-face, and mail survey declined by 6 percent as compared to 2014 (ESOMAR. 2015, 20). The same report is also showing an increased trend of money spent in Automated/Digital and electronic data collection. This category refers to retail audits and media measurement. In other words, market research companies are investing more and more money in Big Data.

Although we do agree with the trend analysis of the ESOMAR reports, we disagree with the implications. The ESOMAR reports capture what is spent by market research companies around the world by contacting country market research associations. The same ESOMAR reports show a change in data collection methods moving more and more to online and smartphone surveys at the expense of other traditional data collection methods such as telephone and face-to-face surveys. What the report cannot capture are two other trends that show increased usage of:

- Do-it-yourself (DIY) web survey platforms
- In-house web survey tools

In the first case (DIY), companies such as SurveyMonkey reported generating 90 million survey completes per month worldwide (Bort 2015). This is not a small number. Qualtrics, another DIY survey tool, distributes one billion surveys annually (personal communication, February 11, 2016). Both survey platforms have major companies as clients in their portfolio.

In the second case, organizations are using in-house web surveys tools, without the need to outsource data collection to market research companies. For example, Google collects customer feedback at scale for all its products using probability-based intercept surveys called Happiness Tracking Survey (Müller and Sedley 2014). Other technology companies has followed suit (Martin 2016).

To summarize, we believe that the real trend in survey-based social and market research is the following:

- From offline data collection methods to web surveys
- From web surveys to mobile web surveys

- From outsourced market research to in-house market research using DIY web survey platforms
- From outsourced market research to in-house market research fully integrated with internal systems

How Surveys and Big Data Can Work Together

Answering the What and the Why

The most commonly shared view among researchers is that surveys and Big Data can and should be used together to maximize the value of each. This is, not surprisingly, one of the takeaways from the American Association for Public Opinion Research task force report on Big Data (Japec et al. 2015).

Looking ahead, the ideal case is to build on the strengths of both data collection methods. Big Data can measure behaviors and tell us the "what" while surveys can measure attitudes and opinions and tell us the "why." A good example of this view comes from a recent Facebook blog post written by two software engineers (Zhang and Chen 2016). The blog post explains the process Facebook used to redesign their News Feed.

> The goal of News Feed is to show you the stories that matter most to you. The actions people take on Facebook – liking, clicking, commenting or sharing a post – are historically some of the main factors considered to determine what to show at the top of your News Feed. But these factors don't always tell us the whole story of what is most meaningful to you. As part of our ongoing effort to improve News Feed, we ask over a thousand people to rate their experience every day and tell us how we can improve the content they see when they check Facebook – we call this our Feed Quality Panel. We also survey tens of thousands of people around the world each day to learn more about how well we're ranking each person's feed.

Surveys Are Just One of a Number of Tools

Sometimes market and survey researchers become so involved in surveys that they forget that surveys are not the *only* tool available to answer research questions (Couper 2013). An illustration can be found when looking at the level of precisions some surveys strive for when asking behavioral questions.

Despite the incredible advances in questionnaire design in past 50 years, asking behavioral questions is and will always be difficult because the answers rely on people's memories. For example, the U.S. Consumer Expenditure Survey used to ask the following questions about clothing purchases: *Did you purchase any pants, jeans, or shorts*? If the respondent said yes, a series of ancillary questions were asked such as: *Describe the item. Was this purchased for someone inside or outside of your household? For whom was this purchased? Enter name of person for whom it was purchased. Enter age/sex categories that apply to the purchase; How many did you purchase?; Enter number of identical items purchased; When did you purchase it?; How much did it cost?; Did this include sales tax?* (Dillman and House 2013, 84).

These questions were repeated for a series of items purchased in the reference month. As the reader can see, the question wording is stretching the limit of the survey tool by asking respondents to remember things with a level of precision that the human memory (in an interview setting) is not very well suited for (see also Eckman et al. 2014 for a discussion on this question wording). In fact, this specific question was an object of a redesign as the committee in charge described it: "This questionnaire structure creates […] *cognitive challenges"* (italics added) (Dillman and House 2013, 84).

We do not envision all behavioral questions being replaced by Big Data collection, but many could be, and there is already some work going in this direction (Sturgis, this volume). For example, Mastrandrea et al. (2015) compared diaries and surveys to wearable sensors and online social media to study social interactions among students in a high school in France. In another application of wearable sensors, Hitachi collected more than a million days' worth of data on employees' activities over the span of 9 years (Yano et al. 2015). The authors were able to correlate the sensor data with happiness measured via questionnaires.

Strengths and Challenges of Surveys and Big Data

Surveys have the advantage of being designed for the researchers to answer the question at hand. They also collect attitudes and opinion data which cannot be readily covered by Big Data. Challenges of surveys are encapsulated in the model of TSEs, and also by the size and coverage of survey data that, unlike few examples (census), are not meant to measure each single member of a particular population.

The most obvious advantage of Big Data collection is that it allows larger sample sizes that support more detailed analysis regarding space, time, and other subgroups. Automated data is also better for measuring certain behaviors (e.g., avoiding recall bias), reducing respondent burden, and avoiding nonresponse bias in some settings. In addition, it can improve turnaround time and facilitate serendipitous findings about variables that no one thought to measure.

On the other hand, Big Data has important challenges. Researchers generally cannot choose what data are collected, or how to gather it. Second, much of Big Data generated is proprietary. Third, the availability of Big Data changes over time. For example, access to Twitter and Facebook changed over time concerning what could be downloaded, who could do it, and the extent of the data over a time period. Finally, as discussed before, Big Data come with Big Noise.

Privacy, Confidentiality, and Transfer of Data

Survey and market researchers have a long tradition and tools in place to handle collection, storage, and processing of survey data in order to guarantee the anonymity and confidentiality of the respondents (ESOMAR 2016; American Statistical Association 2016). Big Data are introducing new questions regarding the collection, storage, and transfer of personal information (Bander et al. 2016). For example, survey research organizations such as the University of Michigan Survey Research Center commonly use multiple databases to augment and enrich telephone and address based samples (Benson and Hubbard 2016). A more powerful example is what political campaigns can do by combining multiple databases and other sources starting from the voter registration databases that exist in each U.S. state. Already in 2008:

> Barack Obama's campaign began the year of his reelection fairly confident it knew the names of every one of the 69,456,897 Americans whose votes had put him in the White House. The votes may have been cast via secret ballot, but because Obama's analysts had come up with individual-level predictions, they could look at the Democrat's vote totals in each precinct and identify the people most likely to have backed him. (Issenberg 2012)

Four years later the Obama campaign created *Narwhal*, a software program that combined and merged data collected from multiple databases and financial sources. The Obama campaign began with a 10 Terabyte database that grew to 50 Terabytes by the end of the campaign (Nickerson and Rogers 2014). This accumulation of data using a census-like approach has privacy advocates worried describing it as "the largest unregulated assemblage of personal data in contemporary American life" (Rubinstein 2014, 861; also Bennett 2013 for an international view).

If we think about the IOT or just about our smartphones, the implications for collecting, storing, and processing personal information are huge. For example, wearable activity bands and smart watches store a large amount of health and personal information that are transferred to apps and stored by the companies producing the devices. Questions such as: who owns these data, what happens to them when a company goes bust or is being acquired by another company? Do not have an easy answer. Privacy, ethical, and legal requirements for collecting, storing, and analyzing Big Data are questions that are here to stay (Lane et al. 2014).

Looking at the Future of Big Data and Surveys

The contemporary social researcher needs to look at Big Data as another source for insights together with surveys and other data collection methods. Unfortunately, at the time of this writing, there is little training available in survey and market research about Big Data. There are however some signs of growth such as the creation of the International Program in Survey and Data Science.[1]

What survey and market researchers can bring to the table is our ability to understand the research questions in greater depth. In the future, the answers to many research questions will not always come only from a survey or some qualitative data collection, but will be increasingly augmented and in some cases replaced by Big Data. The other main strengths that survey and market researchers can bring to the table are the concepts of TSE and BDTE. Shedding light on the limitations and challenges inherent in each data source is key to understanding a phenomenon and validly interpreting research findings. As we stated at the beginning of this chapter, Big Data does not necessarily imply high quality, and surveys can be used to check the quality of Big Data and vice versa.

We encourage survey researchers and practitioners to move the conversation from Big Data to *Rich Data*. We propose the term, rich data, to emphasize the importance of a mindset that focuses on not the mere size of data but their substance and utility. "Big" is never the end goal for research data collection. In fact, Big Data, when thoughtlessly collected and used, may lead to losses in both accuracy and efficiency (Poeppelman et al. 2013). Richness in the data,

[1] http://survey-data-science.net/

on the other hand, captures our methodological aims, namely to enhance and ensure the validity of the research conclusions and inferences as well as the utility of their applications. Specifically, richness means

- a comprehensive coverage of the constructs relevant to a research program.
- the inclusion of multiple complementary indicators that enable accurate and efficient quantification of the target constructs and their relationships.
- the application of appropriate tools to extract information from data, derive defensible and useful insights, and communicate them in compelling fashion.

The new and enhanced data sources and technologies discussed earlier in this chapter provide unprecedented opportunities for researchers and practitioners to improve the richness of their data – through tapping into hard to capture or previously not understood constructs, integrating a multitude of diverse signals (surveys, behavioral data, social media entries, etc.), and leveraging new analytic and visualization tools.

Examples include

- Using high-quality surveys to validate the quality of Big Data sources. This is the case of using surveys to validate the accuracy of voter registration records as reported by Berent et al. (2016).
- Using Big Data to ask better questions in surveys. Big Data can be used as validation data (true value) and different question wording can be tested to determine what is closer to the "true value." The idea is to extend the traditional validation data used in many medical studies such as physicians or nurse tests (e.g., Kenny Gibson et al. 2014) with validation data collected from wearables, or other IOT devices at scale.
- Augment Big Data with survey data such as the Google Local Guides.[2] This opt-in program asks its users to answer few "Yes, No, Not Sure" questions about locations such as restaurants, stores, or point of interest. For example, users can be asked if the restaurant they just visited is family friendly, or has Wi-Fi.

Big Data has opened the door for rich data. It is now time to move beyond the fixation on the size of data and take a more critical view of the new tools and opportunities to advance the science of measuring and influencing human thoughts, emotions, and actions.

[2] https://www.google.com/local/guides/

Acknowledgements We would like to thank Robert Bell (Google), Frauke Kreuter (University of Maryland) and David Vannette (Qualtrics) for their generous comments on the initial draft chapter.

References and Further Reading

American Statistical Association. 2016. "Committee on Privacy and Confidentiality." ASA. http://community.amstat.org/cpc/home.

Baker, Reginald P. 2017. "Big Data: A Survey Research Perspective." In *Total Survey Error: Improving Quality in the Era of Big Data*, edited by Paul P. Biemer, Edith De Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West, 47-70. Hoboken, NJ: Wiley.

Bakker, Bart F. M., Johan Van Rooijen, and Leo Van Toor. 2014. "The System of Social Statistical Datasets of Statistics Netherlands: An Integral Approach to the Production of Register-Based Social Statistics." *Statistical Journal of the IAOS* 30(4): 411–24. doi:10.3233/SJI-140803.

Bander, Stefan, Ron S. Jarmin, Frauke Kreuter, and Julia Lane. 2016. "Privacy and Confidentiality." In *Big Data and Social Science: A Practical Guide to Methods and Tools*, edited by Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, 299–311. Boca Raton, FL: CRC Press.

Bennett, Colin. 2013. "The Politics and the Privacy of Politics: Parties, Elections and Voter Surveillance in Western Democracies." *First Monday* 18(8). doi:10.5210/fm.v18i8.4789.

Benson, Grant, and Frost Hubbard. 2017. "Big Data Serving Survey Research: Experiences at the University of Michigan Survey Research Center." In *Total Survey Error: Improving Quality in the Era of Big Data*, edited by Paul P. Biemer, Edith De Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West, 478-486. Hoboken, NJ: Wiley.

Berent, Matthew K., Jon A. Krosnick, and Arthur Lupia. 2016. "Measuring Voter Registration and Turnout in Surveys. Do Official Government Records Yield More Accurate Assessments?." *Public Opinion Quarterly*, advance access. doi:10.1093/poq/nfw021.

Beyer, Mark A., and Douglas Laney. 2012. *The Importance of "Big Data": A Definition*. G00235055. Stamford, CT: Gartner.

Biemer, Paul P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74(5): 817–48. doi:10.1093/poq/nfq058.

Biemer, Paul P. 2014. "Dropping the 's' from TSE: Applying the Paradigm to Big Data." Paper presented at the 2014 International Total Survey Error Workshop

(ITSEW 2014), Washington, DC: National Institute of Statistical Science. https://www.niss.org/sites/default/files/biemer_ITSEW2014_Presentation.pdf.

Biemer, Paul P. 2016. "Errors and Inference." In *Big Data and Social Science: A Practical Guide to Methods and Tools*, edited by Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, 265–97. Boca Raton, FL: CRC Press.

Bort, Julie. 2015. "How ditching law school and quitting a bunch of good jobs led Dave Goldberg to tech fame and fortune – Business Insider." April 19. http://uk.businessinsider.com/the-incredible-career-of-david-goldberg-2015-4.

Callegaro, Mario. 2013. "Paradata in Web Surveys." In *Improving Surveys with Paradata: Analytic Use of Process Information*, edited by Frauke Kreuter, 261–79. Hoboken, NJ: Wiley.

Chamberlin, Graeme. 2010. "Googling the Present." *Economic & Labour Market Review* 4(12): 59–95. doi:10.1057/elmr.2010.166.

Chang, Robert. 2015. "Doing Data Science at Twitter: A Reflection of My Two Year Journey So Far. Sample Size N = 1." *Medium*. June 20. https://medium.com/@rchang/my-two-year-journey-as-a-data-scientist-at-twitter-f0c13298aee6#.t9wiz09mt.

Choi, Hyunyoung, and Hal Varian. 2012. "Predicting the Present with Google Trends." *Economic Record* 88(S1): 2–9. doi:10.1111/j.1475-4932.2012.00809.x.

Committee on Commerce, Science and Transportation. 2013. "A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes." United States Senate. http://educationnewyork.com/files/rockefeller_databroker.pdf.

Couper, Mick P. 2013. "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys." *Survey Research Methods* 7(3): 145–56. doi:http://dx.doi.org/10.18148/srm/2013.v7i3.5751.

Deming, Edward W. 1944. "On Errors in Surveys." *American Sociological Review* 9(4): 359–69.

Dillman, Don A., and Carol C. House. eds. 2013. *Measuring What We Spend: Toward a New Consumer Expenditure Survey. Panel on Redesigning the BLS Consumer Expenditure Surveys*. Washington, DC: National Academies Press.

Dutcher, Jennifer. 2014. "What Is Big Data? – Blog." September 3. https://datascience.berkeley.edu/what-is-big-data/.

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser. 2014. "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys." *Public Opinion Quarterly* 78(3): 721–33. doi:10.1093/poq/nfu030.

Edjlali, Roxanne, and Ted Friedman. 2011. Data Quality for Big Data: Principles Remain, But Tactics Change. G00224661. Stanford, CT: Gartner.

ESOMAR 2016. "ESOMAR Data Protection Checklist." ESOMAR. https://www.esomar.org/knowledge-and-standards/research-resources/data-protection-checklist.php.

ESOMAR. 2015. "Global Market Research 2015." ESOMAR.

Ferguson, Mike. 2014. "Big Data – Why Transaction Data Is Mission Critical to Success." Intelligence Business Strategies Limited. https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14442usen/IML14442USEN.PDF.

Foster, Ian, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane. eds. 2016. *Big Data and Social Science: A Practical Guide to Methods and Tools*. Boca Raton, FL: CRC Press.

Gershenfeld, Neil, Raffi Krikorain, and Danny Choen. 2004. "The Internet of Things." *Scientific American* 291(4): 76–81. doi:10.1038/scientificamerican1004-76.

González-Bailón, Sandra, and Georgios Paltoglou. 2015. "Signals of Public Opinion in Online Communication. A Comparison of Methods and Data Sources." *The ANNALS of the American Academy of Political and Social Science* 659(1): 95–107. doi:10.1177/0002716215569192.

Gray, Emily, Will Jennings, Stephen Farrall, and Colin Hay. 2015. "Small Big Data: Using Multiple Data-Sets to Explore Unfolding Social and Economic Change." *Big Data & Society* 2(1). doi:10.1177/2053951715589418.

Groves, Robert M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75(5): 861–71. doi:10.1093/poq/nfr057.

Hewson, Claire, Carl Vogel, and Dianna Laurent. 2016. *Internet Research Methods*. 2nd ed. London: Sage.

Hsinchun, Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *Mis Quarterly* 36(4): 1165–88.

Issenberg, Sasha. 2012. "How Obama's Team Used Big Data to Rally Voters. How President Obama's Campaign Used Big Data to Rally Individual Voters." *MIT Technology Review*. https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/.

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul P. Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher. 2015. "Big Data in Survey Research. AAPOR Task Force Report." *Public Opinion Quarterly* 79(4): 839–80. doi:10.1093/poq/nfv039.

Jungherr, Andreas, Harald Schoen, Oliver Posegga, and Pascal Jürgens. 2016. "Digital Trace Data in the Study of Public Opinion an Indicator of Attention Toward Politics Rather Than Political Support." *Social Science Computer Review* doi:10.1177/0894439316631043.

Kenny, Gibson, William, Hilary Cronin, Rose Anne Kenny, and Annalisa Setti. 2014. "Validation of the Self-Reported Hearing Questions in the Irish Longitudinal Study on Ageing Against the Whispered Voice Test." *BMC Research Notes* 7(361). doi:10.1186/1756-0500-7-361.

Kreuter, Frauke, and Roger D. Peng. 2014. "Extracting Information from Big Data: Issues of Measurement, Inference and Linkage." In *Privacy, Big Data, and the Public Good. Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Benden, and Helen Nissenbaum, 257–75. New York: Cambridge University Press.

Lane, Julia, Victoria Stodden, Stefan Benden, and Helen Nissenbaum. eds. 2014. *Privacy, Big Data, and the Public Good*. New York: Cambridge University Press.

Martin, Jolie M. 2016. “Combining ‘small Data’ from Surveys and ‘big Data’ from Online Experiments at Pinterest.” In *ALLDATA* 2016: The Second International Conference on Big Data, Small Data, Linked Data and Open Data (includes KESA *2016)*, edited by Venkat Gudivada, Dumitru Roman, Pia Di Buono Maria, and Mario Monteleone, 33–34. Lisbon: IARA. http://toc.proceedings.com/29767webtoc.pdf.

Mastrandrea, Rossana, Julie Fournet, and Alain Barrat. 2015. “Contact Patterns in a High School: A Comparison Between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys.” *PloS One* 10(9): e0136497. doi:10.1371/journal.pone.0136497.

Müller, Hendrick, and Aaron Sedley. 2014. “HaTS: Large-Scale in-Product Measurement of User Attitudes & Experiences with Happiness Tracking Surveys.” In *Proceedings of the 26th Australian Computer-Human Interaction Conference (OzCHI 2014)*, 308–15. New York, NY: ACM.

Nickerson, David W., and Todd Rogers. 2014. “Political Campaigns and Big Data.” *Journal of Economic Perspectives* 28(2): 51–74. doi:10.1257/jep.28.2.51.

Poeppelman, Tiffany, Nikki Blacksmith, and Yongwei Yang. 2013. “‘Big Data’ Technologies: Problem or Solution?.” *The Industrial-Organizational Psychologist* 51(2): 119–26.

Poynter, Ray. 2014. “No More Surveys in 16 Years? NewMR.” August 27. http://newmr.org/blog/no-more-surveys-in-16-years/.

Rubinstein, Ira S. 2014. “Voter Privacy in the Age of Big Data.” *Wisconsin Law Review* 2014(5): 861–936.

Schober, Michael F., Josh Pasek, Lauren Guggenheim, Cliff Lampe, and Frederick G. Conrad. 2016. “Social Media Analyses for Social Measurement.” *Public Opinion Quarterly* 80(1): 180–211. doi:10.1093/poq/nfv048.

Scott, Steve, and Hal R. Varian. 2015. “Bayesian Variable Selection for Nowcasting Economic Time Series.” In *Economic Analysis of the Digital Economy*, edited by Avi Goldfarb, Shane M. Greenstein, and Katherine E. Tucker, 119–35. Chicago, IL: Chicago University Press.

Statistics Finland. 2004. *Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland*. Helsinki: Statistics Finland.

Stephens-Davidowitz, Seth, and Hal Varian. 2015. “A Hands-on Guide to Google Data.” http://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf.

Swan, Melanie. 2013. “The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery.” *Big Data* 1(2): 85–99. doi:10.1089/big.2012.0002.

Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. “Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment.” In *Fourth International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441.

Waldherr, Annie, Daniel Maier, Peter Miltner, and Enrico Günther. 2016. “Big Data, Big Noise. The Challenge of Finding Issue Networks on the Web.” *Social Science Computer Review*, advance access. doi:10.1177/0894439316643050.

Wallgren, Andrew, and Britt Wallgren. 2014. *Register-Based Statistics: Statistical Methods for Administrative Data*. 2nd ed. Chichester, UK: Wiley.

Warshaw, Christopher. 2016. “The Application of Big Data in Surveys to the Study of Public Opinion, Elections, and Representation.” In *Computational Social Science. Discovery and Prediction*, edited by R. Michael Alvarez, 27–50. New York: Cambridge University Press.

Yano, Kazuo, Tomoaki Akitomi, Koji Ara, Junichiro Watanabe, Satomi Tsuji, Nabuo Sato, Miki Hayakawa, and Norihko Moriwaki. 2015. “Measuring Happiness Using Wearable Technology. Technology for Boosting Productivity in Knowledge Work and Service Businesses.” *Hitachi Review* 64(8): 517–24.

Zhang, Chen, and Si Chen. 2016. “News Feed FYI: Using Qualitative Feedback to Show Relevant Stories.” February 1. http://newsroom.fb.com/news/2016/02/news-feed-fyi-using-qualitative-feedback-to-show-relevant-stories/.

Mario Callegaro is Senior Scientist at Google, London, in the User Research and Insights team, Brand Studio. He focuses on measuring brand perception and users’ feedback. Mario consults on numerous survey and market research projects.

Mario holds a MS and a PhD in Survey Research and Methodology from the University of Nebraska, Lincoln.

Prior to joining Google, Mario was working as survey research scientist for Gfk-Knowledge Networks. He is associate editor of *Survey Research Methods* and in the advisory board of the *International Journal of Market Research.*

Mario has published numerous books, book chapters, and presented at international conferences on survey methodology and data collection methods.

He published (May 2014) an edited book with Wiley titled *Online Panel Research: A Data Quality Perspective,* and his new book coauthored with Katja Lozar Manfreda and Vasja Vehovar: *Web Survey Methodology* is available from Sage as of June of 2015.

Yongwei Yang is a Research Scientist at Google. He works on brand and user research, as well as general research on measurement and survey methodology. Yongwei enjoys figuring out how to collect better data and make better use of data, as well as to implement evidence-based interventions and evaluate their business impact. His research interests include survey and test development and validation, technology-enhanced measurement and data collection, survey and testing in multi-population settings, concepts and models for understanding consumer and employee behaviors, and utility analysis of organizational interventions. He holds a PhD in Quantitative and Psychometric Methods from the University of Nebraska-Lincoln.

24

Getting the Most Out of Paradata

Frauke Kreuter

Paradata is a term used to describe all types of data about the process and context of survey data collection. The term “paradata” was first used in a talk by Mick Couper to describe automatically generated process data such as data from computerized interviewing software (Couper 1998). Examples of paradata include

- Listing information:
  - Day, time, edits
- Keystrokes:
  - Response times, back-ups, edits
- Vocal characteristics:
  - Pitch, disfluencies, pauses, rate of speech
- Contact data:
  - Day, time, outcomes

F. Kreuter (✉)
University of Maryland, College Park, USA
Mannheim University, Mannheim, Germany
Institute for Employment Research, Mannheim, Germany
e-mail: fkreuter@umd.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_24

- Interviewer observations:
  - Sample unit characteristics

Conceptually, these are important data because they allow insights into many sources of total survey error. Some of the most common uses of paradata to examine survey errors include using keystrokes to evaluate measurement error and data validity, or using contact data and observations to examine nonresponse error and adjustment error. These data are distinct from metadata, which are a class of data about the characteristics of the data and data capture systems or "data about the data." Examples of metadata include technical reports, survey instruments, interviewer instructions, show cards, and other documentation of the survey process or variables.

Given the relatively recent discovery of the uses and benefits of paradata there is not a widely accepted set of best practices for how and when to use all of the different types of information collected as paradata. However, the existing literature does provide considerable information about some specific types of paradata. One of the most commonly used and studied forms of paradata is response time to a question. Current uses of response times tend to be post hoc and focused on response error. For example, examining the characteristics of the survey instrument and setting (Bassili & Scott 1996; Draisma & Dijkstra 2004; Tourangeau, Couper & Conrad 2004; Yan & Tourangeau 2008). Factors that tend to increase response times are poor wording, poor instrument layout, question length, and question complexity. Factors that tend to decrease response times are logical ordering of questions, respondent practice at survey completion, and decreasing motivation on the part of the respondent. Response times have also been used to evaluate interview administration and associated errors (Olson & Peytchev 2007; Couper & Kreuter 2012). In the extreme case, response times have been used to identify interviewer falsification, when the interview was completed in less time that it would have taken to read each question, much less hear and record the response (Johnson et al. 2001; Penne, Snodgrass & Baker 2002). Some novel applications of response times have examined the potential for concurrent use to trigger interventions in self-administered web surveys if respondents answer too fast or slow (Conrad, Schober & Coiner 2007; Conrad et al., 2017).

Call record data are another form of paradata that have received considerable attention and research. These data have been used to focus on improving efficiency through identifying optimal schedules for interviewers to reach respondents (Weeks et al. 1980; Greenberg & Stokes, 1990; Bates 2003;

Laflamme 2008; Durrant et al. 2011). These data have also been used for identifying potentially important predictors of response. For example, examining the number of contact attempts it took to reach a certain person, when was that person contacted last time, and what is the probability for that person to be at home, or not at home, the next day or the next time you try to reach that particular respondent (Campanelli et al. 1997; Groves & Couper 1996; Bates & Piani, 2005). Call records have also been used to examine error features such as nonresponse bias analyses and nonresponse bias adjustment (Politz & Simmons 1949; Kalton, 1983; Beaumont, 2005; Biemer & Link 2007), and increasingly for applications of adaptive and responsive design (Groves and Heeringa 2006, Wagner 2013, for an overview see Tourangeau et al. 2017). Quite extensive research has been done in more recent years on the use of interviewer observations, both in responsive design but also for nonresponse adjustment (West et al. 2014; Krueger & West 2014).

However, despite having examined certain aspects and uses of paradata extensively there are still important areas for future research. For example, there is still very little research in developing systematic approaches for handling keystroke data. This may be due in part to the particularly messy nature of these data, but rather than stymying research altogether this should be taken as an opportunity for interdisciplinary research that integrates text analysis and survey research to examine keystroke data systematically. Similarly, there has not been sufficient research on examining face-to-face contact protocol data generated by interviewers after each contact attempt with a household. This lack of research is likely due to the amount of missing data and the very complex hierarchical structure of the data, but this is again simply another opportunity for additional interdisciplinary research to model these data, identify problematic areas and ways to make them more consistent in the future. These types of findings could then easily link back to ways to improve field practice and inform supervision.

Looking beyond the current problems with what is known about uses for paradata, it will be important for future research to examine a couple of important general themes. First, research is needed on ways to enhance real-time use of paradata and the development of best practices around concurrent use of paradata during data collection. Second, research is needed on identifying additional forms of paradata that are available across different modes; there may be additional paradata that could be collected but the opportunities simply haven't been identified or exploited yet. Third, paradata-driven indicators of survey quality (i.e., process quality) need to be explored and developed. Other issues that warrant further discussion and development of best practices include the potential for requiring paradata in data management plans submitted to funding agencies and potential confidentiality issues in the release of paradata.

Areas for future research:

- Expanding the use of keystroke data
  - Development of open-access code repositories
- Development of appropriate statistical methods for handling face-to-face contact protocol data
- Development of better applications for interviewers to use for contact record data
- Identify approaches to enhance real-time use of paradata during surveys
- Identify new forms of paradata in different modes
- Development of paradata-driven indicators of survey quality
- Identify potential confidentiality issues around the release and use of paradata

References and Further Reading

Bassili, J. N., & Scott, B. S. (1996). Response Latency as a Signal to Question Problems in Survey Research. *Public Opinion Quarterly*, *60*(3), 390–399.

Bates, N. A. (2003). *Contact Histories in Personal Visit Surveys: the Survey of Income and* Program Participation (SIPP) Methods Panel. *Demographic Surveys Division*.

Bates, N. A., & Piani, A. (2005). *Participation in the National Health Interview Survey: Exploring reasons for reluctance using contact history process data*.

Beaumont, J.-F. (2005). Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(3), 445–458.

Biemer, P. P., & Link, M. W. (2007). Evaluating and Modeling Early Cooperator Effects in RDD Surveys. In Advances in Telephone Survey Methodology (pp. 587–617). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9780470173404.ch26

Campanelli, P., Sturgis, P., & Purdon, S. (1997). Can you hear me knocking? and investigation into the impact of interviewers on survey response rates.

Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing features of human dialogue to web surveys. *Applied Cognitive Psychology*, *21*(2), 165–187. http://doi.org/10.1002/acp.1335

Conrad, F., Tourangeau, R., Couper, M. P., & Zhang C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, 11(1), 45–61.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association.*

Couper, M. P., & Kreuter, F. (2012). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series a (Statistics in Society)*, *176*(1), 271–286. http://doi.org/10.1111/j.1467-985X.2012.01041.x

Draisma, S., & Dijkstra, W. (2004). Response latency and (para) linguistic expressions as indicators of response error. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: John Wiley & Sons.

Durrant, G. B., D'Arrigo, J., & Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society: Series a (Statistics in Society)*, *174*(4), 1029–1049. http://doi.org/10.1111/j.1467-985X.2011.00715.x

Greenberg, B. S., & Stokes, S. L. (1990). Developing an optimal call scheduling strategy for a telephone survey. *Journal of Official Statistics*, 6(4), 421–435.

Groves, R. M., & Couper, M. P. (1996). Contact-level influences on cooperation in face-to-face surveys. *Journal of Official Statistics*, 12 (1), 63–83.

Groves R. M., Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 169(3), 439–457.

Johnson, T. P., Parker, V., & Clements, C. (2001). Detection and prevention of data falsification in survey research. *Survey Research*, 32(3), 1–2.

Kalton, G. (1983). Models in the Practice of Survey Sampling. *International Statistical Review / Revue Internationale De Statistique*, *51*(2), 175–188.

Krueger, B. S., West, B. T. (2014). Assessing the Potential of Paradata and Other Auxiliary Data for Nonresponse Adjustments. *Public Opinion Quarterly,* 78(4), 795-831

Laflamme, F. (2008). Understanding Survey Data Collection Through the Use of Paradata at Statistics Canada. *Proceedings of the American Statistical Association.*

Olson, K., & Peytchev, A. (2007). Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes. *Public Opinion Quarterly*, *71*(2), 273–286. Retrieved from http://poq.oxfordjournals.org.ezproxy.stanford.edu/content/71/2/273.short

Penne, M. A., Snodgrass, J. J., & Barker, P. (2002). Analyzing audit trails in the National Survey on Drug Use and Health (NSDUH): means for maintaining and improving data quality. Presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods, Charleston, SC.

Politz, A., & Simmons, W. (1949). An Attempt to Get the "Not at Homes" Into the Sample without Callbacks. *Journal of the American Statistical Association*, *44*(245), 9–16.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, *68*(3), 368–393.

Yan, T. & Tourangeau, R. (2008). Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology,* 22(1), 51–68.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, 7, 45–55.

Weeks, M. F., Jones, B. L., Folsom, R. E., Jr, & Benrud, C. H. (1980). Optimal Times to Contact Sample Households. *Public Opinion Quarterly*, *44*(1), 101–114.

West, B.T., Kreuter, F., & Trappmann, M. (2014). Is the Collection of Interviewer Observations Worthwhile in an Economic Panel Survey? New Evidence from the German Labor Market and Social Security (PASS) Study. *Journal of Survey Statistics and Methodology*, 2(2), 159–181.

Frauke Kreuter is Director of the Joint Program in Survey Methodology at the University of Maryland; Professor of Statistics and Methodology at the University of Mannheim, Germany; and head of the Statistical Methods Research Department at the Institute for Employment Research (IAB) Nuremberg, Germany. Her prior appointments were with the Ludwig-Maximilians University of Munich and the University of California, Los Angeles. Frauke Kreuter is a Fellow of the American Statistical Association and recipient of the Gertrude Cox Award. Her recent textbooks include *Data Analysis Using Stata* (Stata Press); *Big Data and Social Science* (CRC Press); and *Practical Tools for Designing and Weighting Survey Samples* (Springer). Her Massive Open Online Course on Questionnaire Design has taught roughly 100,000 students. Currently, she is building an International Program on Survey and Data Science with funding from the German government's Open University program.

25

Open Probability-Based Panel Infrastructures

Marcel Das, Arie Kapteyn and Michael Bosnjak

Introduction

Although probability-based survey panels that collect high-quality, representative data either solely or partly through online questionnaires have been around for a couple of decades (see e.g. Hays et al. 2015) they are still relatively rare. Examples are the CentERpanel and the Longitudinal Internet Studies for the Social sciences (LISS) Panel in the Netherlands; the German Internet Panel and the GESIS Panel in Germany; ELIPSS in France; GfK Knowledge Panel, Pew American Trends Panel Survey, American Life Panel, and Understanding America Study (UAS) in the United States.

Yet, with the advent of more probability-based Internet panels in different countries new opportunities for cross-national and cross-cultural research are opening up. Internet panels are a natural environment for including new forms of data collection, such as data capture from wearables,

M. Das (✉)
Tilburg University, Tilburg, The Netherlands
e-mail: das@uvt.nl

A. Kapteyn
University of Southern California, Los Angeles, USA
e-mail: kapteyn@usc.edu

M. Bosnjak
ZPID - Leibniz Institute for Psychology Information, Trier, Germany
e-mail: michael.bosnjak@zpid.de

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_25

experimentation on population representative panels, and for quick turnaround data collection. Although cross-cultural research has a long history, probability-based Internet panels can provide considerable flexibility in the timing and design of new data collection across countries. As with any kind of data collection, it is essential to be as transparent as possible about every step of the research process and share data as widely as possible.

Here we will illustrate the potential for international comparative research by discussing three collaborating probability-based Internet panels, namely LISS, the GESIS panel, and UAS. These are taken as examples of how cross-country Internet panels can support new and exciting research, but with an eye on expansion to as many panels and countries as possible. We will discuss the advantages, challenges, and suggest topics for future research.

While the specific recruitment and implementation procedures of the three panels differ, to address country-specific requirements and restrictions, they all share three common characteristics: *openness* in terms of being accessible for academic researchers from any substantive area to field primary studies and to use the data collected; *probability-based* and therefore optimized for yielding unbiased population estimates in the respective countries; and *transparency* in terms of the processes by which these infrastructures have been built and are being operated. In addition, the data collection process and its deliverables are transparent, facilitating the replicability of processes and outcomes.

This chapter provides a brief overview of (1) the type of primary research facilitated by the three open probability-based panels, and (2) the data provided to the scientific community for secondary analysis purposes. Next we discuss opportunities for expanding internationally comparative research within the framework of an international alliance of open probability-based panels.

Three Examples of Open Probability-Based Panels

The LISS Panel,[1] maintained by CentERdata at Tilburg University (The Netherlands), has been operational since 2007. The panel consists of 5,000 households and is representative of the Dutch-speaking population. Panel members answer monthly interviews over the Internet (for about 30 minutes in total). The LISS Panel is based on a probability sample drawn from the

[1] www.lissdata.nl

population register, in close collaboration with Statistics Netherlands. Sampled households were contacted either face-to-face or by phone. Households without Internet who are willing to join the panel are given a (easy-to-handle) personal computer and broadband access. More details can be found in Scherpenzeel and Das (2011).

Questionnaires for cross-sectional and longitudinal studies as well as experiments can be proposed by any interested researcher. The core questionnaire, designed with assistance from international experts in the relevant fields, contains questions on topics such as health, work, income, education, ethnicity, political opinion, values, norms, and personality. Designed to follow changes over the life course of individuals and households, it is repeated annually. The average household attrition rate has been about 10 percent per year. Refreshment samples were added in 2009, 2011, and 2014.

The GESIS Panel,[2] located at GESIS – Leibniz-Institute for the Social Sciences in Mannheim (Germany), has been operational since 2014 and is a panel of German-speaking individuals aged between 18 and 70 years old (at the time of recruitment), permanently residing in Germany. By the end of the recruiting phase in February 2014 the GESIS Panel consisted of almost 5,000 panel members. The panel is a probability-based mixed-mode infrastructure. Members complete an omnibus survey bi-monthly (for about 20 minutes). The GESIS Panel is based on a sample clustered in randomly drawn communities from the municipal population register. Individuals are randomly selected within the drawn communities. The bi-monthly waves are collected using two self-administered survey modes: online and by mail. After the setup of the panel approximately two-thirds of the members participate online and one-third participates by mail. From 2016 onwards, refreshment samples will be added. Detailed information about the setup of the GESIS Panel can be found in Bosnjak et al. (in press).

Similar to the LISS Panel a longitudinal core study is run in the GESIS Panel. Within each wave about 5 minutes are blocked for the modules in the core study. The remainder of the time that is available in the particular wave is reserved for external researchers. Proposals can be submitted throughout the year.

[2] www.gesis-panel.org

More recently, the UAS,[3] maintained by the Center for Economic and Social Research at the University of Southern California (USC), was set up as a household panel representing the 18+ population in the United States. Currently, the panel comprises approximately 6,000 households. Similar to LISS, USC provides equipment to those households who do not have access to the Internet at the time of recruitment. These households receive a tablet and a broadband Internet subscription. Panel members are recruited using address-based sampling. Recruitment takes place in replicates (batches). Zip-codes are drawn randomly, after which a vendor of postal addresses draws approximately 40 addresses randomly from within the zip-codes. The first few batches of zip-codes were simple random draws from the universe of zip-codes. Later draws take into account the existing demographic and socio-economic distribution of the existing sample. The probability of selecting a zip-code is determined by an algorithm that takes into account the differences between the socio-economic and demographic distributions of the existing panel and the population. Zip-codes that may help to bring the sample composition more in line with the population distribution have a higher probability of being selected.

Questionnaires and experiments in the UAS are fielded in two languages: English and Spanish. Any researcher can use the infrastructure. Once a survey is ready, respondents are notified via email that a survey is waiting for them and sent a link to their personalized panel page so that they may begin the particular survey. UAS also enables researchers to leverage a rich collection of core data, including data from the Health and Retirement Study (HRS), which is administered bi-annually to all panel members; cognitive capability and numeracy measures; financial literacy; subjective well-being; and personality (big five).

Primary Research Conducted in Open Probability-Based Panels

All three panels mentioned in the previous section have collected (and are still collecting) a vast amount of data. The data infrastructures are used to collect cross-sectional and longitudinal data. Repeated measurements can be taken annually, as is the case in the longitudinal core studies in LISS and

[3] https://uasdata.usc.edu

GESIS, or bi-annually as with the HRS instrument that is administered in the UAS. Shorter time spans are also possible, for example, several months or even weeks or days. The panels are ideally suited to run experimental studies. Instead of using small convenience samples of university students, one can run experiments with large heterogeneous samples in a very cost-efficient way. Moreover, the Internet mode allows for inclusion of pictures, movies, and audio.

The panels are also useful in exploiting technological advances in data-collection techniques such as the collection of non-reactive data. An example of the collection of non-reactive data is the use of accelerometry. Both LISS and UAS have administered questionnaires about physical activity in combination with respondents wearing accelerometers for a period of a week. When regressing both self-reports and objective measures of physical activity on a number of demographic and socio-economic variables it is found that self-reports and objective measures of physical activity tell a strikingly different story about differences between the Netherlands and the United States: At the same level of self-reported activity, the Dutch are significantly more physically active than the Americans (Kapteyn et al. 2016). It appears, in other words, that the Dutch and Americans have significantly different standards as to what counts as physical activity. This is of great importance as physical activity is an important determinant of health and until now comparisons across countries purporting to understand determinants of obesity for instance had to rely on self-reports.

Data Access and Secondary Research

All data collected in the three panels are available to any interested researcher, free of charge, and can be used for secondary research with opportunities to combine variables from different fields of study, which are usually not collected within one specific project. The data are also an excellent source for educational purposes and for student projects. In addition, much attention is paid to documenting the process of data collection and metadata to make sure that all data collected is accessible, understandable, and useful to researchers.

LISS data are made available via the LISS Data Archive (www.lissdata.nl/dataarchive). Data collected in the GESIS Panel are available at www.gesis-panel.org and UAS data can be accessed at https://uasdata.usc.edu/surveys.

The three panel infrastructures invest in opportunities to link the data collected in the panels to supplementary data sources. LISS data can be linked to administrative data from Dutch national statistical registers. These include health status, the use of health-care facilities, wealth, pensions, income, and mortality. In the GESIS Panel, the extended data file accessible exclusively via the GESIS Secure Data Center can in principle be linked via geographic variables to environmental data sources (e.g., noise information, pollution).

A different source of supplementary data collected in the UAS is the result of an NSF-funded project, in which UAS asks panel members for permission to collect their electronic financial transactions (e.g., debit and credit card transactions, online banking) through the use of an intermediary (a financial aggregator). This information is then linked to other information collected from the respondents directly through surveys.

An International Alliance of Panels

Each of the three probability-based panels offers unique possibilities for academic research. Their use is not limited to researchers from the country where the infrastructure is located. Twenty percent of the projects conducted in the LISS panel were initiated by researchers from outside the Netherlands. When combined in a network the opportunities extend to conducting multinational, multiregional, and multicultural research. The description of the physical activity measurement project across LISS and UAS mentioned earlier would be an example.

In 2016 the three probability-based panels described in this chapter initiated the Open Probability-based Panel Alliance (OPPA).[4] The Alliance facilitates cross-cultural survey research with probability-based Internet panels across the globe, endorsing joint methodological standards to yield representative data. OPPA is driven by research demand. Primary researchers are provided a one-stop entry point to submit proposals and then choose whether the data should be collected in all countries participating in the network, or in subsets of it. As an open network, OPPA aims to include additional panel infrastructures from around the globe. The partnership model is light: minimum standards need to apply to join the alliance.

[4] http://www.openpanelalliance.org

Essential conditions are transparency on all fronts, a probability-based panel and openness to any discipline or research team that wants to collect data or use collected data for secondary analyses. Joint methodological and substantive research (proposals) are encouraged, but are not a condition to join the network.

Current and Future Developments

The increasing prominence of probability-based Internet panels and attempts to forge alliances across countries foreshadows a number of exciting new opportunities for research in the social sciences. In 2009, LISS ran a feasibility study on using self-administered tests – collection of blood cholesterol, saliva cortisol, and waist measurement – to gather biomarkers (Avendano et al. 2011). A selection of LISS panel members also received an advanced bathroom scale with a wireless Internet connection; their weight and body fat measurements were then transmitted to the database without any respondent intervention. First empirical analyses were based on almost 80,000 measurements collected in 2011 (Kooreman and Scherpenzeel 2014). A more general discussion about the use of biomarkers in representative surveys and leave-behind measurement supplements can be found in Chapter 29 (by David Weir) and Chapter 27 (by Michael W. Link) of this volume.

Another new development is the use of smartphones; they are increasingly becoming platforms for collecting data about behavior, either actively (by sending questions or prompts asking for action) or passively (by connecting to monitoring devices worn by respondents). See also the Sturgis and Lessof chapter in this volume on new measurement tools. This opens possibilities for more frequent and yet less burdensome data collection. For example, in LISS data were collected on time use (Scherpenzeel and Fernee 2013) and travel behavior using GPS-enabled smartphones (Geurs et al. 2015). One can interview respondents frequently over the Internet about a broad range of topics, while asking them to wear devices measuring physical activity, sleep, biometrics (e.g., heart rate variability, blood pressure, skin conductance to measure stress; e.g., Picard et al. 2016), and social interactions (e.g., Mehl et al. 2012). Additionally, researchers can apply so-called burst designs, whereby during brief periods (e.g., a week) very rich data is collected both by asking questions and from wearables. Bursts may be scheduled at regular intervals (e.g., once a year) or may be triggered by events such as changes in work, health, family composition, or residence. During these short bursts, data may be collected (both from measurement devices and Internet interviews) at very high frequencies (e.g., at the end of every day or through brief questions several times a day, while worn devices record outcomes continuously). See also Chapter 28 by Arthur A. Stone in this volume on experience sampling and ecological momentary assessment.

As with any collected survey data, links to other sources of information may essentially enrich the information collected. With the consent of respondents, one may collect information on respondents' location, their financial behavior (as in the project described before), and merge in contextual data, such as information on weather, pollution, noise, traffic patterns, etc.

Challenges

Primary research conducted in a panel may also have disadvantages. Respondents in a panel become experienced respondents after a while; hence their response to questions may differ systematically from the response of individuals who are not experienced respondents. Toepoel et al. (2008) found little evidence that survey experience influences the question-answering process. A significant panel conditioning effect, however, is reported on (repeated) knowledge questions (Das et al. 2011; Toepoel et al. 2009) and household-saving behavior (Crossley et al. 2016). Surveys that focus on issues such as social interaction and social exclusion may also give biased population estimates when fielded in a panel since, in general, the more socially engaged individuals are more likely to participate in panels. This may also hold for cross-sectional surveys but active participation in a panel requires more effort of the individual.

The most obvious challenge in performing simultaneous research in different countries lies in the adequate translation of questionnaires or experimental instructions. To enhance the reliability and validity of survey data by minimizing undesired culture-driven perception shifts as well as language-driven meaning shifts, OPPA has partnered with a translation agency having a proven track record in ensuring translation accuracy in large-scale international surveys. As the example of accelerometry across the United States and the Netherlands illustrates there may also be large differences in the use of response scales. However this challenge is also a research opportunity.

Areas for future research:

- Internet penetration varies across countries. Although a probability-based Internet panel needs to cover the whole population (either by using a mixed mode approach or by supplying Internet access to respondents without prior access), the difference in coverage may introduce differences in estimates across countries. The comparative advantages of using mixed mode or providing Internet access to all respondents needs to be studied both from the viewpoint of operational efficiency and from the viewpoint of maximizing representativeness.

- Countries do not only differ in Internet penetration, but also in how the Internet is accessed and used. Increasingly, mobile devices are an important part of the technology used by respondents. Understanding the implications of these "mixed devices" differences and how to optimize data collection across platforms merits more and ongoing research.
- Partly the notion of joining an alliance of similar infrastructures is to establish best practices. There are many dimensions affecting the quality of Internet panels, including recruiting procedures and panel retention. This provides ample opportunities for learning from each other, but also to conduct experiments where, for instance, the practice in one country is tested experimentally in a different country.
- As noted, response scales and language peculiarities may affect comparability of responses across countries; this is a ripe area for research.

References and Further Reading

Avendano, M., Scherpenzeel, A., Mackenbach, J. P. (2011). Can Biomarkers be Collected in an Internet Survey? A Pilot Study in the LISS Panel. In: *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (eds. Das, M, P. Ester, & L. Kaczmirek), New York: Taylor & Francis, 77–104.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K.W. (in press). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review*, http://dx.doi.org/10.1177/0894439317697949

Crossley, T. F., De Bresser, J., Delaney, L., & Winter, J. (2016). Can Survey Participation Alter Household Saving behaviour? *Economic Journal*, http://onlinelibrary.wiley.com/doi/10.1111/ecoj.12398/abstract.

Das, M., Toepoel, V., & Van Soest, A. (2011). Non-parametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys, *Sociological Methods and Research*, 40(1), 32–56.

Geurs, K. T., Thomas, T., Bijlsma, M., & Douhou, D. (2015). Automatic Trip and Mode Detection with Move Smarter: First Results from the Dutch Mobile Mobility Panel, *Transportation Research Procedia*, 11, 247–262.

Hays, R. D., Liu, H., & Kapteyn, A. (2015). Use of Internet Panels to Conduct Surveys, *Behavior Research Methods*, 47, 685–690.

Kapteyn, A., Saw, H.-W., Banks, J., Hamer, M., Koster, A., Smith, J. P., Steptoe, A., & Van Soest, A. (2016). What They Say and What They Do: Comparing Physical Activity Across U.S., England, and the Netherlands, Working Paper, Center for Economic and Social Research, University of Southern California.

Kooreman, P., & Scherpenzeel, A. (2014). High Frequency Body Mass Measurement, Feedback, and Health Behaviors, *Economics & Human Biology*, 14, 141–153.

Mehl, M. R., Robbins, M. L., & Deters, F. G. (2012). Naturalistic Observation of Health-Relevant Social Processes: The Electronically Activated Recorder Methodology in Psychosomatics, *Psychosomatic Medicine*, 74, 410–417.

Picard, R. W., Fedor, S., & Ayzenberg, Y. (2016). Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry, *Emotion Review*, 8, 62–75.

Scherpenzeel, A., & Das, M. (2011). True Longitudinal and Probability-Based Internet Panels: Evidence from the Netherlands. In: *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (eds. Das, M, P. Ester & L. Kaczmirek), New York: Taylor & Francis, 77–104.

Scherpenzeel, A., & Fernee, H. (2013). New and Emerging Methods. The Smartphone in Survey Research, Experiments for Time Use Data, *The Survey Statistician*, 67, 19–25.

Toepoel, V., Das, M., & Van Soest, A. (2009). Relating Question Type to Panel Conditioning: A Comparison Between Trained and Fresh Respondents, *Survey Research Methods*, 3(2), 73–80.

Toepoel, V., Das, M., & Van Soest, A. (2008). Design Effects in Web Surveys: Comparing Trained and Fresh Respondents, *Public Opinion Quarterly*, 72(5), 985–1007.

Marcel Das holds a PhD in Economics from Tilburg University, The Netherlands (1998). In 2000, he became the director of CentERdata, a survey research institute specialized in web-based surveys and applied economic research. As a director of CentERdata he has managed a large number of national and international research projects. He is one of the principal investigators of the Dutch MESS project for which CentERdata received major funding from the Dutch Government. Since February 2009, Das is Professor of Econometrics and Data collection at the Department of Econometrics and Operations Research of the Tilburg School of Economics and Management at Tilburg University. He has published a number of scientific publications in international peer reviewed journals in the field of statistical and empirical analysis of survey data and methodological issues in web-based (panel) surveys.

Arie Kapteyn PhD, is a Professor of Economics and the Executive Director of the Center for Economic and Social Research (CESR) at the University of Southern California. Before founding CESR at USC in 2013, Prof. Kapteyn was a Senior Economist and Director of the Labor & Population division of the RAND Corporation.He has about 20 years of experience in recruiting and managing population representative Internet panels, including the CentERpanel (2,000 respondents; the first probability Internet panel in the world) and LISS panel (7,500 respondents, role co-PI) in the Netherlands, as well as the American Life Panel (6,000 respondents) and the Understanding America Study (6,000 respondents) in the United States. He has conducted numerous experiments with the panels, concerning methods (e.g., optimal recruiting and survey design), substantive studies (including health and decision-making), and measurement (self-administered biomarkers, physical activity). He has been involved in telephone and in-person surveys on various continents.

Michael Bosnjak is director of ZPID - Leibniz Institute for Psychology Information in Trier, Germany, and Professor of Psychology at the University of Trier. Before joining ZPID in July 2017, he was team leader for the area Survey Operations at GESIS - Leibniz Institute for the Social Sciences in Mannheim, Germany, and Full Professor for Evidence-Based Survey Methodology at the University of Mannheim, School of Social Sciences. Between 2013 and 2016, he was the founding team leader of the GESIS Panel, a probabilistic mixed-mode omnibus panel for the social sciences. His research interests include research synthesis methods, survey methodology, and consumer psychology.

26

Collecting Interviewer Observations to Augment Survey Data

Brady T. West

Interviewer observations are an exciting form of paradata that have recently received more focus among survey researchers. There are two general categories of interviewer observations: the first are observations recorded by survey interviewers for all sampled units that describe selected features of the sampled units, including attempts at recruitment, neighborhood descriptions, and similar observations. The second type of observations are those recorded by survey interviewers for respondents that describe aspects of the survey interview. Observations like: *Did the respondent understand what the survey questions were trying to get at? Did the respondent seem to take enough time? Did they run into cognitive challenges with the actual survey?* Both types of observations can be thought of as observational paradata, or data that describe the process of collecting survey data, that are not automatically generated but rather observed and recorded by the interviewer.

Several large and important surveys already make use of interviewer observations. For example, the Los Angeles Family and Neighborhood Survey collects interviewer observations on the sampled area for variables such as evidence of crime or social disorder. The National Survey of Family Growth (NSFG) collects data on the sampled household, in particular for key household features (e.g., presence of young children) that may not get

B.T. West (✉)
University of Michigan, Ann Arbor, USA
e-mail: bwest@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_26

reported on rosters for households that refuse screening interviews. The National Health Interview Survey (NHIS) has started to collect housing unit observations related to the health conditions of inhabitants, such as the presence of wheelchair ramps or cigarette butts. The NSFG and Panel Study of Income Dynamics both use interviewer observations on the survey respondent for variables such as interviewer opinions of the quality of data provided by the respondent.

There are a number of reasons why interviewer observations are important to collect and study. First, they are an inexpensive and potentially useful source of auxiliary information on all sample units, and this has been demonstrated in face-to-face surveys. Future research should also examine the utility of collecting these data for telephone surveys as well (e.g., interviewer judgments of respondent gender). The second reason that they are important is that prior research has demonstrated that interviewer observations can be correlated with both response propensity and key survey variables, making them useful for nonresponse adjustments and for responsive survey designs (e.g., case prioritization).

However, there may be some drawbacks to attempting to collect and use interviewer observations. For example, not all interviewers view them as easy to collect or worthy of their time and effort. Further, the existing literature on the subject has shown that these observations can be error-prone and are frequently missing. There is also a concern about asking interviewers to essentially become weak proxy respondents for nonresponding households or individuals so that researchers can hopefully learn something about nonrespondents. If the interviewers aren't incentivized to take this task seriously because they would rather focus on completing interviews, then they might not take the time necessary to record high quality information. Lastly, some types of observations (e.g., post-survey) take large amounts of time for interviewers to record, begging the question: do the benefits of the observations outweigh the costs?

Research on interviewer observations has allowed the development of a few best practices. Examples include the recommendation that every observation that interviewers are asked to record should have a specific purpose, such as

- Nonresponse adjustment
- Prediction of response propensity
- Profiling of active cases for possible targeting
- Assessment of data quality

Collecting observations with no *a priori* purpose is likely to be a waste of interviewer time and researcher money. Furthermore, observations collected on all sample units should be correlated with key variables and/or response propensity. This suggests that one opportunity for future research is to focus on identifying and understanding these associations. Once these associations are empirically established, then the observation requests to interviewers should ideally be designed as proxies for key measures in the survey so that they can act as a form of validation or data quality check. This would drastically increase the benefit that researchers get from interviewer observations.

Once interviewer observations have been collected, it is important that the survey organization *actually analyzes* the observation data. Even if the data are of questionable quality, there may be important information that could point toward operational improvements that the organization can make. More specifically, these observations may point to problems with the questionnaire in general or indicate data quality issues that should be addressed. As part of this point that observations should be analyzed, when possible, the quality of observations should be assessed using validation data. The easiest validation data might be the actual survey reports of the same phenomena observed, but there may be opportunities to validate using administrative records or even observations by another interviewer.

Methods for standardizing the ways in which observations are collected is another key area where best practices have been developed and implemented but further research may be warranted. For example, concrete training approaches such as using visual examples of how to make effective observations should be implemented as part of interviewer training. An example of this can be found in the European Social Survey, and the NHIS has started to employ this kind of training. Interviewers should also be provided with known and observable predictors of features that they are being asked to observe; this approach has been implemented by the NSFG. Lastly, interviewers may be asked to provide open-ended justifications for why they record a particular value for an observation; this is also an approach that has been used by the NSFG.

However, there are several areas where future research is needed before additional best practices can be defined. First, more work is needed to assess the validity and accuracy of interviewer observations across a variety of different face-to-face surveys. Second, future research is also needed to identify correlates and drivers of observation accuracy, such as features of interviewers, respondents, or areas that can predict interviewer observation accuracy or inaccuracy, or alternative strategies used by interviewers that can lead to higher accuracy.

Third, future research needs to examine what interviewer observations add to existing auxiliary variables such as those in commercial databases.

It is important to know if the observations explain additional variance in survey outcomes on top of existing variables. Fourth, the statistical and operational impacts of observation quality on responsive survey design approaches should be examined to determine if decisions based on the observations (e.g., targeting certain cases) backfire at a lower threshold of quality or if the observations serve to improve efficiency regardless of quality.

Fifth, additional research is needed to evaluate the ways that error-prone interviewer observations affect statistical adjustments for survey nonresponse; only weighting adjustments have been studied thus far, meaning that there is a need for further research. Sixth, research is needed on effective design strategies for improving the quality of interviewer observations; for example, does providing interviewers with known auxiliary predictors of features improve the accuracy of their observations or not?

Seventh, work is needed to understand how post-survey observations might be used to improve survey estimates. For example, observations could be used to design calibration estimators or as indicators of data quality to inform researchers about which cases to consider dropping due to poor response quality. Eighth, more work is needed to identify and understand the sources of interviewer variance in observations and observation quality. For example, qualitative studies of different strategies used to record observations in the field and interviews with interviewers to determine what their actual practices are. Lastly, and perhaps most importantly, it will be important for future research to identify and measure the empirical trade-offs between the costs of collecting interviewer observations versus improvements in survey estimates from collecting the observations.

In summary, these are important areas for future research on interviewer observations:

- Assessing the validity and accuracy of interviewer observations
- Identifying correlates and drivers of observation accuracy
- Determining the added value of observations over other existing auxiliary variables from commercial databases
- Examining the statistical and operational impacts of varying observation quality
- Understanding the effects of potentially error-prone observations on statistical adjustments for nonresponse
- Designing optimal strategies for improving the quality of interviewer observations
- Identifying ways to use post-survey observations to calibrate weights or give indications of data quality

- Understanding sources of variation in observations and observation quality between interviewers
- Identifying and measuring trade-offs between costs and benefits of collecting interviewer observations

Brady T. West is a Research Associate Professor in the Survey Methodology Program, located within the Survey Research Center at the Institute for Social Research on the University of Michigan-Ann Arbor (U-M) campus. He also serves as a Statistical Consultant on the U-M Consulting for Statistics, Computing, and Analytics Research (CSCAR) team. He earned his PhD from the Michigan Program in Survey Methodology in 2011. Before that, he received an MA in Applied Statistics from the U-M Statistics Department in 2002, being recognized as an Outstanding First-year Applied Masters student, and a BS in Statistics with Highest Honors and Highest Distinction from the U-M Statistics Department in 2001. His current research interests include the implications of measurement error in auxiliary variables and survey paradata for survey estimation, survey nonresponse, interviewer variance, and multilevel regression models for clustered and longitudinal data. He is the lead author of a book comparing different statistical software packages in terms of their mixed-effects modeling procedures (*Linear Mixed Models: A Practical Guide using Statistical Software, Second Edition*, Chapman Hall/CRC Press, 2014), and he is a co-author of a second book entitled *Applied Survey Data Analysis* (with Steven Heeringa and Pat Berglund), which was published by Chapman Hall in April 2010 and has a second edition in press that will be available in mid-2017. Brady lives in Dexter, MI with his wife Laura, his son Carter, his daughter Everleigh, and his American Cocker Spaniel Bailey.

27

"Just One More Thing": Using Leave-Behind Measurement Supplements to Augment Survey Data Collection

Michael W. Link

Leave-behind measurement supplements, as the name implies, are surveys or survey-related tools left with respondents after an interview has been completed. These tools are frequently used by research organizations but have not been the subject of much empirical research. Much is known about the components of leave-behind materials but very little about the methodology and best practices, and not much has been published on this subject despite the widespread use of the approach.

Leave-behind measures have a number of defining characteristics. First, they almost always involve self-administration. Second, the data collection mode is often different than the initial mode meaning that when they are used leave-behinds are often components of a multi-mode design. The type of data that these measures provide is typically supplemental for a study but on rare occasions they have been used to collect the primary data. Another unique feature is that the leave-behind task is completed after the end of an initial survey or interview, meaning that this is a multi-stage (not simply multi-mode) approach to data collection. Leave-behinds are nearly exclusively implemented by large surveys that often involve complex data collection efforts, rarely are they part of data collection with smaller or more straightforward studies. Leave-behinds may take many forms though

M.W. Link (✉)
Data Science, Surveys & Enabling Technologies Division, Abt Associates, Cambridge, MA, USA
e-mail: Linkmi01@gmail.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_27

additional surveys, diaries, electronic monitors, and physical specimen collection devices are most common. Diaries are perhaps the most popular of the leave-behind methods in survey research and The Nielsen Company's "people meters" which are used to generate television ratings may be the most prominent example of the method being applied and used as the primary source of data collection. Not included in this definition of leave-behind measures are traditional mail or panel surveys, self-initiated online surveys, or ACASI segments conducted during a larger face-to-face interview.

There are many different data collection purposes that leave-behinds are suited for. At the most basic level, they are useful for collecting more information to expand the initial data collection effort, but they are also useful for reducing respondent burden by allowing them to complete a portion of the data collection on their own schedule. They are also very well suited for sensitive topics or when privacy concerns may be an issue, in these contexts they may lead to less social desirability and higher-quality data. For some types of data collection using a leave-behind may provide an opportunity to improve data quality by reducing the need for respondents to recall things such as daily activities that are more accurately captured by use of an activity diary left with the respondent after the initial interview.

While these methods have been widely applied in practice there is an apparent lack of empirical research on the methodology from which best practices might be derived. When leave-behinds do appear in the literature they tend to be discussed as a component of a broader study and not the specific focus. This may be because leave-behinds are perceived to be adjunct data and not the primary focus of the study, thus they are not explicitly examined or the findings regarding the leave-behinds themselves are not reported widely. This presents an important opportunity for future research to examine fundamental questions about how these methods are being used, how effective they are, and how they can be improved.

One important and exciting area for future research is identifying ways that new technologies can enable new and improved methods of leave-behind measurements. Mobile platforms, apps for tablet devices or smartphones, and other new technologies offer innovative approaches to this suite of methodologies that may expand their applicability and utility for surveys. One of the appeals of incorporating these new technologies, as part of the leave-behind approach, is that they could be used to facilitate quick and easy communication with respondents rather than expensive and potentially intrusive face-to-face or telephone contact methods. These technologies could enable respondents to be prompted to enter data or to upload their data for researchers to analyze. Furthermore, GPS-enabled devices have become ubiquitous and, if respondents consent to being tracked using their smartphone or another device, this could allow incredibly rich validation data to be collected in conjunction with time-use diaries and other self-report methods. Bluetooth-enabled devices could similarly revolutionize data collection as they enable researchers to passively and semi-passively capture a wide array of respondent activities, for example, it is possible to passively record blood glucose, blood oxygen, and pulse data passively using Bluetooth devices. Other novel opportunities for collecting data using new technologies include image or video collection, audio capture, and text entry.

Beyond exploring the promise of new technologies, additional research is needed at a more fundamental methodological level. For example, what lessons can be drawn from focusing on leave-behinds as a distinct methodological approach? Can generalizations be made across different approaches or are the techniques used in practice too varied to allow for comparison? How do data quality concerns around leave-behinds differ from "primary" modes of data collection? Do leave-behinds actually reduce respondent burden? What respondent compliance concerns are associated with leave-behinds? Is satisficing behavior influenced by leave-behinds? Specifically regarding data quality issues, do timing and context change responses that might have been obtained otherwise? Are data collected with leave-behinds comparable with other forms of data?

Areas for future research:

- Applications of new technologies to improve and expand leave-behind measurement
- Examinations of leave-behind measurement tools as distinct methodological approach
  - Generalizing across different techniques
  - Data quality concerns unique to the methodology
  - Impacts on respondent burden
  - Issues surrounding respondent compliance
  - Effects on data quality in terms of measurement error when compared with other methodologies

Further Reading

Ashbrook, D., & Starner, T. (2002). Learning significant locations and predicting user movement with GPS (pp. 101–108). Presented at the Sixth International Symposium on Wearable Computers (ISWC 2002), IEEE. http://doi.org/10.1109/ISWC.2002.1167224

Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing* 7(5): 275–286.

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C.,* 17(3): 285–297.

Krenn, P. J., Titze, S., Oja, P., Jones, A., & Ogilvie, D. (2011). Use of Global Positioning Systems to Study Physical Activity and the Environment. *American Journal of Preventive Medicine*, *41*(5), 508–515. http://doi.org/10.1016/j.amepre.2011.06.046

Ohmori, N., Nakazato, M., Harata, N., Sasaki, K., & Nishii, K. (2006). Activity diary surveys using GPS mobile phones and PDA. Presented at the Transportation Review Board Annual Meeting.

Michael W. Link, Ph.D. is Division Vice President of the Data Science, Surveys & Enabling Technologies (DSET) Division at Abt Associates, a leading global providers of policy-based research and evaluation for government, academic, and commercial clients. He is also a past President of the American Association for Public Opinion Research, 2014–2015. Dr. Link's research efforts focus on developing methodologies for confronting the most pressing issues facing measurement and data science, including use of new technologies such as mobile platforms, social media, and other forms of Big Data for understanding public attitudes and behaviors. Along with several colleagues, he received the American Association for Public Opinion Research 2011 Mitofsky Innovator's Award for his research on address-based sampling. His numerous research articles have appeared in leading scientific journals, such as *Public Opinion Quarterly, International Journal of Public Opinion Research,* and *Journal of Official Statistics*.

28

Ecological Momentary Assessment in Survey Research

Arthur A. Stone

Ecological momentary assessment (EMA (Shiffman, Stone, & Hufford, 2008; Stone & Shiffman, 1994); also known as Experience Sampling (Csikszentmihalyi & Larson, 1987)) is a family of measurement techniques where in the content of survey research respondents, who have previously been recruited to be part of a panel, are contacted according to a predetermined schedule and asked to report information about their current state. This may include questions about the psychological state of the respondent, their experiences, behaviors, symptoms, features of their environment, and even physiological metrics. The goal of these methods is to capture data on experiences and behavior as precisely as possible. The gains in accuracy stem from experiences and behavior being measured with minimal recall bias or respondent forgetting, enabling researchers to generate high levels of ecological validity or correspondence with what respondents actually experienced.

EMA methods allow researchers to study time usage in more depth and with greater precision than many other approaches. Using these approaches researchers are able to study patterns of experience or behavior, particularly those that fluctuate rapidly like emotions or symptoms that are harder to recall later. EMA enables research on within-day contemporaneous and lagged associations between experiences and behaviors, for example, how

A.A. Stone (✉)
Department of Psychology, Dornsife Center for Self-Report Science,
University of Southern California, California, USA
e-mail: arthur.stone@usc.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_28

does having a cigarette craving associate with the behavior of having a cigarette? These methods have also been used to link experiential and behavioral data with real-time physiological data such as blood glucose levels, EKG, EEG, and other cardiovascular measures.

Conceptually, EMA methods are an important class of measurements because they allow access to information that can be challenging to capture using other approaches given how how and where they are stored in memory. Immediate information is stored in experiential memory, which captures information about what respondents are experiencing at the moment. This type of memory tends to be very short-lived, making it hard to access later such as when a more traditional survey might ask about the experience. Episodic memory captures memories of experiences whereas semantic memory stores a different kind of information, typically content like attitudes and beliefs about experiences. When a recall period expands from "*How are you doing right now?*" to "*How have you done over a day or over a week or over a month?*" there is a shift in the kind of memory that respondents access in order to create the answer and as the memory period increases because the recall period increases, respondents' answers tend to shift to semantic memory and beliefs. So in the case of EMA, rather than attempting to measure beliefs about things that happened, the goal is to measure the experience itself.

Researchers who apply EMA methods are typically concerned about distortion associated with "long" recall periods, which may include periods as short as a week or even a day. There are a variety of cognitive heuristics that the brain uses in order to summarize information in ways that may reduce response accuracy over longer recall periods. A key point is that respondents are not aware heuristics when they activated. For example, a well-known heuristic is the "peak-end" rule, which describes people's tendency to remember peaks of experience (salient moments) and things that are relatively proximal to when they are completing the questionnaire. A similar heuristic appears when respondents report current levels either as a proxy for the recall period asked about or to alter their response by reporting the past experience relative to the current one. Essentially, these issues strongly suggest that asking people about certain kinds of experiences, symptoms, or behaviors over relatively long periods may be fraught with bias. These have been recognized for a long time including by survey researchers (Bradburn et al., 1987 in *Science*). However, EMA methods are just beginning to be used in the broader survey research community.

While ecological validity isn't typically a major concern for survey researchers, there are some particular contexts in which it is an important consideration. For example, in health studies or drug trials it is often desirable to capture data on what actually happened to respondents (symptoms, side effects, etc.) rather than the belief-biased responses that respondents may provide later. The underlying concept is that to understand the experiences that people are actually having researchers need to representatively sample situations from the relevant universe of situations that includes the experiences of interest. The approach then is to contact respondents at a representative set of random intervals and ask them about periods of time immediately preceding the contact.

The features of EMA have a number of benefits and drawbacks that are important for researcher to consider before applying the methods. In terms of real-time data capture, the primary benefits are that recall bias is significantly reduced or eliminated because the sampling period occurs concurrently with or immediately following the experience. It can also allow a window into daily patterns and rhythms of experience for respondents. However, the drawbacks of real-time data capture are that it only captures point estimates rather than more global evaluations, the sampling framework is complex and challenging to implement, important events that occur may often happen outside of the very short window of the recall period, and lastly, the approaches tend to be expensive and burdensome for respondents(Stone et al. 2006a).

More broadly, there are a number of concepts that apply to typical questionnaire research that EMA researchers have identified as best practices for reducing bias due to recall errors on the part of respondents. First, it is typically best practice to limit the recall period to a very short amount of time immediately preceding the contact with the respondent. A second technique is to elicit reconstruction of the recall period by respondents. Third, researchers can use very precise questions to ease the recall process for respondents and similarly researchers can limit queries to information that can easily be recalled, such as salient events. For example, EMA methods are useful for collecting in-depth information about a single day or a portion of a day, but, by definition, have limited value for collecting data over longer recall periods. Furthermore, the burden on respondents participating in EMA studies is sufficiently high that long-term panel participation with high levels of daily measurement is likely to be infeasible. For some surveys this approach may be valuable but others might find it untenable.

The Day Reconstruction Method (DRM) was developed in response to some of these challenges with EMA approaches. The DRM was designed to allow researchers to reproduce the results that would have been achieved by EMA through having respondents systematically reconstruct the recall period (Kahnemann, et al 2004). The American Time Use Survey and the Princeton Affect & Time Use Survey have both implemented this

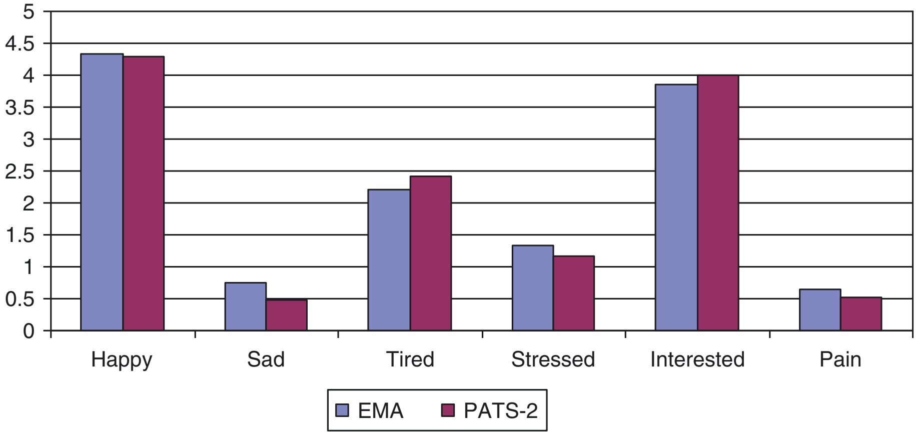


Fig. 28.1 Comparison of data from DRM and EMA

approach successfully. Preliminary research has indicated that the DRM produces substantively similar results to EMA, as seen in Fig. 28.1 (Stone et al. 2006b).

However, the DRM method requires a technologically sophisticated administration approach and it is still time-consuming for respondents, often taking 30–45 minutes. Thus, while the DRM approach is an improvement in terms of its applicability to surveys, future research is needed to create versions of the DRM task that are more amenable to the survey context. For example, versions of the DRM that are implemented over the Internet are now available in addition to paper-and-pencil and interviewer-administered versions.

In summary, there is considerable interest in characterizing daily experiences and behaviors in real time or near-real time. Many subject areas studied with survey data could potentially benefit from these kinds of approaches, but more research is necessary to identify optimal ways to integrate these approaches with large-scale data collection operations. This is not to imply that surveys should move to begin collecting these data, given the logistical and cost challenges with collecting EMA and DRM data; in fact, there should be a clear rationale for attempting to collect detailed daily information in large-scale surveys. Several studies have been successful with alternatives to EMA, which is promising because it makes these types of data more feasible to collect in conjunction with more traditional survey data.

Areas for future research:

- Identifying large-scale surveys that could benefit from EMA or DRM data and conducting feasibility testing
- Determining best practices for implementing EMA or DRM methods in conjunction with traditional survey data collection
- Continued research comparing EMA and DRM methods
- Identifying ways to reduce the time and costs associated with DRM methods
- Identifying novel ways to leverage new technologies to collect these data at lower cost and in larger quantities
- Measuring respondent burden when these methods are applied
- Addressing data consistency and quality concerns

References and Further Reading

Bradburn, N. M., Rips, L. J., & Shevell, S. (1987). Answering Autobiographical Questions: the Impact of Memory and Inference on Surveys. *Science*, *236*(4798), 157–161. http://doi.org/10.1126/science.3563494

Csikszentmihalyi, M., & Larson, R. (1987). Validity and Reliability of the Experience-Sampling Method. *The Journal of Nervous and Mental Disease*, *175*(9), 526.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science*, *306*(5702), 1776–1780. http://doi.org/10.2307/3839780?ref=search-gateway:8cb2fab2e33d2829b6c8d8113cdb2349

Shiffman, S. S., Stone, A. A., & Hufford, M. R. (2008). Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. http://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Stone, A. A., & Shiffman, S. S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, *16*(3), 199–202.

Stone, A. A., Kessler, R. C., & Haythomthwatte, J. A. (2006a). Measuring Daily Events and Experiences: Decisions for the Researcher. *Journal of Personality*, *59*(3), 575–607. http://doi.org/10.1111/j.1467-6494.1991.tb00260.x

Stone, A. A., Schwartz, J. E., Schwarz, N., Schkade, D. A., Krueger, A. B., & Kahneman, D. (2006b). A Population Approach to the Study of Emotion: Diurnal Rhythms of a Working Day Examined with the Day Reconstruction Method. *Emotion*, *6*(1), 139–149.

Arthur A. Stone is Professor of Psychology and Director of the Dornsife Center for Self-Report Science at the University of Southern California. Stone's early work was concerned with improving the measurement of life events and coping with the goal of understanding how events and coping impact our susceptibility to somatic illnesses. These studies led to an interest in psychobiology with a particular emphasis on how environmental events affect biological processes. Concurrently, he was researching how people self-report information about their psychological and symptom states. This led to the development of diaries measuring within-day phenomena, ultimately yielding a set of techniques known as Ecological Momentary Assessment. Stone has been involved with alternative methods for capturing the ebb and flow of daily experience for large-scale surveys, including the development of the Day Reconstruction Method. He is also been involved with the development of questionnaires for use in clinical trials (the PROMIS project).

29

Biomarkers in Representative Population Surveys

David Weir

Biomarkers are a class of measures that are collected via physical specimens provided by respondents. These are typically direct measures of biological states, including disease, physiological functioning, and physical traits such as respondent height and weight. Prior to implementation in health surveys, these types of measures were typically collected in the clinical context and often only using convenience samples rather than representative samples of the population. Now there are at least four large nationally representative surveys currently collecting biomarker data: National Health and Nutrition Examination Survey (NHANES), Add Health, National Social Life, Health, and Aging Project (Williams and McDade, 2009), and the Health and Retirement Study (HRS) (Weir, 2008; Sakshaug, Couper, and Ofstedal 2010). Of these surveys, NHANES is considered the gold standard as it is essentially a study designed to collect biomarker data on large and representative samples of the population.

There are several different types of biomarker measures that are commonly collected, as mentioned earlier. These are typically minimally invasive and range from physical measurements of height and weight to biochemical assays from blood, other fluids, or body parts (McDade, Williams, and Snodgrass, 2007; Sakhi, Bastani, and Ellingjord-Dale 2015). Occasionally advanced imaging technology is involved such as X-ray or MRI, but this is far

D. Weir (✉)
University of Michigan, Ann Arbor, USA
e-mail: dweir@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_29

less common. Lastly, a relatively new measure that is expected to have great utility in the future is DNA samples (Schonlau, et al 2010; Calderwood, Rose, and McArdle, 2013).

Biomarkers are valuable measures to collect for a number of reasons (Freese, Li, and Wade, 2003). First, they provide objective measures of health that are more accurate and less subject to bias than self-report data, they are also able to measure biological traits and states that respondents themselves are often unaware of during the survey. For example, as can be seen in Fig. 29.1, the HRS compared self-reported height with measured height and found that, across ages and gender, measured values were consistently lower than self-reported values.

Second, biomarkers enable researchers to generate descriptive statistics about the health of the population. Third, researchers are able to use direct measures of health as dependent variables to identify important

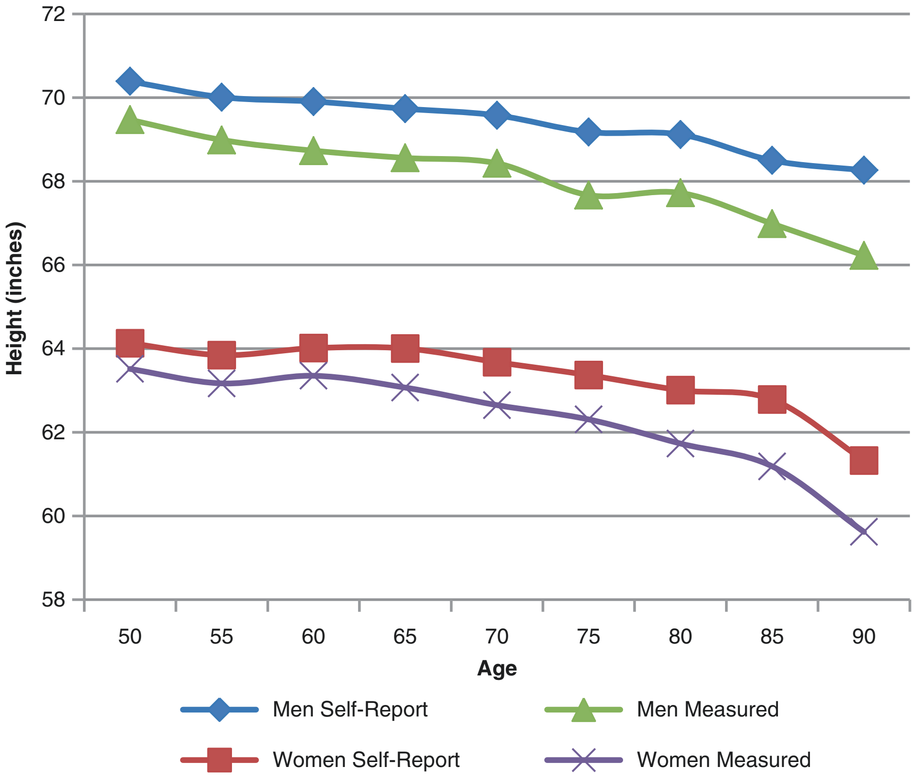


Fig. 29.1 Comparison of self-reported versus measured height by age and gender

predictors of specific health states. For example, in epidemiology there are theories that social economic status affects health outcomes through stress-related pathways, biomarkers of those systems allow researchers to test that mechanism directly. And lastly, economists use objective health metrics as predictors of other important variables such as productivity and economic outcomes.

There are a number of cautionary points and best practices for biomarker collection that have been indicated in prior work. The first is that participation in biomarker collection may be related to the health state that the researcher is trying to measure, meaning that nonresponse bias is a concern for biomarker measures. This can be addressed by pairing physical measures with self-report items because it allows imputation or re-weighting solutions to capture the full range of variation in the biomarker. Second, researchers need to be aware of the effects of applying different cut-points to the data, meaning that slight differences in where cut-points are assigned may have significant substantive effects on the results obtained. For example, body mass index (BMI) is calculated by dividing weight by height-squared. Weight estimates tend to be fairly accurate, but the errors in self-reported height above will be magnified in self-reported BMI estimates. The result is that self-report BMI is about 4 percent low on average with 29.5 percent of respondents being classified as obese. But the mean of BMI is very close to the cutoff for being obese, and measured BMI indicates that 38.2 percent of respondents are obese. This means that the 4 percent error in self-reported BMI translates into a 29 percent increase in the fraction of the population that actually qualify as being obese when BMI is measured using biomarkers. So the cut-off for being classified as obese has significant substantive implications for how both the self-report and biomarker data are interpreted.

The third best practice relates to interviewer training. There is a range of concerns that arise here, from addressing respondent confidentiality concerns to maintaining a sterile environment when biological specimens are being collected and transported. Properly training interviewers to handle the unique demands of collecting biomarkers is important. Preliminary evidence indicates that interviewer uncertainty and unfamiliarity with collecting biomarkers may be associated with lower levels of compliance among respondents. This implies that biomarker collection cannot simply just be added onto an existing survey without investing in significant training efforts to not only teach interviewers how to collect the biomarker data but also to make them comfortable doing it.

It is important to note that biomarker collection has two potentially important impacts on surveys that must be considered before implementing these methods. First, collecting biomarkers is very time-consuming. This drives up respondent burden and means that interviewers are able to collect fewer interviews in the same amount of time. Second, and relatedly, the amount of effort and cost that must be put into actually recruiting respondents and conducting the interview also increases. The good news is that, when the additional effort is invested in recruitment, response rates seem to not be affected by the addition of biomarker data collection. However, there do seem to be significant racial differences in respondent cooperation specifically and uniquely with regard to biomarker data collection, indicating that further research is needed into how to overcome this potential nonresponse bias.

Two recent developments in biomarker collection promise to vastly expand the breadth and depth of biological information that can be captured by health surveys: dried blood spot (DBS) and DNA samples. DBS are easily collected by using a small lancet to prick the respondent's finger and then very small drops of blood are collected on filter paper where it can be stored in dry conditions. In addition to the ease of data collection, DBS enable collection of a blood-based biomarker by regular interviewers, that is, interviewers without the phlebotomy training and certification that is required to draw whole blood from respondents. Storage and handling of DBS is also considerably easier than with whole blood, which requires careful temperature control and rapid processing after collection. DBS also has the advantage of being much cheaper to collect and process than whole blood.

The tradeoff that comes with DBS is that there are a limited number of biological assays that can be used to analyze the DBS, meaning that a smaller range of analyses can be performed relative to whole blood. There are also some concerns about the quality of these measures. This is an area where future research is needed; increasing the range of analyses that can be performed on DBS and improving the quality of the measures will help this method to achieve its promise as a key innovation in biomarker research. Lab validation studies comparing within-subject DBS with whole blood samples, test–retest reliability of DBS assays, and comparisons of population distributions of estimates attained between studies using whole blood and DBS are still needed.

The second recent development in biomarker research that holds great promise for the future is in the area of genetic biomarkers. In recent years, collecting DNA from respondents has become very easy and inexpensive with a number of vendors producing cost-effective and easy to use kits that regular interviewers can use during interviews. Currently analysis costs are still high and the range of useful analyses that can be conducted on DNA is relatively low, however the costs are dropping rapidly and analyses are becoming increasingly useful. These are minimally invasive tests typically only involving a saliva sample, meaning that field implementation is also relatively easy for interviewers. The DNA analysis approaches still require significant future research before the potential of genetic biomarkers as part of survey data collection is fully realized but this is arguably a major component of the future of biomarker research.

Respondent consent, confidentiality, along with notification guidelines and concerns are other important areas requiring future attention. Researchers need to define the ethical framework for how to ensure that respondents are aware of the implications of their consent and how the data will and will not be used. For example, most researchers do not notify respondents if biomarker data indicate disease or health risk factors; this is an area where the ethics of handling these unique types of data have not been fully explored and best practices defined. Confidentiality is also a concern as genetic data and data from other biomarkers can be extremely sensitive and identifiable; continued discussion of best practices for respondent confidentiality is necessary as biomarker data become increasingly powerful and prevalent.

In summary, applications of biomarkers in survey research have only scratched the surface of their potential. The existing methods have demonstrated the feasibility of combining biological data collection with population surveys but further research is needed to develop the tools even further to increase the amount of data collected, the quality of those data, and the utility of those data for analysts. New technologies and methods will reduce costs and continue to drive biomarker research into new promising areas of population health research, but more work is needed to fully maximize the potential of these approaches.

Areas for future research:

- Identifying approaches to reduce the nonresponse bias associated with race in response to biomarker requests
- Methods for improving the quality of biomarker measures
- Improving interviewer training to ensure data quality
- Expanding comparisons of similar methods such as whole blood and DBS to identify areas of inaccuracy

- Testing and implementation of new technologies for collecting and analyzing biomarker data
- Maximizing the potential of newly cost-effective biomarker data such as DNA

References and Further Reading

Calderwood, L., Rose, N., & McArdle, W. (2013). Collecting saliva samples for DNA extraction from children and parents: findings from a pilot study using lay interviewers in the UK. *Survey Methods: Insights From the Field (SMIF).*

Freese, J., Li, J.-C. A., & Wade, L. D. (2003). The Potential Relevances of Biology to Social Inquiry. *Annual Review of Sociology*, *29*, 233–256. http://doi.org/10.2307/30036967?ref=search-gateway:7dfdc81e23833b50858bac7cf6be2fe4

McDade, T. W., Williams, S. R., & Snodgrass, J. J. (2007). What a Drop Can Do: Dried Blood Spots as a Minimally Invasive Method for Integrating Biomarkers into Population-Based Research. *Demography*, *44*(4), 899–925. http://doi.org/10.2307/30053125?ref=search-gateway:50cdd3c9e61d93d73634a76cf5c5fe3e

Sakhi, A. K., Bastani, N. E., & Ellingjord-Dale, M. (2015). Feasibility of self-sampled dried blood spot and saliva samples sent by mail in a population-based study. *BMC Cancer*, *15*(1), 327. http://doi.org/10.1186/s12885-015-1275-0

Sakshaug, J. W., Couper, M. P., & Ofstedal, M. B. (2010). Characteristics of Physical Measurement Consent in a Population-Based Survey of Older Adults. *Medical Care*, *48*(1), 64–71. http://doi.org/10.2307/27798406?ref=search-gateway:25d2fe2091212e7ad549e95e4ee84791

Schonlau, M., Reuter, M., Schupp, J., Montag, C., Weber, B., Dohmen, T., et al. (2010). Collecting Genetic Samples in Population Wide (Panel) Surveys: Feasibility, Nonresponse and Selectivity. *Survey Research Methods*, *4*(2), 121–126.

Weir, D. W. (2008). Elastic Powers: The Integration of Biomarkers into the Health and Retirement Study. In M. Weinstein, J. W. Vaupel, & K. W. Wachter (Eds.), *Biosocial Surveys*. Washington, DC: National Academies Press (US).

Williams, S. R., & McDade, T. W. (2009). The Use of Dried Blood Spot Sampling in the National Social Life, Health, and Aging Project. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *64B*(Supplement 1), i131–i136. http://doi.org/10.1093/geronb/gbn022

David R. Weir is a Research Professor in the Survey Research Center at the Institute for Social Research at the University of Michigan and Director of the Health and Retirement Study (HRS). He received his PhD in Economics from

Stanford University and held faculty positions at Yale and the University of Chicago before returning to Michigan in 1999. He has led the transformation of the HRS into a world-leading biosocial survey combining its traditional excellence as a longitudinal survey with direct biological measures of health, genetics, linked medical and long-term care records from the Medicare system, and enriched psychological measurement. His research increasingly includes comparative analyses from the international family of HRS studies that now cover more than half the world's population.

30

Measuring Past Events Using Calendar Aids

Robert F. Belli

Respondents to surveys often provide retrospective reports of events or behaviors that have high levels of error, indicating that the responses are often not reliable and raising questions about the quality of the data being collected. Survey researchers have responded by developing a number of methods and tools for maximizing the quality of these retrospective reports. These innovations have led to number of best practices being developed but there is still a need for improvement through continued research on developing new methods and improving those currently being used.

In conventional interviewing practice, one important goal is to measure only the variance in respondent reports, and this often means minimizing other exogenous factors that could influence these reports, such as interviewer effects. One important method toward reducing these exogenous effects has been the development of standardization practices so that respondents are all exposed to the same stimuli prior to providing their responses. However, standardization does little to help with autobiographical memory recall, so other methods were developed including calendar interviewing and time use diaries.

Calendar interviewing refers to the use of event history calendars that are used as visual aids to assist respondents with recalling with greater precision when a particular event occurred. This method typically includes using a

R.F. Belli (✉)
University of Nebraska-Lincoln, Lincoln, USA
e-mail: bbelli2@unl.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_30

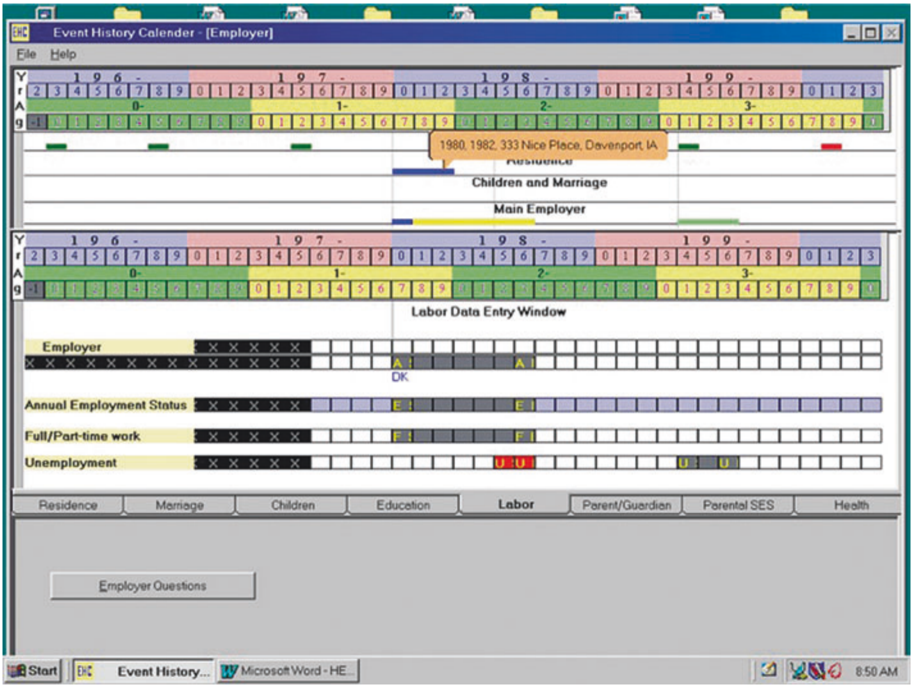
physical calendar as a visual aid for the respondent to improve the precision of their estimates. It is worth noting that calendar interviewing is not generally compatible with standardized interviewing because the steps of the process will look different for each respondent, even if the procedure is the same.

Underlying this method is the idea that one of the best ways to get respondents to provide more accurate reports, in terms of reconstructing their past more accurately, is to use whatever information the respondent can provide from memory as cues for remembering other thematically or temporally related events (Belli, 1998). For example, a respondent may tell the interviewer that they had a child in June 1984 and then the interviewer will use that event and date as an anchor to determine when other events in that time period occurred. These are often events that are harder to remember such as the date that the respondent moved to a new home.

Because most of the cues collected and used for calendar interviewing are idiosyncratic to individuals, standardization is not a practical approach when using the event history calendar method to improve recall accuracy. This means that flexible interviewing approaches that emphasize conversational interviewing are typically used when calendar methods are being used. Calendar methods are often used to assist with collecting data over very long reference periods such as years, decades, or the entire lifespan of the respondent, as can be seen in the example calendar CAPI interface in Fig. 30.1.

Interviewers are trained to apply at least three types of memory retrieval strategies using the cues they collect with the calendar method. First is the sequential retrieval strategy, in which the interviewer helps the respondent to identify a salient event and then uses that event as an anchor to move forward or backward in time to identify other temporally proximal events. An example of this would be asking a respondent to reconstruct their employment history forward through time starting from the start of their first job. Second is parallel retrieval in which the interviewer uses a cue provided by the respondent to identify other events that co-occurred with the cue, for example, having lived in a particular place could act as a cue for who the respondent's employer was at the time. Lastly, there is top-down retrieval which identifies a general event and then drills down to collect more specific information, for example, after having identified an employer for a specific period, the interviewer can then probe about whether or not there were any changes in income.

Existing research on the results of the calendar method has indicated that it generates more accurate reports, particularly for temporally remote events,



1980, 1982, 333 Nice Place, Davenport, IA

Fig. 30.1 An example of a CAPI event history calendar

than conventional questionnaires alone (Belli et al. 2007). Reports of sensitive events also seem to be more accurate when the event history calendar is used than a conventional questionnaire alone.

Time-use diaries are also commonly used for improving respondent recall accuracy. In many cases these are self-administered by the respondent; however they can also be used by interviewers to aid respondent recall of events from the past day or week. In either context these tools have been shown to increase recall accuracy of past events. These and similar approaches are discussed in more detail in the Weir chapter on "*Leave-Behind Measurement Supplements*" and the Stone chapter on "*Experience Sampling and Ecological Momentary Assessment*" in this volume.

However, these measurement tools are costly to implement due to the amount of interviewer time they require and the potential of adding burden on respondents. Future research is needed on ways to computerize more of these tools and make them adaptive and intelligent. Moving toward more self-administration with computerized instruments that are easy to use and capable of collecting data at the same level of accuracy and precision as interviewer-administered tools would drastically increase the use of these tools.

There are many sources of data that future computerized tools can draw on and methods for making use of these data to develop better interfaces and tools for self-administered questionnaires. However, much more research is needed before these ideas can become reality. Figure 30.2 shows one general model of easily collected data (in yellow ovals), existing methods for processing these data (in light blue squares), and theorized tools (in dark blue squares) that could be developed to ease implementation of specialized measurement tools for past events.

Once systems like this are developed, a considerable amount of research will need to be done to evaluate the reliability and validity of the data collected when compared to traditional methods. Effects of self-administration on accuracy, comprehensiveness of data collection, and compliance will be of particular concern. The exploration and development of on-the-fly data quality metrics will also be important to making smart survey instruments feasible.

In summary, specialized tools for measuring past events are uniquely capable of improving respondent recall accuracy for events that occurred years or decades in the past. Event history calendars have demonstrated superior performance to conventional questionnaires, indicating that these tools do work. Time-use diaries have similarly been shown to improve recall of episodes from the previous day or week when compared with conventional questionnaires. These methods have drawbacks in terms of cost and the amount of time that they add to the interview process, but future research on computerizing these tasks and developing smart survey instruments could

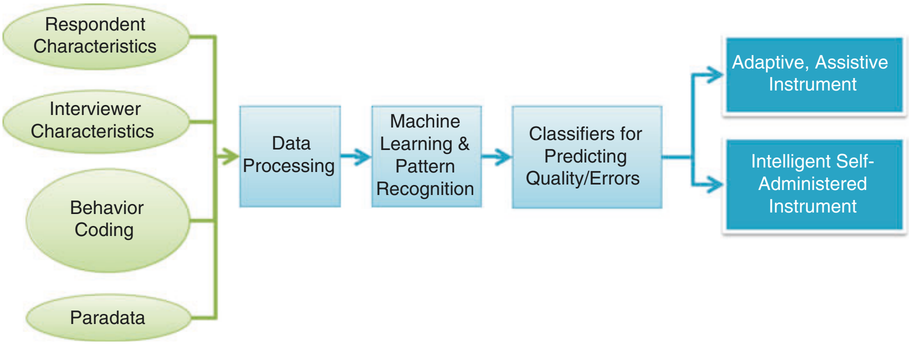

Respondent Characteristics
Interviewer Characteristics
Behavior Coding
Paradata
Data Processing
Machine Learning & Pattern Recognition
Classifiers for Predicting Quality/Errors
Adaptive, Assistive Instrument
Intelligent Self-Administered Instrument

Fig. 30.2 A model of data types, data collection modes, and new tools

make it possible to collect highly accurate and precise data using self-administered modes.

Areas for future research:

- Identifying best practices for reducing the time associated with event history calendars and time-use diaries
- Developing and testing computerized systems to explore whether self-administration is feasible for these tasks

References and Further Reading

Belli, R. F. (1998). The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys. *Memory*, *6*(4), 383–406.

Belli, R. F., Smith, L. M., Andreski, P. M., & Agrawal, S. (2007). Methodological Comparisons between CATI Event History Calendar and Standardized Conventional Questionnaire Instruments. *Public Opinion Quarterly, 71*(4), 603–622. http://doi.org/10.2307/25167583?ref=search-gateway:84a4cdb1103e0262c7f8b5723003d767

Robert F. Belli is Professor of Psychology at the University of Nebraska-Lincoln. He served as North American Editor of Applied Cognitive Psychology from 2004–2009. He received his Ph.D. in experimental psychology from the University of New Hampshire in 1987. Dr. Belli's research interests focus on the role of memory in applied settings, and his published work includes research on autobiographical memory, eyewitness memory, and the role of memory processes in survey response. The content of this work focuses on false memories and methodologies that can improve memory accuracy. His current research is examining the electrophysiological correlates of suggestibility phenomena, and the conversational and memory processes that optimize the quality of retrospective survey reports. Teaching interests include courses on basic and applied cognitive psychology, and on the psychology of survey response.

31

Collecting Social Network Data

Tobias H. Stark

Introduction

For decades, sociologists have been interested in the effects of social networks on people's behavior, attitudes, and economic success (Borgatti et al. 2009; Granovetter 1973). But also scholars in other fields such as medicine (Christakis and Fowler 2007), public health (Cornwell et al. 2014), and social psychology (Wölfer et al. 2015) have acknowledged the importance of social networks for phenomena in their discipline. Researchers in all of these fields use similar methods to assess the characteristics of people's network contacts and to get an understanding of the social structure that surrounds people. While social network data can be collected in many different ways, including archival records, tracking devices, or from mining the Internet, the vast majority of social network researchers still make use of surveys to gather information about the social connections of their subjects.

Comprehensive reviews of survey methods for social network data have appeared recently (Cornwell and Hoaglin 2015; Marsden 2011). Accordingly, the goal of this chapter is not to give an additional overview of these methods. Rather, the focus is on the challenges survey researchers

T.H. Stark (✉)
Utrecht University, Utrecht, The Netherlands
e-mail: t.h.stark@uu.nl

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_31

working with social networks are experiencing. Promising areas for future research are also identified that might help overcome some of these challenges.

In principle, it is possible to distinguish two types of network studies. In a *whole network study*, each member of a predefined social network completes an interview in which he or she indicates with which other persons in the network a relationship exists. A necessary prerequisite for this type of network study is that every member of a predefined social network can be identified in advance and that everyone is reachable for an interview. Accordingly, whole network studies are typically limited to one social context such as a school or a school class. The advantage of whole network studies is that researchers can get a very accurate picture of the social structure of the entire network (i.e., who is connected with whom). This contrasts with *ego-centered network studies*. Here, respondents (called egos) are asked to name their social contacts and these contacts do not need to be part of a predefined social network or belong to the same social context. Instead of interviewing all contacts, the survey respondents are asked to answer proxy questions about their network contacts. With this approach, only the direct network of respondents can be mapped and the larger, surrounding network that also includes the contacts of respondents' contacts remains invisible. The advantage of ego-centered network studies is that they can be included in regular surveys because the focus is not on a predefined social network of which all members need to be interviewed. Many large-scale national representative studies such as the American National Election Study, the General Social Survey (GSS), the Netherlands Life Course Survey, or the German Socio-Economic Panel have from time to time implemented ego-centered networks in their study design.

Whole Network Studies

Whole network studies typically focus on small social settings with clear boundaries to identify all members of the underlying social network. These can, for instance, be business leaders who interact in the same industry, employees within one company, or students in the same schools. All relationships that these people might have with people outside of the particular social setting are outside of the scope of a whole network study. This means every person of the sampling frame needs to be interviewed in a whole network study. As a consequence, whole network studies are typically case studies and cannot be implemented as part of a representative survey.

Survey researchers have to make a number of decisions when designing a whole network study. Typically, respondents are presented with a list of names of all members of the social network and are then asked to identify with whom they have a relationship. Two approaches have been used in the past. Some researchers print ID numbers next to the names of all network members and asked the respondent to write down the ID numbers of their contacts. Other researchers prefer to provide a name list with check boxes and ask the respondent to check the box next to the names of their contacts. No research has evaluated which method might yield more accurate representations of a network even though both methods face potential problems. The ID number method might encourage underreporting of network contacts because considerably more (mental) effort is needed when contacts have first to be found on one list and an associated ID number has then to be copied to another form. Moreover, this method adds a potential source of measurement error because copying ID numbers from one sheet to another may be prone to more error than simply checking a box. The check-box method, in contrast, might invite overreporting of network contacts because it is very simple to mark a large number of names on a list. For instance, in my own research, I have encounter students in school studies who report to be "best friends" with all of their 30 classmates. While possible, this seems highly unlikely. One possible solution to this problem is to only study network connections that are based on mutual nominations of both persons involved in a relationship. However, whether this is possible depends on the type of relationship that is studied (e.g., bullying relationships tend to be one-sided) and on the research questions (e.g., sometimes it is of interest under which circumstance a network nomination is reciprocated).

Researchers have also to decide whether they want to allow respondents to identify up to a certain number of network contacts or that respondents can identify as many contacts as they please. For instance, the National Longitudinal Study of Adolescent to Adult Health (AddHealth) allowed students to only identify their five best male friends and their five best female friends out of all students of their high school. European researchers often focus on school classes instead of entire schools and typically allow students to identify as many contacts among their classmates as they wish (e.g., Stark 2015). Research has found that both methods yield similar results but the unlimited nominations approach seems to be more valid when nominations are made that gauge social status (Gommans and Cillessen 2015).

Additional challenges for survey researchers

- Gaining access to name records in advance: To prepare the name lists, researchers need to have access to the names of all members of the social setting that is being studied. This is problematic if participant consent can only be obtained at the moment of data collection when the name lists have to already be prepared.
- Informed consent: Respondents who complete a social network questionnaire give information about "secondary subjects" (their network contacts). These secondary subjects have to be identifiable by the researcher in order to assess the structure of the social network. This means that data might be collected on people who have refused to participate in the study and have thus not given permission for any data being collected about them. Typically, network researchers have to make a strong argument with Institutional Review Boards (IRB) to justify their method.
- Cognitive burden: In large social settings (e.g., schools in AddHealth), respondents have to go through a long list of names to identify their contacts. Digital questionnaires (e.g., computer-assisted self-interviewing (CASI) or Internet surveys) can help reduce the cognitive burden if they are linked to a database with all names. Respondents start typing the names of their contacts and the computer can suggest matching names.

Ego-Centered Network Studies

Ego-centered network studies focus on the "core personal networks" (Marsden 2011) of respondents instead of the complete social network in a given social setting. In a first step, respondents are asked to identify their network contacts. This is done with "name generator questions." The most well-known name generator is used in the GSS and asks respondents, "From time to time, most people discuss important matters with other people. Looking back over the last six months – who are the people with whom you discussed matters important to you? Just tell me their first names or initials." Other name generator questions ask for names of people whom the respondents feel close to or from whom they could borrow money. The choice of name generator depends on the type of social relationship a researcher is interested in. Some researchers rely on one of these questions to assess the core personal network of their respondents but asking at least two different name generator questions seems to produce more accurate measures of network size (Marin and Hamilton 2007).

Name generators can be compromised if people forget network contacts (recall bias, see Bell et al. 2007; Brewer 2000), if people misinterpret the name generator question (e.g., Bearman and Parigi 2004; Brashears 2011; Small 2013), and due to random error (Almquist 2012). Unfortunately, there is no gold standard to achieve highest validity and reliability of name generator questions. Probes that call respondents' attention to different contexts and to people that may be close to already named contacts seem to reduce forgetting social contacts (Marin 2004). Recent research also suggests that asking respondents to go through their cell phone book to check for names they might be forgetting reduces the recall bias (Hsieh 2015).

To assess the dynamics of ego-centered social networks over time, such as with longitudinal surveys, Cornwell and colleagues (2014) developed a roster matching technique for name generator questions. After respondents have been interviewed for the second time in a longitudinal study, the names of the network contacts reported in the previous wave are matched to the new answers given. Respondents can then be asked to verify the matches and correct mistakes. This technique allows following up with questions about the reasons for changes in the network compositions that would otherwise only be detected in the data analysis phase of a research project.

Because the network contacts of survey respondents are typically not interviewed in an ego-centered network study, respondents have to report the characteristics of their contacts. These proxy reports are done in follow-up questions about each contact that are called "name interpreter questions." These questions either ask about characteristics of each of the contacts (e.g., "Is [NAME 1] a man or a woman?", "Is [NAME 2] a man or a woman?") or about relationships between the contacts (e.g., "Does [NAME 1] know [NAME 2]?"). Thus, information about people in the network and the structure of the social network relies entirely on the perception of the survey respondents.

There has been extensive research on the quality of answers given about network contacts in such follow-up questions. In this volume the chapter by Cobb (2017) focuses entirely on the quality of these proxy reports. In general, answers given by respondents about their contacts do often not correspond with the answers these contacts give themselves when they are also interviewed. This is particularly true for questions about non-observable characteristics, such as network contacts' attitudes. However, the accuracy of proxy reports about network members is often less of a concern in social network analysis than it is when proxy reports are used in a regular survey to replace a hard-to-reach target respondent. The reason for this is that, in network analysis, researchers are typically interested in how the network influences their respondents. For this purpose, the perception of the network by the respondent might often be more important than the objective characteristics of the network (Cornwell and Hoaglin 2015). Yet, the ultimate test of this assumption is still lacking. A study that compares the impact of perceptions of a person's network on that person's attitudes or behaviors to the impact of objective measures of the network by also interviewing the network connections has, to the best of my knowledge, not been conducted yet.

Additional challenges for survey researchers:

- Cognitive burden: Repeatedly answering the same follow-up questions for each network connection or pair of network contacts may impose a substantial cognitive burden on respondents and reduce the data quality (Hsieh 2015; Matzat and Snijders 2010; Tubaro et al. 2014; Vehovar et al. 2008).
- Size of the network: Answering multiple follow-up questions about each network contact takes up valuable interview time. Moreover, the number of pairs of contacts that have to be evaluated to assess the structure of a person's social network (i.e., who knows whom) increases exponentially with the size of the network (McCarty et al. 2007). Accordingly, most researchers limit their respondents to a maximum of five network contacts and this number allows producing reliable estimates of many network characteristics such as network composition and network density (Marsden 1993).
- Mode effects: Research found that the number of connections between respondents' network contacts (e.g., "Does [NAME 1] know [NAME 2]?") was exaggerated in an online survey compared to a face-to-face survey (Matzat and Snijders 2010). More research is needed comparing different modes.
- Interviewer effects: Interviewers learn that more network contacts increase the length of an interview due to the follow-up question about each contact and have been found to shorten an interview by falsely reporting no or very few network contacts (Eagle and Proeschold-Bell 2015; Paik and Sanchagrin 2013). An interesting approach for future research might be to make use of mixed-mode designs. The names of the network contacts could be collected in CASI mode while an interviewer could collect the follow-up questions about the contacts in computer-assisted personal-interviewing (CAPI) mode.

Data Collection

Despite potential interviewer effects, face-to-face interviews or telephone interviews in which interviewers can motivate respondents to answer repetitive follow-up questions effortfully are still considered the best way to collect ego-centered network data (Marsden 2011). However, independent of the mode, a computer is necessary to handle the complexity of an ego-centered social network questionnaire because the names of the network contacts have to be pasted into the follow-up questions about the contacts (e.g., "Is [NAME 1] a man or a woman?"). Thus computer-assisted telephone-interviewing (CATI), CASI, CAPI, or Internet surveys can be used for ego-centered network questionnaires.

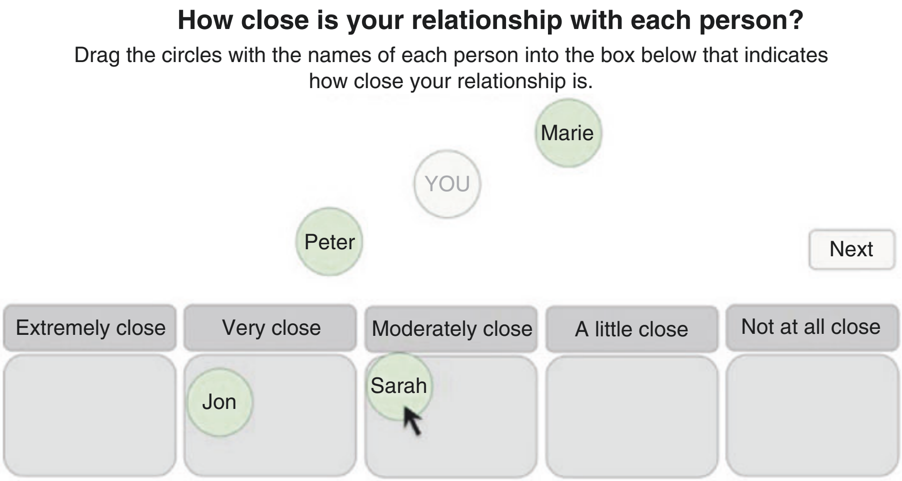
When a research question requires data on whole networks, paper-and-pencil questionnaires can be used in addition to the computer-assisted modes. The reason is that researchers using a whole network design typically only want to know who the network contacts are. Follow-up questions about these contacts are not necessary because these people are also part of the study and complete a questionnaire on their own.

Some design recommendations have been made for ego-centered network studies that use self-administration of the questionnaire (CASI, Internet). Most importantly, asking about one attribute of all network contacts in follow-up questions before asking about another attribute of all contacts leads to less item nonresponse, less drop out (Vehovar et al. 2008), and more reliable data (Coromina and Coenders 2006) than does asking all follow-up questions about one contact before asking all questions about the next contact. Also the number of name boxes displayed under the name generator question should be well considered because respondents tend to match the number of names they give to the number of name boxes they see (Vehovar et al. 2008).

Recently, graphical software tools have been developed that make use of these design recommendations and try to make the process of answering ego-centered network questionnaires less repetitive and thus more enjoyable. These tools make use of visual aids to reduce the cognitive burden for respondents. The survey tool PASN (Lackaff 2012) derives names of respondents' social networks by accessing their Facebook profiles whereas the tool TellUsWho (Ricken et al. 2010) mines respondents' email accounts for names. Subsequently, respondents can answer questions about their contacts by dragging and dropping the names or the Facebook profile pictures of their contact into answer boxes. The software ANAMIA EGOCENTER (Tubaro et al. 2014) lets respondents draw a picture of their network in great detail, which give valuable information about connections and cliques but may make completing a network survey rather complex.

A new class of graphical data collection software lets respondents answer follow-up questions about their network contacts through interacting with a visual representation of the network. Such approaches have been implemented in the programs OpenEddi (Fagan and Eddens 2015), netCanvas (Hogan et al. 2016), and GENSI (Stark and Krosnick 2017). Questions about the network contacts can be answered by either clicking on the names of the contacts (for dichotomous questions) or by dragging and dropping the names into answer boxes (an example using GENSI is shown in Fig. 31.1). Connections between the network contacts (i.e., who knows whom) can be indicated by the traditional approach of asking separately for each pair of network contacts whether a relationship exists or by drawing lines between the names of two or more contacts in the figure of the network (Fig. 31.2). OpenEddi also allows indicating connections by sorting the names of network contacts into piles.

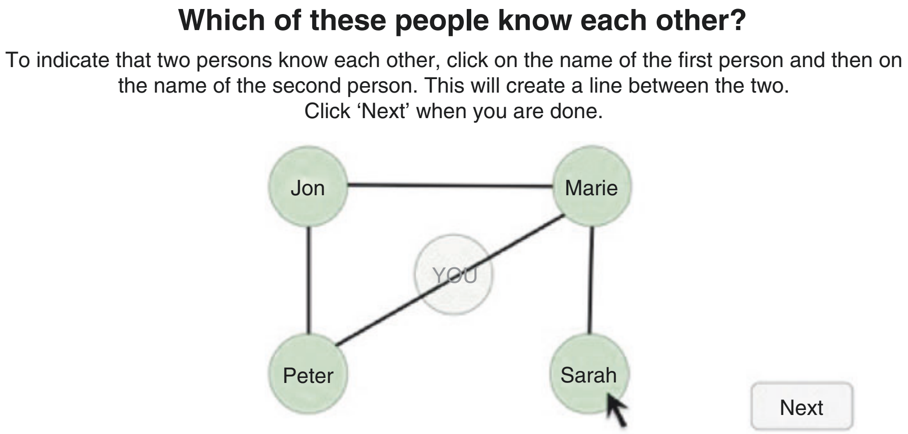
All of this new software has been developed to reduce respondent burden and increase data quality. However, very little research has been done to examine respondents' perception of these graphical tools and the quality of data collected. An evaluation study found that GENSI produces data of equal quality as a traditional ego-centered questionnaire (Stark and Krosnick 2017). However, respondents enjoyed completing the questionnaire more



How close is your relationship with each person?

Drag the circles with the names of each person into the box below that indicates how close your relationship is.

Fig. 31.1 Drag-and-drop question in GENSI. Answers are given by moving name circles into answer boxes



Which of these people know each other?

To indicate that two persons know each other, click on the name of the first person and then on the name of the second person. This will create a line between the two.
Click 'Next' when you are done.

Fig. 31.2 Question for network relationships in GENSI. Relationships are indicated by drawing lines between related network contacts

with GENSI than with the traditional design. The new tool also seems to solve what past researcher have considered to be a problem with online administration; exaggerated numbers of connections between the contacts of a respondent (Matzat and Snijders 2010). Given these promising results, future research that compares the various programs with each other and with traditional ways to collect ego-centered network data seems highly valuable.

Social Media

With the abundant availability of social network data produced by social media, one might wonder why few network researchers make use of social media data but rather use surveys to collect their data. A central challenge for this research is that many social media websites that have a large amount of information on their users (e.g., Facebook, Instagram) do not allow access to these data. For instance, Facebook recently changed its application programming interface (API) to no longer support automatic downloads of user data. It is thus only possible to access and code public user profiles by hand but not in a less time-consuming automatized fashion. In contrast, social media data that are publicly available and can be automatically downloaded typically do not offer much information on a particular user (e.g., Twitter), which makes it difficult to link this data to survey data.

There are a number of additional limitations that make surveys still a preferred mode for collecting social network information. First, social media networks are limited to relationships between people who are users of the social media website. This means that important parts of a population might be missed when a study relies completely on social media data. Whether this is a problem or not depends, of course, on the research question at hand. A study of, for instance, social influence through social media will not be affected by this limitation whereas a study that is interested in social influence between people in general might overlook influential actors that are not members of the social media network. Another limitation of social media data is that there is typically no way to understand the nature of the relationship between two people on the website. The pure existence of a connection says little about the actual relationship because on most websites everybody is linked in the same way (e.g., friends on Facebook, or contacts on LinkedIn). Researchers could enrich these data by accessing and coding the communication between connected people on the website. However, interpreting the meaning of the communication is typically difficult. Moreover, important communication may take place outside of social media. It is possible that people who interact on a daily basis, and are thus very relevant network contacts, make less use of communication through social media. A final limitation is that the available information on social media websites is restricted to users' behavior on the website whereas other information that is of interest for many researchers (e.g., age, sex, attitudes) is often missing. Recently, researchers have started to overcome this problem partially by combining social media information with survey data (Schober et al. 2016). This seems to be a valuable direction for future research.

Areas for future research:

- The existing best-practice recommendations for network questionnaire designs are based on small samples that are not representative for any population. Tests of these recommendations with data from random probability samples are needed.
- It is still unclear whether people's perceptions of their social network or objective measures of their networks have more impact on people's attitudes and behaviors.
- Tests of mode effects are needed, both for the collection of whole network data and ego-centered network data.
- Evaluation studies of the various existing graphical tools to collect ego-centered network data in comparison to traditional survey tools are needed. Do the graphical tools reduce cognitive burden and produce better measures of social networks?

- A combination of whole networks to gauge relationships within a social setting with ego-centered networks to assess relationships outside of this setting might overcome the weaknesses of both approaches. Such data are rarely collected because different research questions typically motivate the collection of whole or ego-centered networks. One noteworthy exception forms the CILS4EU study for which classroom network data were collected among more than 18,000 students from 958 classrooms in four European countries. Students were also asked to complete ego-centered network questions about up to five friends outside of their classroom.
- Research that links social media information with network survey data might give insights in how relevant these ties are compared to the network that is typically assessed with traditional approaches. The chapters in this volume by Pasek (2017), and Blaermire (2017) respectively give an overview of the opportunities and challenges associated with linking survey data with such external data.

References and Further Reading

Almquist, Z. W. (2012). Random Errors in Egocentric Networks. *Social Networks*, *34*(4), 493–505.

Bearman, P., & Parigi, P. (2004). Cloning Headless Frogs and Other Important Matters: Conversation Topics and Network Structure. *Social Forces*, *83*(2), 535–557.

Bell, D. C., Belli-McQueen, B., & Haider, A. (2007). Partner Naming and Forgetting: Recall of Network Members. *Social Networks*, *29*(2), 279–299.

Blaermire, B. (2017). Linking Survey Data with the Catalist Commercial Database. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research*. New York: Palgrave.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network Analysis in the Social Sciences. *Science*, *323*(5916), 892–895.

Brashears, M. E. (2011). Small Networks and High Isolation? A Reexamination of American Discussion Networks. *Social Networks*, *33*(4), 331–341.

Brewer, D. D. (2000). Forgetting in the Recall-Based Elicitation of Personal and Social Networks. *Social Networks*, *22*(1), 29–43.

Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *The New England Journal of Medicine*, *357*, 370–379.

Cobb, C. (2017). Proxy Reporting. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research*. New York: Palgrave.

Cornwell, B., Schumm, L. P., Laumann, E. O., Kim, J., & Kim, Y. J. (2014). Assessment of Social Network Change in a National Longitudinal Survey. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences*, *69*, S75–S82.

Cornwell, B., & Hoaglin, E. (2015). Survey Methods for Social Network Research. In T. P. Johnson (Ed.), *Health survey methods* (pp. 275–314). Hoboken, NJ: Wiley.

Coromina, L., & Coenders, G. (2006). Reliability and Validity of Egocentered Network Data Collected Via Web – A Meta-Analysis of Multilevel Multitrait, Multimethod Studies. *Social Networks*, *28*(3), 209–231.

Eagle, D. E., & Proeschold-Bell, R. J. (2015). Methodological Considerations in the Use of Name Generators and Interpreters. *Social Networks*, *40*, 75–83.

Fagan, J., & Eddens, K. (2015). *OpenEddi*. Paper presented at the XXXV Sunbelt Conference. Brighton, UK.

Gommans, R., & Cillessen, A. H. N. (2015). Nominating Under Constraints. A Systematic Comparison of Unlimited and Limited Peer Nomination Methodologies in Elementary School. *International Journal of Behavioral Development*, *39*(1), 77–86.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*, 1360–1380.

Hogan, B., Melville, J. R., Ii, G. L. P., Janulis, P., Contractor, N., Mustanski, B. S., & Birkett, M. (2016). *Evaluating the Paper-to-Screen Translation of Participant-Aided Sociograms with High-Risk Participants*. Paper presented at the Human Factors in Computing, San Jose, CA.

Hsieh, Y. P. (2015). Check the Phone Book: Testing Information and Communication Technology (ICT) Recall Aids for Personal Network Surveys. *Social Networks*, *41*, 101–112.

Lackaff, D. (2012). New Opportunities in Personal Network Data Collection. In M. Zacarias & J. V. De Oliveira (Eds.), *Human-Computer Interaction*. Berlin: Springer.

Marin, A. (2004). Are Respondents More Likely to List Alters with Certain Characteristics?: Implications for Name Generator Data. *Social Networks*, *26*(4), 289–307.

Marin, A., & Hamilton, K. (2007). Simplifying the Personal Network Name Generator: Alternatives to Traditional Multiple and Single Name Generators. *Field Methods*, *19*, 163–193.

Marsden, P. V. (1993). The Reliability of Sociocentric Measures of Network Centrality. *Social Networks*, *15*, 399–422.

Marsden, P. V. (2011). Survey Methods for Network Data. In J. Scott & P. J. Carrington (Eds.), *The SAGE Handbook of Social Network Analysis* (pp. 370–388). London: Sage.

Matzat, U., & Snijders, C. (2010). Does the Online Collection of Ego-Centered Network Data Reduce Data Quality? An Experimental Comparison. *Social Networks*, *32*(2), 105–111.

McCarty, C., Killworth, P. D., & Rennell, J. (2007). Impact of Methods for Reducing Respondent Burden on Personal Network Structural Measures. *Social Networks*, *29*, 300–315.

Paik, A., & Sanchagrin, K. (2013). Social Isolation in America: An Artifact. *American Sociological Review*, *78*(3), 339–360.

Pasek, J. (2017). Linking Knowledge Networks Web Panel Data with External Data. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research*. New York: Palgrave.

Ricken, S. T., Schuler, R. P., Grandhi, S. A., & Jones, Q. (2010). TellUsWho: Guided Social Network Data Collection. *Proceedings of the 43rd Hawaii International Conference on System Sciences*.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social Media Analyses for Social Measurement. *Public Opinion Quarterly*, *80*(1), 180–211.

Small, M. L. (2013). Weak Ties and the Core Discussion Network: Why People Regularly Discuss Important Matters with Unimportant Alters. *Social Networks*, *35*(3), 470–483.

Stark, T. H. (2015). Understanding the Selection Bias: Social Network Processes and the Effect of Prejudice on the Avoidance of Outgroup Friends. *Social Psychology Quarterly*, *78*(2), 127–150.

Stark, T. H., & Krosnick, J. A. (2017). GENSI: A New Graphical Tool to Collect Ego-Centered Network Data. *Social Networks*, *48*, 36–45.

Tubaro, P., Casilli, A. A., & Mounier, L. (2014). Eliciting Personal Network Data in Web Surveys Through Participant-Generated Sociograms. *Field Methods*, *26*(2), 107–125.

Vehovar, V., Manfreda, K. L., Koren, G., & Hlebec, V. (2008). Measuring Ego-Centered Social Networks on the Web: Questionnaire Design Issues. *Social Networks*, *30*(3), 213–222.

Wölfer, R., Faber, N. S., & Hewstone, M. (2015). Social Network Analysis in the Science of Groups: Cross-Sectional and Longitudinal Applications for Studying Intra- and Intergroup Behavior. *Group Dynamics-Theory Research and Practice*, *19*(1), 45–61.

Tobias H. Stark is an Assistant Professor at the European Research Centre on Migration and Ethnic Relations (ERCOMER), Utrecht University, The Netherlands. He studies how social networks affect the development and spread of prejudice, as well as how prejudice hinder the development of interethnic relationships. He uses insights of this research to develop anti-prejudice interventions in schools. His dissertation "Integration in Schools. A Process Perspective on Students' Interethnic Attitudes and Interpersonal Relationships" was awarded with a Research Prize of the Erasmus Prize Foundation and won the prize for best dissertation of the Dutch Sociological Association (NSV). Dr. Stark's research has been funded by the

European Commission (Marie-Curie International Outgoing Fellowship) and the Netherlands Organisation for Scientific Research (Veni grant). His work has appeared in sociological and social-psychological high-impact journals such as *Social Networks, Social Psychology Quarterly, Public Opinion Quarterly*, and *Personality and Social Psychology Bulletin*.

Section 3

Linking Survey Data with External Sources

32

Methods of Linking Survey Data to Official Records

Joseph W. Sakshaug

Government agencies, medical facilities, and market researchers are among the entities that collect vast amounts of data in the form of administrative records. These records are rich sources of data that can potentially be linked to survey data in valuable ways. Official records such as government documents, medical records, or academic transcripts can provide not only a source of supplemental data but can act as a gold standard to validate the accuracy of self-report data from surveys.

Administrative record linkage refers to appending datasets based on one or more linking variables that exist in both datasets. These linking variables could be at the person, household, or establishment level and could be such variables as names, addresses, Social Security numbers, tax identification numbers, or other variables that can reliably associate disparate datasets.

There are a number of reasons that linking administrative records is important. First, as mentioned before, are the methodological purposes of assessing data accuracy and the reliability of self-report data from surveys. Another methodological benefit is that these data provide a way to assess nonresponse bias by comparing the survey data collected from respondents with the record data from non-respondents (Kreuter, Müller, & Trappmann, 2010). There are also substantive benefits to linking administrative records

J.W. Sakshaug (✉)
University of Manchester, Manchester, England
Institute for Employment Research, Nuremberg, Germany
e-mail: joe.sakshaug@manchester.ac.uk

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_32

with survey data. For example, it can allow for longitudinal analysis because many types of records are in time-series form such as medical or tax records. Linking also permits researchers to investigate complex policy-oriented questions such as trends in healthcare spending among older populations or lifetime earnings and retirement planning.

Government records are some of the most important types of records that are commonly linked to survey data. Popular administrative databases include Social Security records, which contain detailed earnings and benefit histories (Olson, 1999); Medicare claims records, which document Medicare enrollment and detailed healthcare expenditure records among Medicare beneficiaries. Last, is the National Death Index, which is a database that collects death certificate records from the vital statistics offices of each state, these are aggregated by the National Center for Health Statistics and made publicly available.

Three approaches to linking are commonly used. The first is *exact linkage*, which involves linking administrative records to survey data using a common variable that acts as a unique identifier. These unique identifiers are things like Social Security numbers, Medicare numbers, or tax identification numbers. Respondents are typically relied upon to provide the unique identifier and must also provide informed consent before the linkage can be made. There are some practical concerns associated with attempts to do exact linkage including the fact that consent rates vary across studies and sets of records. This can lead to reduced sample size and biased inference if the people who do not consent are systematically different from those who do on some unmeasured dimension (Yawn et al., 1998; Baker et al., 2000).

The second approach is *probabilistic linkage*; this can be used when there is no unique identifier on which to match the datasets, in this case there are other potential identifiers that can be used together to link records with a certain probability of accuracy (Fellegi & Sunter, 1969). Commonly used identifiers are names, dates of birth, and addresses. These identifiers are used to calculate the probability that an administrative record and a survey report belong to the same unit with the match status being determined by a prespecified probability threshold and decision rule. This approach is commonly used by government agencies such as the Census Bureau and the Centers for Disease Control and Prevention with both of these organizations having developed specialized software packages designed to implement the procedures (Winkler, 1999). Some of the practical issues surrounding using probabilistic linkage are that it is very difficult to estimate the frequency of false matches and false non-matches (Belin & Rubin, 1995). Furthermore, the matching variables may have varying levels of accuracy or missing data themselves, which can add more uncertainty. Lastly, linking three or more databases can be very problematic and there are no accepted best practices for doing so.

The third approach to linking administrative records to survey data is *statistical matching*. This is another method that is used when exact record linkage is not possible. Statistical matching takes the two datasets and uses statistical models to predict the similarity between records and attempts to merge the datasets without regard for identifying cases that actually belong to the same unit. Essentially, this approach uses variables that the two datasets have in common to link statistically similar records; these matching variables may be age, gender, race, and other similar socio-demographic variables. Metrics such as Euclidean distance, predictive mean matching, and propensity score matching can then be used to identify similar records in each dataset and match them. Statistical matching has many practical problems that must be addressed when it is being implemented, but one of the most basic issues is that it makes very strong statistical assumptions (e.g., the conditional independence of variables Y and Z, given common variable X) that are difficult to test with the actual data and therefore may be unjustified. Considerably more research is needed to evaluate whether statistically matched records actually reflect true values in the population.

Despite having developed these important and useful approaches to linking administrative records data with survey data, there are still many opportunities for basic and important future research on this subject. At a very fundamental level, research on how the properties of these different linkage types influence data quality is still needed. It is unclear how linkage errors affect subsequent statistical analyses or whether the strengths of one technique can be used to overcome the weaknesses of another. For example, exact matching and probabilistic linkage require informed consent from respondents, whereas statistical matching does not, under what conditions does this make statistical matching preferable over the additional effort and potential biases involved with obtaining informed consent? Also on the subject of consent, how low do consent rates need to get before researchers begin considering alternative (non-exact) approaches to linkage? What are the theoretical mechanisms that drive the linkage consent decision? How does this decision differ from the decision to participate in a survey? And how should researchers balance the tradeoffs between data utility and data confidentiality in the unique context of linking administrative records? These are all important questions that future research is needed for.

Methods for linking three or more data sources simultaneously are currently imperfect and in need of additional research. The most common method is called "chaining" where data sources are linked sequentially starting with the most reliable data source. Very little is known about how well this approach works or whether or not it is an optimal method. More evaluation is needed on how linking multiple data sources may affect subsequent statistical analyses.

Lastly, one idea that is starting to gain some attention is the notion of starting with the administrative records data and designing the survey around them, this approach turns the traditional view of linking records on its head and treats the official records as the primary dataset and the survey data as supplemental. This approach could lead to reduced data collection costs, reduced respondent burden, more efficient survey design and use of records, greater transparency in how records are collected, and expanded opportunities for scientific research on linking records and survey data. But so far existing research testing this approach has been limited. Future research should examine this idea to explore any promise that it may hold.

Areas for future research:

- Identifying the ways that linkage errors influence subsequent statistical analyses
- Identifying methods to assess the accuracy of administrative records
- Best practices for selecting the most appropriate linkage method
- Effects of consent rates and consent bias on the quality of linked records
- Methods for optimally linking three or more datasets
- Methods for making better use of administrative records prior to survey data collection
- Improving statistical matching algorithms to overcome low linkage consent rates
- Incentivizing interviewers to obtain higher linkage consent rates
- Identifying optimal approaches to using multiple linkage techniques in sequence

References and Further Reading

Baker, R., Shiels, C., Stevenson, K., Fraser, R., & Stone, M. (2000). What proportion of patients refuse consent to data collection from their records for research purposes? *British Journal of General Practice*, *50*(457), 655–656.

Belin, T. R., & Rubin, D. B. (1995). A Method for Calibrating False-Match Rates in Record Linkage. *Journal of the American Statistical Association, 90*(430), 694–707. http://doi.org/10.1080/01621459.1995.10476563

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Kreuter, F., Müller, G., & Trappmann, M. (2010). Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. *Public Opinion Quarterly, 74*(5), 880–906. http://doi.org/10.1093/poq/nfq060

Olson, J. A. (1999). Linkages with data from Social Security administrative records in the Health and Retirement Study. *Soc Sec Bull*, 62(2): 73–85.

Winkler, W. E. (1999). *The State of Record Linkage and Current Research Problems. Statistical Research Division*. U.S. Census Bureau Statistical Research Division.

Yawn, B. P., Yawn, R. A., Geier, G. R., Xia, Z. S., & Jacobsen, S. J. (1998). The impact of requiring patient authorization for use of data in medical records research. *Journal of Family Practice*, *47*(5), 361–365.

Joseph W. Sakshaug is a Senior Lecturer in Social Statistics in the School of Social Sciences at the University of Manchester (UK), a Senior Researcher in the Department of Statistical Methods at the Institute for Employment Research (Germany), an Adjunct Research Assistant Professor in the Survey Research Center at the Institute for Social Research at the University of Michigan, and a faculty member in the International Program in Survey and Data Science offered through the University of Mannheim and the Joint Program in Survey Methodology. He received his Ph.D. and M.S. degrees in Survey Methodology from the University of Michigan and a B.A. in Mathematics from the University of Washington. He conducts research on record linkage, nonresponse and measurement errors in surveys, small area estimation, and statistical disclosure control.

33

Linking Individual-Level Survey Data to Consumer File Records

Josh Pasek

In addition to official administrative records, such as those collected by government agencies, medical facilities, or employers, another class of records data also exists in the form of consumer file data collected by market research organizations. These data can be purchased from clearinghouses that aggregate the data and make it available to marketers and market researchers. InfoUSA, one of the largest data clearinghouses, has a database of over 3.5 billion records from over 75 sources that they sell access to; other sources of these large datasets include Experian, Acxiom, and Marketing Systems Group. These databases typically contain individual-level demographics, addresses, and other general information.

Survey researchers are interested in these data because, to the extent that they can be linked to the individuals selected using traditional survey sampling procedures, it may be possible to develop a better understanding not only of respondents but also of non-respondents. There are several potential benefits that linking consumer databases with survey sampling frames could confer. First, it could help improve the efficiency of sampling; by integrating the demographic and other ancillary data with the sampling frame information it may be easier to accurately oversample hard-to-reach populations. This could lead to lower costs of data collection and more representative samples of the

J. Pasek (✉)
University of Michigan, Ann Arbor, USA
e-mail: jpasek@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_33

population being surveyed, which might reduce bias and the need for large *post hoc* survey weights. Another potential benefit from these data is that they can provide substantially more information about non-respondents than is commonly available, which has implications for how non-response bias calculations are made and might even enable researchers to correct for biases in the sampling frame. Taken together, the potential benefits for using these data in conjunction with surveys are difficult to overstate.

However, all of these potential benefits are predicated on the assumption that the data in the consumer databases are of sufficient quality to serve these purposes. The two primary dimensions of quality that need to be considered are accuracy and completeness. If the consumer file data are inaccurate, then linking with survey data could have deleterious rather than beneficial effects because it could lead to incorrect inferences or might lead us to misstate the nature of non-response biases. If the consumer data are incomplete then there could be differential levels of certainty in the quality of the match with the survey data, which adds another layer of complexity to the already challenging problem of missing data. These are important questions that need to be addressed by future research.

The current state of research seems to indicate that market research databases are not yet of sufficient quality to justify their widespread use in conjunction with survey data. The consumer file data are not particularly accurate, estimates derived from these data frequently differ from self-reports, missing consumer data are systematic and non-ignorable, and standard imputation algorithms do not fully correct these biases. However, the potential benefits from these databases certainly warrant considerable future research into how these sources of information can best be used in valid ways. Research is needed to identify appropriate and valid uses for these data along with novel ways for correcting the biases and inaccuracies in the consumer marketing data.

One important feature of these market research databases that severely limits their utility for serious research is the lack of transparency with regard to how the data are collected, matched internally, and manipulated prior to bundling for sale. The companies that sell access to these data nearly universally claim that this is proprietary “trade-secret” information, and they refuse to give access to these key features of the data that would enable researchers to evaluate their quality. Until these practices change it is unlikely that these databases will be able to provide their maximum benefit to researchers. Serious researchers then should restrict their use of these commercial data to experiments and exploratory analyses; applications for substantive analyses should be avoided until greater research transparency can be achieved.

Areas for future research:

- Identifying organizations with commercial databases that are willing to engage in transparent research practices
  - When the records were obtained
  - From whom the records were obtained
  - How the records are cleaned
  - Full disclosure about the matching algorithms including the criteria for determining a match
- Assessing the correspondence of data from consumer databases and survey self-reports
- Evaluating the nature of missing data in the consumer databases
- Exploring approaches to correct for missing data in consumer databases
- Using consumer data to examine variables that might not be collected by surveys as a supplemental form of data and conducting sensitivity analyses on these data
- Determining whether consumer file data identify differences between respondents and non-respondents that can be used to improve survey weights

Further Reading

Pasek, J., Jang, S. M., Cobb, C. L., III, & Dennis, J. M. (2014). Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data. *Public Opinion Quarterly, 78*(4), 889–916. http://doi.org/10.1093/poq/nfu043

Sinibaldi, J., Trappmann, M., & Kreuter, F. (2014). Which Is the Better Investment for Nonresponse Adjustment: Purchasing Commercial Auxiliary Data or Collecting Interviewer Observations? *Public Opinion Quarterly, 78*(2), 440–473. http://doi.org/10.1093/poq/nfu003

Valliant, R., Hubbard, F., Lee, S., & Chang, C. (2014). Efficient Use of Commercial Lists in U.S. Household Sampling. *Journal of Survey Statistics and Methodology, 2*(2), 182–209. http://doi.org/10.1093/jssam/smu006

West, B. T., Wagner, J., Hubbard, F., & Gu, H. (2015). The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology, 3*(2), 240–264. http://doi.org/10.1093/jssam/smv004

Josh Pasek is an Assistant Professor of Communication Studies, Faculty Associate in the Center for Political Studies, and Core Faculty for the Michigan Institute for Data Science at the University of Michigan. His research explores how new media and psychological processes each shape political attitudes, public opinion, and political behaviors. Josh also examines issues in the measurement of public opinion including techniques for reducing measurement error and improving population inferences. Current research explores the origins and effects of both accurate and inaccurate political beliefs and assesses the conditions under which nonprobability samples, such as those obtained from big data methods or samples of Internet volunteers can lead to conclusions similar to those of traditional probability samples. His work has been published in *Public Opinion Quarterly, Communication Research*, and the *Journal of Communication*among other outlets. He also maintains R packages for producing survey weights (anesrake) and analyzing weighted survey data (weights).

34

Linking Survey Data to Administrative Records in a Comparative Survey Context

Annelies G. Blom and Julie Korbmacher

Introduction

In social survey projects, administrative records linked to survey data can offer a wealth of additional information about the respondents. This additional information can be used to *augment* survey data on topics that are difficult or burdensome to collect during the survey interview because the information may be difficult to recall for respondents or may be unknown to them at the level required by the researcher. The types of administrative records that will be linked to the survey data will depend on the subject area. For example, in medical research such records may contain information about hospitalization periods, diagnoses, or prescribed medicines; while economists tend to link detailed information on employment spells, taxable income, benefits receipt, or pension rights.

The linked information may also be used to *validate* the accuracy of the survey data, such that the survey data collection can be improved. Such validation assumes that the administrative records are more accurate than the

A.G. Blom (✉)
School of Social Sciences and Collaborative Research Center "Political Economy of Reforms" (SFB 884), University of Mannheim, Mannheim, Germany
e-mail: blom@uni-mannheim.de

J. Korbmacher
Max Planck Institute for Social Law and Social Policy, Munich, Germany
e-mail: Korbmacher@mea.mpisoc.mpg.de

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_34

respective survey data. This assumption may be met by some types of records (e.g., doctors' diagnoses and medicines prescribed), while other types may not always be so accurate (e.g., voting records in the United States). Moreover, where researchers receive access to both matched and unmatched administrative records, a comparison of the two can deliver insights into nonresponse and coverage biases.

In an increasing number of European countries, the Survey of Health, Ageing and Retirement in Europe (SHARE) links survey data with information from administrative records, for example, on employment- and pension-related variables. SHARE is a multidisciplinary and cross-national panel database of micro-data on social and family networks, health, and socioeconomic status of approximately 123,000 individuals aged 50 or older in 21 European countries (Börsch-Supan et al. 2013). Individual employment histories and pension rights are a key element of the SHARE research endeavor; however, the accurate collection of such information in a survey setting is a challenge. A special linkage project therefore aims to meet this challenge. In 2009 a pilot study started in Germany and linked SHARE data to administrative records of the German pension fund. After the success of the German pilot study, SHARE decided to continue the project in Germany and to scale it up to additional SHARE countries.

The challenges to be met when linking administrative records to survey data, however, differ across European countries. Some of these differences are due to differences in administrative data infrastructures; some of them are due to differences in data protection laws; each challenging the development of uniform procedures for the linkage of survey data with administrative records (Schmidutz et al. 2013). In this chapter, we provide an overview of the opportunities for and obstacles to collecting and conducting research with survey data linked to administrative data across countries (see also Korbmacher and Schmidutz 2015). While the SHARE linkage project has delivered some insights into steps that need to be taken when linking survey and administrative data, a key lesson is also that a limited amount of knowledge is transferable across countries due to country-specific differences. Rather than a set of best practices, this chapter aims to portray some of the challenges in obtaining linked survey and administrative data across countries.

From a researcher's perspective, once appropriate datasets and variables that are suitable and available for linking are identified, data linkage involves two main stages: the *process of linking the survey data to the administrative records* and the *process of making the linked data available for research*. With respect to the former, as outlined in the chapter by Sakshaug in this volume, there are three main ways in which a survey can be linked with administrative data (Sakshaug, this issue). The key difference between the three linking methods lies in the researcher's objective: while exact and probabilistic linkage aims to link the survey data of a *specific individual* to the administrative records of this individual, statistical matching links the survey data of a survey respondent to administrative records of a *similar individual*, where similar is defined by means of key respondent characteristics identified by the researcher. The SHARE linkage project links the survey data and administrative records of the same person, preferably via exact linkage. In this way the linked dataset provides the most accurate information and allows for both the augmentation of the survey data and the validation of the survey data through the administrative records.

A difficulty associated with direct linkage is that most European countries require the survey respondents' informed consent before linkage can take place. While this safeguards respondent interests, the current fragmentation of European data protection law brings about differences in the ways in which respondents can give informed consent.[1] This means that fieldwork for the record linkage has to be set up differently across countries. In the strictest countries, respondents must provide their signature on a consent form (*written consent*). In other countries *verbal consent*, which can be given during a face-to-face or telephone interview, is sufficient. Finally, there are countries where implicit consent is assumed, such that survey data can be linked to administrative records, unless respondents explicitly *opt-out*. Strict consent procedures will lead to a reduction in the size of the linked sample and can, in addition, lead to non-consent bias, if respondents' consent decision is not random but correlates with respondent characteristics. With differences across countries in consent regulations, differences in the non-consent bias mechanisms are likely to occur. This means that special care needs to be taken when conducting comparative research based on linked data.

The way in which administrative and survey data can be linked also depends on the availability of a unique identifier available in both datasets. Some European countries (typically, the Northern European countries, like

[1] The central legislative instrument of European data protection law is the "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.“ The Directive includes a minimum set of provisions to be implemented by the Member States and had to be transposed into national law by all EU Member States by the end of 1998. However, the Data Protection Directive (by definition) is not a self-executing legal instrument and therefore leaves the choice of form and methods to the national authorities. As a result, the provisions of the Directive have been implemented in different ways in the Member States, resulting in differences in the level of data protection, both on paper and in practice (Schmidutz et al. 2013).

Demark or Sweden) have central registers which all use the same unique personal identification number that is also available on the survey-sampling frame. This makes it technically very easy to link survey data with administrative data. In other countries (e.g., in Germany), registers are not available centrally, different types of registers are not interlinked, and unique identifiers are not available on the survey-sampling frame. In such situations, respondents have to provide the identifier (e.g., their Social Security Number (SSN)) or the information needed to generate the identifier during the interview.

The ease with which the identifier is retrieved from respondents depends on how commonly it is used in their daily lives. In this respect, most European countries differ strongly from the United States, where citizens are frequently asked to report their SSN and many persons know their SSN by heart. In Europe, citizens seldom know their SSN by heart and across European countries there are differences in how easily respondents can retrieve their SSN. For example, in Austria the SSN is printed on each citizen's health insurance card. Furthermore, the SSN consists of numbers that can be verified during the interview using a special algorithm. In Germany, in contrast, respondents asked to provide their SSN during a survey interview, have to look it up in their tax documents, and exact validation during the interview is not possible. As a result, the accuracy of the identifier needed for linkage and the response burden associated with the request for the identifier and for informed consent differs largely across countries.

In a next step, the linkage requires the exchange of the survey and administrative data and of the identifier information across different institutions, such as the university conducting the survey and the organization holding the administrative records. This is a highly sensitive procedure because the protection of respondents' anonymity is pivotal. Typically, the personal identifier is the key, which allows the de-anonymization of the respondents. Therefore, linkers have to ensure that none of the individuals and institutions involved has simultaneously access to the survey data, to the raw administrative data, and to the personal identifier. A careful and verified documentation of the linkage process in advance of embarking on data linkage is, therefore, essential to guaranteeing the anonymity of linked respondents. The implementation of a linkage project in a cross-national survey like SHARE with different institutions involved, different legal frameworks to adhere to, and different identifiers, results in country-specific linkage procedures, which have to be developed and verified on a country-by-country basis.

The ultimate goal of the linkage project in SHARE is the provision of anonymized administrative data which can be linked to survey data as

Scientific Use Files (SUFs) and made freely available to the research community. Therefore, once the process of linking the data has been completed, the linked data are prepared as SUFs. Again, protecting respondents' anonymity is paramount and also challenging. The great advantage of administrative data – its high level of detail (e.g., on regional information) – increases the risk of the de-anonymization of the respondents, especially when additional information from the survey data is added. For this reason, the risk of de-anonymization has to be evaluated by considering the information available in each individual dataset as well as in the linked dataset. Typically, this means that the detailed information in the administrative data needs to be aggregated before the data can be published as SUFs.

The linkage and aggregation of administrative data across countries is further complicated by the substantial differences across countries in granting research access to administrative data and thus in their experiences with dealing with research interests. While some countries have little experience with and regulations for using administrative data for research purposes, other countries have a longer tradition of anonymizing and providing administrative data to researchers. In the past 10 years, Germany, for example, has founded several research data centers for administrative data to improve the data infrastructure for academics interested in analyzing such data. In other countries, however, the situation is more difficult.

The linking of survey data with administrative records is an important innovation for survey research. Both types of data have different advantages and disadvantages, many of which can be overcome by using them in combination with each other. Surveys are an important instrument within the social sciences as researchers can define the population of interest and the information to be collected. Furthermore, a survey can collect information on a variety of areas, such as respondents' behaviors, personality, attitudes, and expectations. However, survey data suffer from measurement and representativeness errors. Administrative data are very detailed to a degree that is impossible to collect in a survey. Furthermore, because the data are less affected by measurement and representativeness error, individual-level information can be collected more accurately. However, researchers cannot influence the content of the data available on administrative records since researcher interests are at best a secondary purpose. Often, administrative data used by researchers are just a by-product generated by organizations, institutions, companies, or other agencies in the process of monitoring, archiving, or evaluating the function or service they provide (Calderwood and Lessof 2009). Therefore, the combination of both survey and administrative data opens promising possibilities to conduct innovative research, partially because of how they each may offset some of the disadvantages present in the other.

This chapter has demonstrated the many challenges to be faced when embarking on linking survey data with administrative data across countries using the example of the SHARE linkage project. In particular, there are many differences across countries in data protection regulations and their interpretation in a data linkage context and in the experience that institutions holding relevant administrative records have in dealing with research interests. Once these challenges have been confronted, however, the gain in the detail and accuracy of the linked data offer exiting new research perspectives.

As the chapter demonstrates, research into possibilities for data linkage across countries is still in its infancy. In fact, SHARE is the first project endeavoring data linkage on a large cross-national scale. More research is still needed into administrative data sources that may be available for linkage across countries, the possibilities for direct data linkage and associated consent procedures in each country, as well as potential differential consent biases and their consequences for comparative analyses of linked survey and administrative data.

Areas for future research:

- The availability of different types of administrative data available for linkage with survey data across European countries
- Identification of similar administrative data sources to enable cross-national research
- Methods for increasing the proportion of survey respondents that consent to data linkage
- Potential differential consent biases and their consequences for comparative analyses of linked survey and administrative data
- Consequences for data linkage projects of the new EU directive on data protection, which is currently being negotiated at the European Union

References and Further Reading

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., & Zuber, S. (2013). Data resource profile: The Survey of Health. Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, 42(4), pp. 992–1001

Calderwood, L., & Lessof, C. (2009). Enhancing longitudinal surveys by linking to administrative data. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (Pp. 55–72). New York: Wiley.

Korbmacher, J. M., & Schmidutz, D. (2015). A note on record linkage in SHARE. In F. Malter & A. Börsch-Supan (Eds.), *SHARE Wave 5: Innovations & Methodology, MEA, Max Planck Institute for Social Law and Social Policy* (Pp. 57–63). Munich.

Schmidutz, D., Ryan, L., Müller, G., De Smedt, A., & De Smedt, K. (2013). Report about new IPR challenges: Identifying ethics and legal challenges of SSH Research. Deliverable D6.2 of Data Service Infrastructure for the Social Sciences and Humanities (DASISH). Retrieved from: http://dasish.eu/deliverables/.

Annelies G. Blom is a Professor at the Department of Political Science, School of Social Sciences and the Principal Investigator of the German Internet Panel (GIP) at the Collaborative Research Center "Political Economy of Reforms" at the University of Mannheim (Germany). Prior to this she founded of Survex – Survey Methods Consulting. Dr. Blom holds a BA from University College Utrecht (The Netherlands), an MPhil from the University of Oxford (UK), and a PhD from the University of Essex (UK). She is Conference Chair of the 2017 European Survey Research Association conference and member of the Methods Advisory Board of the European Social Survey, the Scientific Advisory Board of the GESIS Panel and the Review Board of the Norwegian Citizen Panel. Her research looks into sources of survey error, including measurement error, interviewer effects, and nonresponse bias.

Julie Korbmacher studied social sciences at the University of Mannheim and completed a PhD in statistics at the Ludwig-Maximilian-University Munich in 2014. She currently works as Senior Researcher at the Munich Center for the Economics of Ageing at the Max Planck Institute for Social Law and Social Policy where she is the head of the unit SHARE research projects. She is responsible for the record linkage projects of the Survey of Health, Ageing and Retirement in Europe (SHARE). Her main research interest is in survey methodology especially on interviewer effects and measurement error.

35

Linking Survey Data with the Catalist Commercial Database

Robert Blaemire

As a political corollary for the consumer marketing databases that have been developed for market research, politically oriented organizations have begun to aggregate data about individuals as well. Rather than simply collecting demographic or purchasing behavior data, these organizations attempt to identify people that may be politically active or amenable to becoming politically active given their past behavior. The types of records that these organizations collect are membership rosters for other groups or clubs such as the Sierra Club, in the case of Democratic organizations or the National Rifle Association, for Republican organizations.

The application that these organizations typically have in mind for such political databases is linking them with official voter registration and turnout records that are made publicly available by each state for non-commercial purposes. By combining these datasets, the notion is that political parties and politicians may be able to micro-target individuals that may have a higher probability of voting for some particular candidate or issue if they were contacted. These databases can be nearly as large as those generated by market research firms, for example, Catalist, the leading Democratic-leaning database creator has records for over 280 million individuals in their database.

R. Blaemire (✉)
Independent Consultant, Indiana, USA
e-mail: bblaemire@gmail.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_35

For many of the individuals represented in these databases there are extensive records of past voting behavior and group membership. So instead of mailing an identical political advertisement to every member of a state's Democratic party, the database organization can subset down to the individuals that perhaps haven't voted consistently in prior elections but who may be easily persuaded to turn out to vote. Going a step further, these organizations can then use the past group membership history for each individual to create subgroups that will receive different political advertisements based on what the organization believes will constitute a persuasive argument. Beyond political advertising, these resources have been used to assist with political fundraising, volunteer solicitation, and get-out-the-vote campaigns.

Recently, there has been considerable interest on the part of some researchers that make use of survey data for political applications, such as political scientists and polling organizations, to attempt to use these databases. Often, the goal is to validate turnout using a one-stop-shop approach rather than having to conduct the extremely varied and often challenging task of collecting voter files from each state after an election. In theory, using a political database organization to do this proverbial "heavy lifting" could make validation studies much easier and give survey researchers and pollsters a better sense for how to handle self-report data regarding voting behavior, which is widely believed to be over-reported.

However, there are significant concerns that must be addressed by those using these sorts of databases for academic research. First, they are constructed very similarly to the commercial consumer marketing databases and often even incorporate some of the demographic data purchased from these databases. This means that they may suffer from some of the same data quality issues associated with the consumer databases. Second, the political databases frequently have to make decisions about conflicting records for the same individual, rather than keeping all of the records, they use a statistical model to determine which record should be kept and then delete the conflicting records. Lack of transparency in this data cleaning process precludes a very important step necessary for serious researchers to evaluate the quality of the data in the database. Third, these databases are designed with a focus on coverage, the goal being to increase the number of people that they can categorize in their models; this may come at the cost of accuracy if many records with very poor quality data are aggregated with those records that may actually be of sufficient quality for serious academic or government research in such a way that the researcher cannot disentangle the two. Lastly, the transparency and rigor in documentation of procedures and changes to the data may be below ideal standards for academic or government research or may be held as proprietary information and not made available to researchers. It is also true, however, that academics could know a lot more about the people they may be contacting than they can by using anonymous surveys. The potential utility for researchers in these databases is certainly great, but much of it will likely go unrealized in the wider academic and government audiences until database organizations are able to provide greater transparency into their methods and allow researchers unencumbered access to evaluate data quality issues.

Areas for future research:

- Increasing transparency into the sources and quality of the data in private commercial or political databases
- Establishing and making public metrics for data quality for each private database
- Identifying the sources of errors and approaches to correcting these errors in private databases
- Developing testing methods to identify the error rates in private databases

I came to Washington when I was 18 years old to go to George Washington University and immediately went down – on the first day in Washington after my parents left – to the senate to volunteer to work for United States Senator Birch Bayh of Indiana and ended up eventually getting a job there and staying with him for the next 13 years until we lost in 1980 to Dan Quayle. During those 13 years I spent a lot of time learning about what the senate offered us in terms of services, and one of the things I learned early on was that there was a computer services division that compiled computer databases for members of Congress or members of the senate to do mailings. And I said, “Well great, what does the Bayh office have?” And they told me, “You have one list of 13,000 names that you mail once a year at Christmas time.”

At that point Bayh had been a senator for nine years. And I can tell you that in the next nine years we built a list that ended up having 2.8 million households in Indiana with 250 identification codes, and for two years in a row we were the biggest mailer in the United States Senate which a lot of people would argue was a big waste of money. For me it was a great badge of pride because not only did we learn tons and tons about building a database for constituents but also about the communication process about what worked, what didn’t work, you know, inviting people to town hall meetings, and if they would come; it was fascinating. The first time we ever did that. We sent a mailing out to the geography around where a town hall meeting was going to be held and went out with Senator Bayh for it, and there were probably four times the number of people who would fit in the room and many of them had the letter in their hand; it was a very satisfying experience.

After we lost the election in 1980, and realizing that I had to get a real job, I ended up getting hired by a firm that wanted to build voter lists to match them to the AFL-CIO membership for the AFL to find out who was registered and who was not. And so the first question came: what do you know about voter files? All I knew about voter files is that we didn't have them in Indiana and that we had to use commercial databases where we did not know in our 1980 campaign whether people were registered or not. We knew certain commercial aspects of them and we did geodemographic targeting using clear task cluster technology, if you're familiar with that, but we didn't know a lot of things we would like to know about them in politics. You know, number one is, are they registered?

I didn't know a lot but spent a lot of time learning, calling around the country, finding out what was computerized, and happened to be in a good time in this particular history because the country was starting to become computerized. And looking back I realize when we were doing this nobody else was doing this. This is kind of the beginning of a business, which has gotten quite a bit larger now, and we didn't know that we were inventing something. But I ended up working in this business with another company for about eight years and then started my own company, again, building voter file databases mostly for Democratic state parties and Democratic campaigns and becoming the vendor providing data to Democrats through the good offices of the party which was a good thing all the way around because it brought the party back into candidate services. It helped candidates have access to computerized voter data that they had not had access to before and it helped vendors reach candidates whom we could otherwise never reach. And that business grew.

I had my company for 17 years and then merged with Catalist in 2007. Catalist was formed by a combination of progressive organizations after the 2004 America Coming Together project. If you're familiar with that, this was an effort by the progressive community to try to coordinate its activities so basically the country would be covered better, people wouldn't be all going after the same voters, and you wouldn't have huge gaps of geography omitted. And at the end of the 2004 campaign when of course we did not win the presidency, a lot of organizations got back together with Harold Ickes who had been the Deputy Chief of Staff for Bill Clinton and is now and still the president of Catalist. The organizations got together and the common theme was a complaint about data that data wasn't available in a consistent, reliable form that all the organizations could use. I was providing a lot of data to ACT but a lot of organizations within ACT would not use our data because they were 501(c)(3)'s and they wouldn't use political party data which like most of mine was owned by political parties. So there was a constant struggle and nobody had the whole country. So Harold put together a bunch of investors that built Catalist and started building the whole country to make it available to the progressive community. So we sit in the center of a lot of progressive organizations like AFL CIO, League of Conservation Voters, Planned Parenthood, SCIU and a number of other such organizations that are almost like a public utility where they have online access to the data; they pull data out of our system and they put response data and IDs that they collect in the phone, in person, and by mail back into the database. We then can take those IDs and all the data that we have and use that soup to create models, and the increasing level of predictability of the models has become sort of the coolest new thing in politics and I think that's something that sort of has given us the edge in this particular business.

Are these databases perfect? No. But they are better now than they were when I got started. I used to build the main voter file for instance from 516 sources – towns all over the state. It was one of the hardest things you could ever try to do. We sent out letters to every town asking to send their voter file for the Democratic Party and right away 300 of them came in and those last couple hundred were killers. You know, you get a message; one of them says, "Call our town board between 2:00 and 3:00 on the first Tuesday of every month," and of course the line was always busy; you couldn't reach them.

And increasingly over those years, particularly after the Gore, Bush Florida debacle and Congress passed the Help America Vote Act, more and more states have become statewide databases, and they've gotten better and better. And the progress in this area has gotten better and I can just tell you that the error rate on companies like ours has also gone down. So while they're not perfect, they certainly are better than anything that has ever been available in this particular business.

Okay, so this is how we build the database basically. The component parts, we start with about 184 million voter file records in every state. Included in that is over a billion items of vote history or ballots cast. Another 80 million unregistered persons – 18 and above – come from a commercial database. And then we have specialty lists matched in things like aviation employees, doctors and nurses, farmers, teachers, hunters, and fishermen. Then we have the IDs from these response data that came from America Coming Together in 2004 and the Kerry IDs from 2004. This is the basis of the database.

Basically, we'll start with the voter data. The voter comes in most cases; all but two states we get from Secretaries of State. That's not true; all but three states we get from Secretaries of State. When I started with Catalist in 2007 there were only 28 states that were statewide. Now all of them are statewide but a couple of them are very restricted on who can get them and we cannot, so we have to get them by county. And we build the voter file multiple times per year.

Every state is at least twice and in the election year we did several states, oh man, 20 times. We were doing Colorado like every couple of days and the last month. Vote history again comes generally with the voter data. Sometimes it's separate; it depends on the state. Infogroup is our vendor. Infogroup is one of the largest list houses in the country where we get the unregistered people, and our unregistered people are basically made up of three types of persons, then when we match our file to their file, they give us back people who are 18 and above who don't match.

We then retain people who were on previous versions of the voter file but not on the current one. We call them dropped voters. And then we also have those who might have filed a change of address and moved into another state. They're available as an unregistered person in that state though they may be registered somewhere else – someone we refer to as a multiple person.

Then we go out when clients want different kinds of specialty lists that we can get. We go out and see where we can get publically available license files like gun owner licenses, hunter licenses, teachers, or farmers. There are some states where you can get these very, very easily and inexpensively; there are other states where you can't get them at all. I remember getting, when I had my company, a list of gun owners in Ohio for $10.00, and in other places, you know, you can't get lists almost of any kind for anything. So essentially very few of these are 100 percent national but they're all large enough that they give us the kind of data we need in order to be able to model in some of these areas.

The ACT IDs and the Kerry IDs are in a whole range or based on party preference, issue preference and candidate preference that were taken throughout 2003 and 2004. Membership records are also collected for any organizations that use Catalist, for all of our organizations, one of the things they do when they become a Catalist client is they take their membership and they match it in. So we have an indication of which of the voters or the unregistered people are members of a given organization. And so there are 54 million people who are a member of one or more of the organizations that are on the database.

Most of this is from the 2008 campaign, and this chapter refers to data mostly from the last presidential election. The most recent election is a little recent for us to have any data.

It's probably 54 million match with the voter file. There are 266 million contacts to almost 126 million individuals so we know whom each organization is contacting for different purposes throughout the campaign in 2008 and what responses came back. And then there are 264 models that had been built by our clients and ourselves scoring 2.5 billion records. In other words everybody is matched multiple times in the database.

So the traditional voter file gives us the name, contact information, address, phone number, gender, and age and sometimes includes party. There are 20 states that do not have party registration. Vote history is also included, and again one of the things about the Catalist database as opposed to if you went out and did it right now yourself, we would have a lot more history because we built it over time. We keep everything we have and we keep building on top of it so our history for most states goes back before 2000.

There are seven southern states that provide race on the voter file which again gives us a huge dataset to be able to model race. Now looking back at Lancaster Pennsylvania, in 2008 and looking at the traditional targeting where we would take precinct level targeting first where everybody in a precinct performs a certain way, that aggregate data used to be all that was available or precinct targeting, you know everyone was a target or they are not. And the voter file helps us exclude Republicans and 65 percent Democratic precincts as one example, or the voter file can also identify Democrats in heavily Republican precincts.

And then we go to the enhanced voter file which is clearly where we are now, where we have all the basic contact information and then commercial data. And the commercial data we get through Infogroup is hundreds of fields of commercial data items that are used for analytics purposes and are not necessarily available for clients for contact purposes. Census data as you know right now is very fresh but part of this is getting old. Historical IDs are not only from our own clients but many of their clients that they collect data from as well. And then the specialty data includes those particular kinds of lists that I was talking about – hunters, fishermen, and so on.

So the enhanced voter file, again Lancaster County, identifies registered Republicans, and then modeling predicts which Republicans strongly oppose the Iraq war, for example. And modeling also identifies Republicans who strongly support choice.

Some of the results come from modeling, having the issue responses. One of our clients for instance is a polling consortium, of pollsters that getting agreement to take their responses on a regular basis and upload them to our database. So we end up having thousands and thousands every week of actual responses on issues in a variety of categories. So when you get a dataset that's big enough and you think you can build a model from it, we go to work.

Now our database is comprised of 280 million voting age Americans with 700 fields of information so it's a lot of data. And it's a lot of data that you can create these models from. And I have to tell you, I look over the years of doing this crazy business. And in many respects we've always tried to predict what we're going to do. We try to say, is a Democrat going to be more likely to behave like a Democrat than a Republican? Does an older person care more about Social Security than a young person?

You know, I think we all agree that there are lots of assumptions we can make about voters, and the more we know about a voter the better we can communicate with them because the better likelihood is we have to make accurate assumptions about what they're likely to feel about. We're not dealing again with perfection; we're dealing with likelihoods. I've joked many times that if you try to get my wife's attention by starting a conversation about the Washington Redskins, you've lost her at the outset. But I'm paying attention. If you try to get a young person to pay attention to your campaign by telling them about the future of Social Security, they don't care.

There are certain things that may be obvious to us on a little bit of data but it can become more nuance and more detailed as we have more data. You know, I often use an example that if we bump into each other on the street, what do we talk about? The weather? We don't know anything about each other. I can perceive your likely gender, your likely age group, your likely race but I don't know anything else about you. Oh I find out that you're meeting at this meeting, I have something to talk about. I find out you're at a baseball park where I am; we can talk about baseball. The more I know about you, the better I can communicate with you and the better chance I have of persuading you to do something and that's what we're about.

We want to get the data as good as we can get it for practitioners, for people that need to use it for political communication. This data is restricted by law as well. We cannot use it for commercial purposes so you cannot use it to sell products. The only clients we have that are for profit clients are consultants and pollsters who agree that it is only being used by a permissible entity and that they have to tell us who it is. And I'll give you an example. I've had to turn down campaigns like a casino gambling campaign because the actual client in this case was Harrah's Casino and if it was a coalition of let's say nonprofit organizations that got together to promote or oppose a casino, that would be different. We could do that but we can't do it if it's indirectly commercial.

We have different restrictions we have to deal with in different states and this is not an easy process. Building these files is difficult and it's fraught with error. I like to believe that one of the advantages that I've been able to bring to Catalist having done this as long as I have is helping us do this better. And hopefully we do. I would like to give you some examples of something I think is pretty cool because for many years I've always recommended this to candidates and sold it to candidates because I believe it makes sense. I believe it works. And I've always believed that but now we have really sort of more empirical proof that it does work.

We can compare people who were contacted to those who weren't contacted to see if there's a difference in turnout and support. So we create models and these models, you know, are tools; they don't predict anything. They help us target better on an individual level. They improve our efficiencies in selecting universes and to expand and contract our universes and to tailor what messages we use. And the models we have are things like national partisanship.

We have a database that has everybody in the country who is registered by party and everybody in the country who votes in a partisan primary and tens of millions of people who have expressed a party preference on the telephone. That is a huge dataset to create a partisan model. It was created in 2007 or early 2008 and it has been revalidated and re-jiggered every year since then. It is probably the most popular, most widely used among models. Second to that is our voting propensity models which we've done every year. That is the one that is easiest to validate the likelihood to vote because we can see after the election whether they did or not.

Did we predict it and did they? You can tweak it very well and that is also used extensively but then we have more esoteric models you might say like likelihood of going to church at least once a week, likelihood of being a hunter, likelihood of being a gun owner, likelihood to support Obama in 2012 against Romney. A series of models like that includes an ideology model which allows us very well to slice and dice different partisan portions of the electorate, find Republicans for instance who have progressive views, and find Republicans who are pro-choice or believe that gay marriage is not a sin. It allows our clients to do a better job of slicing and dicing the electorate.

In 2008 again, 90 organizations used our data to make 335 million contacts, phone, mail and be at the door to 126 million individuals with another 7.4 million unique voter registration applications. We can identify where the contact was more Democratic than 2004 or more Republican and the contact, at least among our clients, was far more Democratic oriented in 2008.

Looking at a time lapse view of the data being pulled out and put into our system throughout the 2007–8 period. We looked at the timeline of the election from the Iowa caucus, New Hampshire primary, Nevada caucus, Florida primary, Super Tuesday. John McCain becomes presumptive nominee, the Obama race speech in Philadelphia, Pennsylvania primary, West Virginia primary, Hillary Clinton concedes, the Obama speech in Berlin, Democratic convention, Fanny Mae and Lehman Brothers collapse, John McCain suspends his campaign, the DOW falls 800 points, the last presidential debate, Obama's prime time TV address and the election. Talking about the early motivation for why Catalist was created and you look at the logic of the contact over time, you see Illinois is an island of non contact. That makes sense. Why would a campaign spend a lot of money contacting Illinois when not only is Illinois normally a blue state but the senator from Illinois was the candidate? Now you see a lot of contact around Chicago.

In terms of fundraising, California demonstrated same thing, an island of non contact. Fundraising in San Francisco, fundraising in L.A. Even in Texas you see fundraising in Dallas. New York, contact around the city for fundraising but not the rest of the state. So there was a logic. This coordination did seem to work. Then we looked at the state of North Carolina, the first state that would provide us a database with vote history uploaded right after the 2008 election.

So we look at the sporadic voters, or also known as surge voters, the voters who would vote in a presidential general election but not in an off-year election – in 2006 or 2010. And we see the turnout of the sporadic voters and the Democratic sporadic voters overall, compared to the turnout of the sporadic and the Democratic sporadic when they'd been contacted by one of the 90 organizations working with our data. The turnout went up. Similarly in Ohio, which is not a party registration state though they do vote in partisan primaries, it's tough to target because only about half of the people vote in primaries. I know because I was the vendor of the Ohio Democratic Party, the Kerry campaign and the coordinated campaign in 2004 and I argued tooth and nail unsuccessfully with them about how they should target in Ohio in 2004.

Now if we're targeting we want to go after the low turnout Democrats and the high turnout moderates using these models. I think you might agree with that logic. And then we look at what actually happened, there was more contact in the areas where there should have been in Ohio in 2008.

The tools and the coordination allow for a more effective campaign and Obama won in Ohio compared to 2004 where you can see the contact was all over the map and heavily focused on the Democrats and many cases the Democrats who were already going to vote which was a waste of money. So again I think it's just a graphic presentation of how this stuff can be made to work. Similarly North Carolina, looking at the turnout model across the bottom and the actual turnout on the bell curve, shows you that the turnout actually tracked the model fairly closely with one exception – this area where the predictor said they wouldn't vote but they did. So you might guess who and what kind of voters those are. Basically Black voters in North Carolina voted for Black candidate where they had not previously voted.

Again this shows the progressive footprint of voter registration contacts and IDs where you can also see those islands of noncontact. And it's fascinating when you talk to different people in the states and they'll readily confirm or/ and agree that this was exactly what did or didn't happen in their state. The first time I showed this was in Tennessee. And the whole reason to have me out there was to talk about the failure in the 2008 campaign to mount an effective campaign while the whole country was going Democrat and they were electing Republicans across the board. Similarly we use these models for advocacy where traditionally when you're advocating an issue we might have a mindset that says "Well, I'm from Indiana." You're not going to advocate pro-choice in Indiana. That would be wrong because there are pro-choice people, pro-environment people, and pro-healthcare people everywhere.

The question is, by modeling can you find them on an individual level instead of these typical geographic ways we might do it? I'll just run these at the same time and show you how we zero in on actual people who are healthcare reform activists used extensively in the 2009 healthcare debate or environmental activists. In Nebraska, not a big liberal state as you know, and the environmental activists there in Arkansas. And the reality is that the organizations that are pushing issues, environmental issues and healthcare issues could use this data to find people everywhere to pressure members of Congress on these issues. So you know there may not be a preponderance or a 50 plus one percent view on an issue in that state but you can still create pressure from voters to the members of Congress on these issues.

Similarly we use this in fundraising. We've created a progressive donor model that basically has been able to show organizations that if they use this they can spend less money per donor and net more by using the progressive donor model. We showed one Democratic organization that if they took their previous two million prospects that they mailed to, ran the donor model against it and cut off the bottom scoring half million pieces, huge savings, it would've only cut their donations by 16 percent. So essentially this stuff works. Is it the level of perfection that you want? I'm sure it's not. Is it better than it used to be? It sure is. I can tell you every year it gets better; every year it can be more precise.

The way we have academics using this around the country is we have an academic subscription where we have a number of universities, Stanford being one, that have access to our data. They have online access to the data to run unlimited counts, queries, and cross tabs of information. We have a number of colleges that send us survey data to match. You know, we can argue whether or not that's worth doing. I happen to think it is. I think that having a lot of information appended to that survey or starting with that information for your survey makes sense.

Regarding the pollsters who subscribe to this kind of data, we have 13 of the most prominent Democratic pollsters are subscribers; they want to make sure they're only sampling people and polling people whom they want to poll. Secondly their questionnaires aren't nearly as long because they know a lot of information about those voters before they talk to them. They don't need to ask the preliminary questions which make the cost of the survey so high. So they may spend more money to get the right sample and less money implementing the poll. But it's been very popular; it's used all over and they take advantage of the models as well.

So that's basically my spiel and this is my contact information if you're interested in any particular questions. But I happen to think this is a huge database; it's hugely valuable. I can't tell you that it's going to solve all the concerns that the academic community wants for your particular kind of studies but there's an awful lot of information on it and there's an awful lot of things you can do with it. For the purposes that we build it, is there a bias because we're progressive? There's a bias in whom we work with. The data would be built exactly the same if we were doing this for both sides or for the other side.

You try to do the best job of building a comprehensive database with as much information as possible so the clients can make the best possible decisions. You know there are lots of adages we throw out in this business saying, you know, you pick your cherries where the cherries are. I've always thought that was relevant because this helps you pick the right cherries and maybe know which ones are sour and which ones work. You know, there are a lot of things that this allows in political communication in the nonprofit and the academic worlds. A lot of things you can do with this that weren't able to be done before. You know, whether you're looking at the impact of redistricting, or looking at old districts, new districts, you want to look at the impacts of race on elections or turnout or race by income or race by gender.

I mean we run; we have people running cross tabs on this data all the time. Any time somebody calls about campaign work after the election it's kind of annoying given what we just went through. But I had somebody call yesterday and they needed to run some counts because there's a runoff in one of the congressional districts in Louisiana and I was pleased by the kind of questions she was asking. I was running cross tabs on ideology by party, ideology by turnout, and partisanship by turnout to let her figure out which kind of universe she wanted because what happens in politics generally is people have a budget that governs the decision-making process. I have enough, I can afford 10,000 pieces of mail or 50,000 phone calls. So you want to take the best 50,000 households or the best 10,000, whatever that number is. And this data helps you whittle those universes down to get to that universe.

Robert Blaemire has been an active participant in politics all of his adult life. Born and raised in Indiana, his career began at the age of 18 upon entering George Washington University. His employment with Senator Birch Bayh (D-IN) began in 1967 during Bob's freshman year and concluded with Bayh's unsuccessful re-election campaign in 1980 against Dan Quayle. Those 13 years saw Bob rise from volunteer worker to Office Manager to Executive Assistant in the Senate Office. His campaign experience with Bayh began by traveling with the candidate throughout the 1974 re-election campaign, continued with a variety of responsibilities in the 1976 Presidential campaign and, finally, Political Director of the 1980 campaign.

Also during this period, Bob completed his BA in Political Science and his MA in Legislative Affairs at George Washington University.

After the 1980 defeat, he founded a political action committee, The Committee for American Principles, an organization seeking to combat the growing role and influence of the New Right in political campaigns.

He began his career providing political computer services in 1982, eventually joining with and starting the Washington Office of Below, Tobe & Associates. During the three plus decades he has worked in this area of politics, his clients have included the Dukakis for President Campaign, both Clinton-Gore campaigns, Gore-Lieberman, Kerry-Edwards and dozens of Governor, Senate, Congressional, Mayoral and local campaigns and state Democratic Parties in 26 states.

In 1991, Bob created Blaemire Communications, a political computer services firm serving Democratic campaigns, progressive organizations and political consultants. During that time, Blaemire Communications managed more Democratic state party voter file projects than any other vendor. In 2003, Blaemire Communications introduced *Leverage*, an online voter file management system that was used extensively across the country in Democratic campaigns. In late 2007, Blaemire Communications was acquired by Catalist. He served there as Director of Business Development until leaving at the end of 2016 to pursue other consulting and writing opportunities.

Bob's first book will be published in 2017, a biography of his former boss, Senator Birch Bayh.

36

Challenges and Opportunities in Collecting Election Administration Data

Michael P. McDonald

Historically, survey respondents' self-reports of voting have exceeded official records of turnout by substantial margins. This led to long-standing skepticism about respondents' reports of their turnout behavior. Some suggest that respondents may think it's embarrassing to admit not voting, and therefore claim to have voted when they didn't. Others propose that the over-reports may be due to people who usually vote in every election but happened to overlook the fact that they uncharacteristically didn't vote in a particular instance. Because the official records are a gold standard, many researchers have argued in favor of replacing the self-report data with the official government records.

There are two general types of voter files that are frequently and reliably compiled into commercial and political databases: voter registration and absentee ballot requests. Vote history is a common component of these data as well but their reliability is variable and access to the general public can be restricted on a state-by-state basis. These voter files are used for a variety of research purposes ranging from vote validation studies and voter mobilization studies to racial bloc voting analyses for voting rights litigation.

Because of the gold standard status of voter files, they hold considerable promise, in theory, for linking with survey data that include variables relating to any of the data contained in these files. However, these files

M.P. McDonald (✉)
University of Florida, Gainesville, USA
e-mail: michael.mcdonald@ufl.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_36

are not without error. The ANES, after a long history of using voter files to conduct turnout validation studies, determined that the inaccuracies and errors were sufficiently egregious and the task of validation sufficiently challenging that they stopped conducting the validation studies altogether.

The challenges that the ANES encountered were not isolated. For voter registration alone there are common errors in the records themselves and issues with match algorithms across databases often lead to false positive and false negative matches, such that a person who did not vote may be identified as a voter or a person who did vote may be identified as a nonvoter. Furthermore, there is significant variability in how states handle voter files. Some states will only release voter files to political campaigns, meaning that they are unavailable to academic or government researchers; in the case of Florida, even federal programs are unable to get access to the absentee files. Similarly, Virginia forbids disseminating or sharing the vote history data outside of political campaigns. Sometimes these files can be exceedingly expensive to purchase outright or to time-consuming to collect, especially when vote history is only available from local election officials.

These complexities are part of the allure of using a pre-packaged database created by a commercial vendor such as Catalist or Aristotle. Allowing the organization to collect and manage all of the data can save researchers considerable time and effort. However, this outsourcing comes at the cost of the data becoming a "black box" with very little information about their quality or the validity of using them for rigorous research. Commercial vendors' business plans can be at odds of research agendas. Vendors want to provide an accurate current snapshot of voter registration to their clients for campaign purposes. The most current voter files available from election officials may exclude individuals that are have become ineligible since the last election, most often because they moved. The most current voter file may thus not contain records of all voters who participated in a given election. Commercial vendors may perform their own purging of records or enhance the data in other ways, such as attempting to match registered voters who moved with the record in their former home. Little is known about the potential biases that may arise from these practices. These potential errors are important to understand as a recent study claiming widespread noncitizen voting exemplifies: are five vote-validated self-reported noncitizens indicative of true rates of noncitizen voting, or are they artifacts of commercial vendors' data processing techniques?

Areas for future research:

- Identifying better methods for collecting and aggregating voter file data from states
- Working with database vendors and data clearinghouses to identify data files that are of sufficient quality for academic and government research
- Understanding better potential errors in database management procedures resulting from election administrators and commercial vendors practices

References and Further Reading

Achen, C. H., & Blais, A. (2010). Intention to vote, reported vote, and validated vote. Presented at the Paper presented at the meeting of the American Political Science Association, Washington, D.C.

Ansolabehere, S. D., & Hersh, E. (2010). The quality of voter registration records: A state-by-state analysis. *Working paper by the Institute for Quantitative Social Science at Harvard University and by the Caltech/MIT Voting Technology Project.*

Ansolabehere, S. D., & Hersh, E. (2012). Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. *Political Analysis, 20*(4), 437–459. http://doi.org/10.2307/23359641?ref=search-gateway:33f283072be3e852342fbb162894a90d

Cassel, C. A. (2004). Voting Records and Validated Voting Studies. *Public Opinion Quarterly, 68*(1), 102–108. http://doi.org/10.1093/poq/nfh007

Jackman, S., & Spahn, B. (2015). Unlisted in America. *Working Paper.*

McDonald, M. P. (2007). The True Electorate: A Cross-Validation of Voter Registration Files and Election Survey Demographics. *Public Opinion Quarterly, 71*(4), 588–602. http://doi.org/10.2307/25167582?ref=no-x-route:d7de9120976ad8af4cd61a27eca26eff

Dr. Michael P. McDonald is an Associate Professor of Political Science at the University of Florida, where he is affiliated with the University of Florida's Informatics Institute. His scholarship has appeared in top scholarly journals and he produces United States voter turnout statistics used widely by academia, the media, and policymakers. He has provided election-relating consulting to ABC, NBC, the Associated Press, the United States Election Assistance Commission, the Federal Voting Assistance Program, and the media's national exit poll organization.

37

Challenges with Validating Survey Data

Matthew K. Berent

Self-reports of voting behavior in many studies, including the American National Election Studies (ANES), are often about 20 percent higher than the official statistics on turnout. This has been a remarkably consistent finding since the 1960s, meaning that as actual turnout increases or decreases there are similar increases or decreases in self-reports of voting. Many people take this information to imply that the estimates generated by self-report survey data are not representative of the population because they don't match the "true" gold standard value provided by the government.

However, there are a number of different ways that this discrepancy could arise that warrants consideration. For example, a respondent might not answer the question because perhaps they believe voting behavior is private and sensitive and are unwilling to report their answer to the survey question, which leads to nonresponse and inaccuracy in the measure that is unrelated to the representativeness of the survey sample.

The second explanation is that there could be survey effects, meaning that participating in the survey could influence the voting behavior that the survey is attempting to measure. Many voting behavior studies involve very in-depth interviews before the election, and then an interview after the election. It is reasonable to expect that, after having answered a long battery of questions about politics before an election, respondents are made aware of

M.K. Berent (✉)
Matt Berent Consulting, Ohio, USA
e-mail: matt@mattberent.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_37

their political attitudes that were perhaps not as salient before the interview. This increased salience could then influence some respondents to go vote when they might not have otherwise. The result is that these surveys could inadvertently be biasing their own sample to over-represent voters.

This raises a third potential mechanism for higher levels of voting behavior in surveys than in the general population, it is possible that survey response propensity is positively associated with the propensity to vote, meaning that the same types of people who are willing to respond to a long interview about politics may be the same types of people who go out and vote; this would be another way that the sample of survey respondents could be biased from the general population, not demographically but behaviorally.

Lastly, it is entirely possible that respondents misreport their voting behavior, meaning that their answers are factually incorrect. This could be caused by respondents misinterpreting the question, misremembering the behavior, reporting based on their typical voting behavior rather than their specific behavior in the focal election, or intentional lying which might be due to social desirability bias if the respondent believes that it would reflect poorly on them to report not voting in the election to the interviewer.

Misreporting is the most commonly cited cause of the discrepancy between survey data and the official records, leading to the typical conclusion that official records should be used instead of self-reports. However, the approach of using official records is not completely straightforward. To employ this method, researchers must obtain the official records of turnout histories, match each respondent to his or her official record, and then determine the "correct" turnout status for each respondent.

There are two primary problems with the validation task. The first is that over 200 million people are currently eligible to vote in U.S. elections, thus the matching task is not trivial simply on the basis of the size of the databases involved. Second, the federal government does not aggregate individual voting records, meaning that researchers must collect the records from each individual state and there is substantial variability between states in the availability, accuracy, and content of these records.

Researchers are increasingly turning to commercial vendors of political and voter file databases to conduct the matching between survey data and the official voter records. However, these researchers face another set of challenges when working with these databases. For example, vendors vary in their level of transparency, which means that researchers need to identify and specify the level of uncertainty that they are willing to accept with regard to the provenance and quality of their data. Vendors are also typically unwilling to provide complete details about their matching algorithms, which makes it impossible to estimate the reliability or validity of the matches.

Some researchers opt to attempt in-house matching by obtaining the government registration and turnout records from a sample of states. Publicly available computer applications, such as LinkPlus from the Centers for Disease Control, can then be used to match survey respondents to their official government records. This approach requires more work on the part of the researcher but the benefits are complete control over the data cleaning and matching processes.

Recent evidence from an ANES in-house matching project indicates that two main factors are contributing to the discrepancy between self-reports and official records of voting behavior (Berent et al. 2016). The first is a downward bias in government records that occurs when the records incorrectly identify some respondents as having not turned out when in fact they did. These are all cases where a record match cannot be found, not cases where the self-report and government record data disagree explicitly. The second factor is an upward bias in the self-reports that occurs because people who participate in surveys are more likely to vote. These biases are additive, not offsetting, and seem to account for nearly all of the discrepancies between the self-report data and government records.

Areas for future research:

- Developing a better set of tools for researchers to conduct their own transparent validation studies without needing to use commercial vendors
- Identifying and understanding the correlates of survey participation and turnout to better understand the potential mechanism that drives the higher levels of turnout among survey respondents

Reference and Further Reading

Berent, M. K., Krosnick, J. A., & Lupia, A. (2016). Measuring Voter Registration and Turnout in Surveys Do Official Government Records Yield More Accurate Assessments? *Public Opinion Quarterly*, 80 (3), 597–621.

Matthew K. Berent received a PhD in Social Psychology from The Ohio State University in 1995. Early in his career, Matthew held faculty positions at Colgate University, Idaho State University, University of California – Santa Cruz, and

Florida Atlantic University. He has published research on survey question design, attitude theory, personality theory, and even audiology. More than 1,000 scholarly publications have cited his work.

Matthew is currently the president of Matt Berent consulting. His consulting clients include academics, high-tech companies, start-ups companies, manufacturing companies, and legal firms. He recently worked with researchers from Stanford University and the University of Michigan to develop and evaluate best practices for survey question design and coding open-ended data, to investigate problems with field interviewer behaviors, and to identify factors that cause distortions in time series data. You can find some of his work on these topics at his consulting website mattberent.com.

38

The Promise of Collaborative Data Sharing Across Research Sectors

Robert M. Groves

The social sciences seem to be at a key point in their history. Part of the apparent transformation is that research that has traditionally been the purview of academia and government is increasingly being done in the private sector. In the past, the social sciences, and survey research in particular, operated under a model in which researchers would create data, analyze them, and then make them publicly available for use, usually because the research had been federally funded. This paradigm seems to be shifting to more of a market-based approach where massive amounts of data are collected about individuals using both traditional social science methods and new methods that are often focused on reducing data collection costs. Meanwhile, the costs associated with generating the traditional high-quality social science data, such as those from nationally representative surveys, are at an all-time high.

Some of the most important features of government and academic produced survey data are that they are incredibly multivariate, they provide uniform and consistent measurement, and researchers control them. The researchers identify a specific construct that they want to measure and then invent the tools to measure these constructs, and when this process is done well these data provide can provide rich insights into the social world. Deep care is exercised in constructing each measurement, assuring that it reflects

R.M. Groves (✉)
Georgetown University, Washington, D.C., USA
e-mail: bgroves@georgetown.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_38

accurately the underlying construct. Further, many measurements are taken on the same unit, so that multivariate models can be estimated from the data. Finally, the measurements are administered in a consistent manner, achieving comparability over units.

Another benefit of these data is that they are typically made publicly available so that other researchers can use them an unlimited number of times. Sociology would not be the same field it is today without the General Social Survey, the same is true of political science with the American National Election Studies and economics with the Panel Survey of Income Dynamics. These specific studies provide the additional benefit of representing a long time series of data that allow longitudinal trends to be easily examined in consistent and valid ways. In most of the social sciences generations of quantitative scientists have "cut their teeth" on these datasets, and thousands of scholarly articles and books document the information extracted from these data. They have yielded important discoveries about the society and its economy.

However, the new data being generated in the commercial sector are often not nearly as rich or powerful as the survey data of the past. These data are much more likely to be univariate, lacking in consistency and uniformity, and "organic" in the sense that they are not controlled by researchers. Private sector firms disproportionately hold these data, and these firms often have no chief mission to benefiting society. There is nothing in their goals making these data freely available for government or academic research purposes. Rather, for an increasing number of these firms, these data are viewed as a potential source of revenue to be guarded carefully and kept proprietary.

At the same time as commercial firms are collecting a disproportionate amount of data, academic and government research costs have become unsustainable with no evident solution in sight. Research costs have risen especially for those methods that rely on the general public to supply self-reports in the data collection. In most developed countries, public participation among sampled persons is declining, forcing data collection agents to increase the number of repeated attempts to gain their participation. At the same time, the quantitative side of the social sciences seems to have won the day with nearly every institution in society now primarily using quantitative information. Taken together with the unsustainable costs of traditional research methods, this raises the specter of privatized and commercialized social science, at least to the extent that firms can profit from statistical information about the population.

"Big data" is a buzzword that is commonly used across research sectors and nearly everyone agrees there is unexplored potential value in these vast stores of data. However, there is little agreement about how government or academic researchers should acquire and process these data, and at a more fundamental social science level identifying appropriate inferential frameworks for these data is also contentious. Further complicating matters, commercial firms that hold much of these data are hesitant to make them available to government and academic researchers for a few reasons that need to be addressed by future work.

First, these companies are concerned about liability if their data are used for linking with government records and then somehow breached. This will remain a concern for these firms until legislation protecting them is put into place, so it will be important for a consortium of commercial, government, and academic organizations and individuals to attempt to make headway on these legislative efforts.

Second, these firms are concerned about increased attention to confidentiality because many of their data collection models depend on much of the population not paying attention to or caring about the confidentiality of their data. This concern needs more attention from the perspective of government and academic researchers who have typically approached confidentiality from the exact opposite position, often even going so far as to make confidentiality an argument for participation. Until common ground on data confidentiality can be addressed commercial data are likely to remain sparsely available.

Lastly, commercial firms are concerned that the use of their data may lead to a potentially profitable product being generated that they will be unable to make money because of the collaborative agreement. This is a concern that government and academic researchers will have a harder time addressing but one that needs to remain at the forefront of thought as these joint data use efforts are arranged. Thus, an important area in need of future research and funding is bringing government, academic, and commercial data interests together. This is a popular topic of conversation among social scientists but one that has not translated well into large-scale fundable research programs.

The starting point for this work needs to be a data integration research program that brings together the highly related but traditionally disparate interests of computer science, statistics, mathematics, and the social sciences. This big data consortium also needs to be structure in such a way that it addresses the concerns of commercial firms and appeals to the interests of government statistical agencies. This will fundamentally be an administrative endeavor aimed at building the infrastructure that will enable science to progress while also benefiting government and commercial interests. This is an area that National Science Foundation (NSF) could directly and immediately spur progress.

Continued funding of piecemeal disciplinary projects purporting to make progress in this area is insufficient and is likely to restrict significant progress more than enable it. Interdisciplinary funding at a higher level than the traditional programs of political science or sociology is needed to make substantive progress in this area. Status quo funding that does not address the larger infrastructure needs will result in continued reinvention of linkage, estimation, and dissemination techniques between disciplines that are unaware of the progress being made in each other. This endeavor is considerably larger than any one discipline and as such requires the development of a cohesive complementary set of research programs that are both interdisciplinary and cross-sector. Funding agencies like the NSF need to recognize the long-term benefits that these efforts could garner and make concerted efforts to incentivize work in this area.

Robert M. Groves is the Gerard J. Campbell, S.J. Professor in the Math and Statistics Department as well as the Sociology Department at Georgetown University where he has served as the Executive Vice President and Provost since 2012. Dr. Groves is a social statistician who studies the impact of social cognitive and behavioral influences on the quality of statistical information. His research has focused on the impact of mode of data collection on responses in sample surveys, the social and political influences on survey participation, the use of adaptive research designs to improve the cost and error properties of statistics, and public concerns about privacy affecting attitudes toward statistical agencies.

Prior to joining Georgetown as provost he was director of the US Census Bureau, a position he assumed after being director of the University of Michigan Survey Research Center, professor of sociology, and research professor at the Joint Program in Survey Methodology at the University of Maryland.

Dr. Groves is an elected member of the US National Academy of Sciences, the National Academy of Medicine of the US National Academies, the American Academy of Arts and Sciences, the American Statistical Association, and the International Statistical Institute.

Dr. Groves has a bachelor's degree from Dartmouth College and master's degrees in Statistics and Sociology from the University of Michigan, where he also earned his doctorate.

Section 4

Improving Research Transparency and Data Dissemination

39

The Importance of Data Curation

Steven Ruggles

Data curation is often overlooked when individual researchers or long-standing studies apply for new or continuing funding. Many researchers are more concerned about maximizing funds that can be applied toward collecting new data rather than appropriately integrating, disseminating, and preserving existing data. With new data collection costs at all-time highs, it may be time for researchers to begin investing in maximizing the utility of existing datasets. There are four major data curation challenges: (1) data integration, allowing interoperability across time and across data sources; (2) electronic dissemination, including online tools for data discovery and manipulation; (3) sustainability, including planning for long-run preservation and access, and the creation of persistent identifiers; and (4) metadata, which is machine-processable structured documentation.

Data integration is important because it enables time-series and comparative analyses to be undertaken without each individual researcher generating their own systems for harmonizing differences in datasets. Long-standing studies evolve over time, modifying their survey instruments, data collection procedures, processing protocols, and archiving methods. These factors make cross-temporal analyses challenging. Researchers are forced to adopt *ad hoc* data integration solutions, which leads to inconsistencies between researchers. Even

S. Ruggles (✉)
University of Minnesota, Minneapolis, USA
e-mail: ruggles@umn.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_39

modest investments in data integration can reduce redundant effort and minimize the potential for introducing error.

For example, in 1991 microdata (individual-level data) samples from the Decennial Census existed for 10 census years dating back to 1850. These data had 10 different codebooks associated with them totaling 2,880 pages of codes and over 1,000 pages of ancillary documentation. Furthermore, of these 10 codebooks, 9 of them used different coding systems for most variables. The Integrated Public Use Microdata Series (IPUMS) project harmonized the codes, generated a consistent record layout, and integrated the documentation, all with no loss of information. In 1999 IPUMS was expanded to include 100 national statistical agencies around the world. The result is that over 500 censuses and surveys have been harmonized for 75 countries and nearly one billion person-records, and this is set to double over the next 5 years.

These approaches and best practices for harmonizing datasets are also being extended to data with different formats and from different scientific domains. For example, many of these individual-level datasets can be appended with community-level data relating to the physical environmental contexts in which individuals were and are living. This might include data about land-use statistics, land cover from satellite imagery, raster data (gridded values linked to spatial coordinates), or historical climate records. Data integration approaches like this can add richness to already fine-grained datasets by providing important contextual variables (Fig. 39.1).

Some of these principles have been applied by surveys such as the American National Election Study (ANES), General Social Survey (GSS), and the Panel Study of Income Dynamics (PSID) to their datasets, meaning that researchers can access cumulative data files that have been harmonized. However, there is very little documentation of cross-temporal comparability or consistency issues. Future funding and research efforts on data integration can focus on designing surveys to maximize interoperability across time and between surveys by developing and implementing standard classifications and coding systems; this would increase the value and use of all types of survey data designed with these features.

The second major topic relating to data curation is dissemination. Efficient data dissemination is important because the large investment of scarce resources for social science infrastructure can only be justified if the data are widely used. Even modest investments in data dissemination can reduce redundant efforts and stimulate new kinds of research. Increased emphasis on ongoing improvements to dissemination methods and platforms is necessary to maximize the value of survey data.

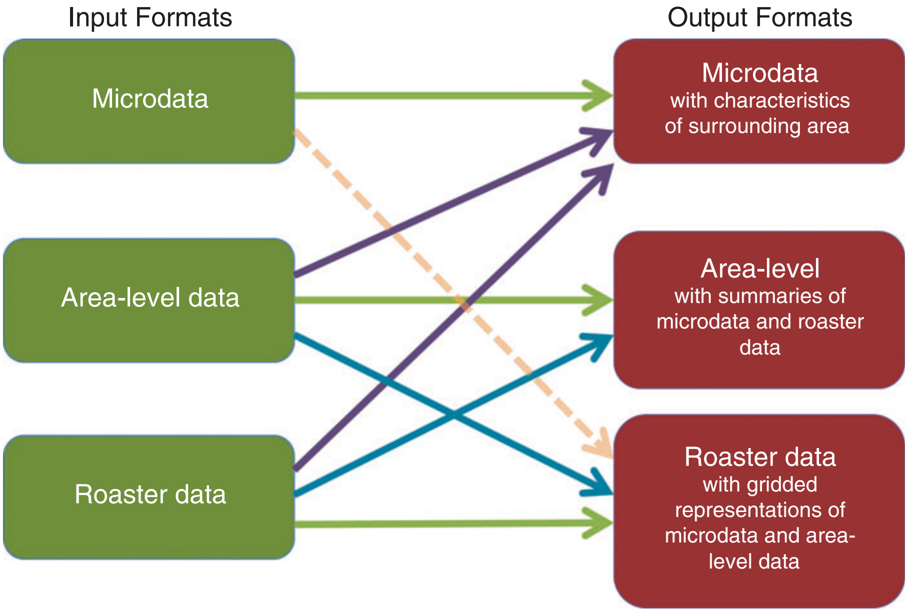

Input Formats
Output Formats
Microdata
Area-level data
Roaster data
Microdata with characteristics of surrounding area
Area-level with summaries of microdata and roaster data
Roaster data with gridded representations of microdata and area-level data

Fig. 39.1 A schematic of common data input and output formats

Best practices for dissemination suggest that data storage and access platforms undergo constant improvements to keep pace with technological advancements and user expectations. For example, IPUMS is now onto its fourth major generation of dissemination tools; these are driven entirely by structured metadata stored in a relational database. The tools include the ability to manipulate the data easily, online analyses, and the use of Data Documentation Initiative (DDI) standards for metadata, which is discussed further next.

Most survey data dissemination platforms, including ANES, GSS, and PSID, do not follow these best practices. The platforms have not kept pace with technological advancements and are for the most part running on antiquated systems that allow efficient discovery or exploration of variables. Furthermore, the metadata for these surveys are typically in PDF documents, effectively locked away and not actionable, which severely hinders the efficiency of any variable identification or analysis tools. The following are some key statistics for the metadata of the “Big 3” NSF-funded surveys:

- ANES
  - ~245 PDF files
  - ~35,000 non-actionable pages

- GSS
  - ~250 PDF files
  - ~25,000 non-actionable pages
- PSID
  - ~230 PDF files
  - ~50,000 non-actionable pages

To begin making these data more available and useful the ANES, GSS, and PSID should implement metadata browsing functionality that will enable researchers to quickly identify variable availability across time and to directly access survey instruments and supporting documentation. Better dissemination software for selecting subsets of variables is also important, especially for PSID. These surveys also need to implement machine-understandable metadata – even in the form of Stata, SAS, and SPSS setup files, the current approaches tend to be inadequate or non-existent. Future funding and research to begin optimizing the usefulness of existing survey data is critically important.

The third key point about data curation relates to sustainability. Funding agencies, universities, and government statistical agencies have invested hundreds of millions of dollars in data infrastructure. The greatest value from this infrastructure and the data generated by it comes from the power for cross-temporal analysis. In general, the older the data, the more rare and irreplaceable it is – and the greater value it has to researchers. This suggests that researchers, survey organizations, and sources of research funding must all work together to ensure effective stewardship of the nation's statistical heritage.

Broadly speaking, sustainability with regard to data curation means preserving data, especially old data. This preservation must also include steps to maintain broad access to the data. Sustainability can be thought of in three key areas: organizational, financial, and technical sustainability that will enable the data to be migrated effectively as technology changes. The ongoing surveys need to decide whether or not they are functioning as effective data archives. If they are, then they need to implement a formal preservation plan, if they are not then they need to arrange a cooperative agreement with a major archive such as the Inter-University Consortium for Political and Social Research (ICPSR). Current integration with ICPSR is varied the major NSF surveys, ANES share the most with 57 of 68 data files being in ICPSR, GSS is next with 16 of 36 files, and the PSID has only 4 of about 100 files in ICPSR.

To move toward a sustainable data curation model surveys also need to expand the use of persistent identifiers. In the current state of affairs identifying information for variables, datasets, and documentation is prone to changing, and in many cases have changed multiple times and exist in multiple iterations. Survey organizations need to adopt consistent standards for unique identifiers that will be consistent over time. The NSF could speed this transition by requiring that funded datasets provide a persistent identifier such as a Digital Object Identifier (DOI) or another recognized persistent identifier, and present a plan for long-term maintenance of the DOIs.

The fourth and final topic for data curation is metadata. As has been mentioned elsewhere in this report, metadata is a class of information that relates to and describes that focal data – it is "data about data." Examples of metadata include technical reports, survey instruments, interviewer instructions, show cards, and other documentation of the survey process or variables. The key point for survey organizations is that machine-actionable structured metadata is vital to enabling efficient data integration and dissemination. Without good metadata, software must be custom-built for each dataset, which is terribly inefficient and not a good use of resources. Metadata following DDI standards are also necessary for preservation purposes. In short, all aspects of data curation require good metadata in order to function properly. Currently all major survey data archives use DDI including ICPSR, UK Data Archive, Odum Institute, and the Institute for Quantitative Social Science. Unfortunately, the ANES, GSS, and PSID likely have over 100,000 pages of documentation that need to be converted to a structured metadata format, which means that significant investment and future work is required. Future research aimed at reducing the costs of conversion through the development of "smart" automated processes could have great impact on this process.

Areas for future research:

- Improving the state of metadata held by the ANES, GSS, and PSID
  - Identifying a set of common standards for generating, structuring, and disseminating new metadata that conform with DDI standards
  - Identifying projects that might reduce the conversion costs for existing metadata
- Improving data integration within and between surveys
- Establishing standards for dissemination across surveys

References and Further Reading

Anderson, M. J., Citro, C. F., & Salvo, J. J. (2012). IPUMS (Integrated Public Use Microdata Series). In *Encyclopedia of the U.S. Census*. 2300 N Street, NW, Suite 800, Washington DC 20037 United States: CQ Press. http://doi.org/10.4135/9781452225272.n94

Block, W., & Thomas, W. (2003). Implementing the Data Documentation Initiative at the Minnesota Population Center. *Historical Methods: a Journal of Quantitative and Interdisciplinary History, 36*(2), 97–101. http://doi.org/10.1080/01615440309601219

Data Documentation Initiative (2012). Data documentation initiative specification. Available at http://www.ddialliance.org/

MacDonald, A. L. (2016). IPUMS International: A review and future prospects of a unique global statistical cooperation programme. *Statistical Journal of the IAOS, 32*(4), 715–727. http://doi.org/10.3233/SJI-161022

Paskin, N. (2016). Digital Object Identifier (DOI ®) System. In *Encyclopedia of Library and Information Sciences*, Third *Edition* (3rd ed., Vol. 6, pp. 1586–1592). CRC Press. http://doi.org/10.1081/E-ELIS3-120044418

Ruggles, S., Sobek, M., King, M. L., Liebler, C., & Fitch, C. A. (2003). IPUMS Redesign. *Historical Methods: a Journal of Quantitative and Interdisciplinary History, 36*(1), 9–19. http://doi.org/10.1080/01615440309601210

Sobek, M., & Ruggles, S. (1999). The IPUMS project: An update. *Historical Methods: a Journal of Quantitative and Interdisciplinary History, 32*(3).

Steven Ruggles is Regents Professor of History and Population Studies and Director of the Institute for Social Research and Data Innovation at the University of Minnesota. Ruggles received his PhD from the University of Pennsylvania in 1984, followed by a postdoctoral National Research Service Award at the Center for Demography and Ecology of the University of Wisconsin. Ruggles has published extensively in historical demography, focusing especially on long-run changes in multigenerational families, single parenthood, divorce, and marriage, and on methods for analysis of historical populations. Over the past 25 years, Ruggles has developed large-scale data infrastructure for economic, demographic, and health research. He is best known as the creator of the worlds largest population database, the Integrated Public Use Microdata Series (IPUMS), which provides data on billions of individuals spanning two centuries and 100 countries.

40

Improving the Usability of Survey Project Websites

David L. Vannette

Website usability is not a subject that is unique to survey research; however, the principles and best practices for making websites usable have often been ignored or poorly implemented by survey organizations. There are two primary reasons that funding agencies, survey organizations, and researchers should care about improving survey website usability. First, and most importantly, is dissemination; by making survey data and documentation easy to access through survey websites we can increase the use of survey data and broaden the influence and importance of survey research. Second, improving survey website usability is important because it will make secondary research easier to conduct because both data and documentation will be easily located, accessed, and used. This will make the process of using survey data more accessible and transparent.

Broadly speaking, website usability means creating the conditions for efficient and effective visits to the website in question. This means satisfying a number of important criteria that define what usability means in terms of broad goals, including

- Providing relevant and easily accessible information
- Enabling learnable routines
- Designing efficient paths

D.L. Vannette (✉)
Department of Communication, Stanford University, CA, USA
e-mail: dave.vannette@gmail.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_40

- Creating memorable patterns
- Minimizing user errors
- Satisfying user experience

Breaking each of these items down further; in terms of providing relevant and easy accessible information to the users, it's important that those things go together. Information on survey websites should be both relevant to user needs and easy to access. Easy access implies using standardized formats and machine-actionable documentation and data so that users do not need to search through thousands of pages of PDF documents to find a single piece of information.

Next, it is key for survey website structure and organization to enable learnable routines. This refers to having similar tasks on a website follow similar routines so that users can apply learning from the browsing or searching that they did for data to the browsing or searching that they do for documentation. Similarly, survey websites need to design efficient paths and this refers to minimizing the number of clicks or the number of search terms that users need to use before they find what they came to the website for. The key is to minimize the distance between when users arrive at the site and when they get to the data or documentation that they came looking for. Also in this vein is the best practice of creating memorable patterns in the ways that webpages are structured. Important elements should not move around on web pages or disappear on some pages and come back on others.

Lastly, in terms of general goals, survey websites should also seek to minimize user errors. An important goal is to make it really hard for users to make mistakes. Users should not get 10 minutes into a task and realize that they should have caught a mistake that they made first arrived at the website. This means making survey websites clear and intuitive for the average user. Taken together, these broad goals are aimed at helping users to have a satisfying experience.

A very small proportion of all websites satisfy the goals outlined earlier; however, survey websites can be particularly egregious in not achieving some of these usability goals. A good starting point for assessing website usability is asking what their users want, and in the case of survey websites this is thankfully not a challenging task. Survey website users are typically looking for three broad categories of information: (1) data, (2) process and data documentation, and (3) examples of how these data are used such as publications.

After identifying the goals of website usability and the particular needs of survey website users, the next step is to identify best practices that can guide the design decisions implemented by survey websites. There are many best practices that have been defined for website usability but they fall into two broad categories of principles: (1) optimizing for memory and (2) optimizing for visual perception.

In terms of optimizing for memory it is important that websites standardize task sequences. This means that finding variables and documentation on a survey website should follow the same procedure across all possible pages and tasks. Standardized task sequences allow users to apply what they learned about conducting a task in one part of the website to all other parts of the website where that task is performed. At the highest level, there are two classes of tasks that users can engage in on a website: (1) searching and (2) browsing. Survey websites need to ensure that no matter what part of the site the user is in these two tasks are structured the same. This means that once a user knows the task sequence for searching for data they also automatically know the task sequence for searching for documentation. This is an area that survey websites such as the ANES, GSS, and PSID need considerable development and improvement.

Reducing user workload and designing websites for working memory limitations are also key best practices. For example, the GSS website uses a system called Nesstar for online data retrieval and analysis; this system requires knowledge of a 51 page user guide and this knowledge is not transferable because it is specific to the GSS and even there only to the Nesstar portion of the website. This design approach does not minimize user workload or acknowledge the working memory limitations that new users have when coming to the site. Survey websites should take care to display directly usable information to users. For example, on the starting pages for datasets important survey design and dataset features should be prominently displayed so that users can quickly evaluate whether this is indeed the dataset that they wanted to access. The websites for the ANES, GSS, and PSID can all improve in applying these best practices.

Survey websites often need to house tens of thousands of pages of technical documentation and code books; however, this information is most commonly archived in PDF documents that are not machine actionable – which in a very basic sense means that a website search does not include searches inside these documents. These documents are also poorly linked and can be nearly impossible to use without extensive prior knowledge of the particulars of the documentation. Best practices for survey documentation suggest using organizing features such as linked tables of contents and structured documents with labeled sections. Tying in with the importance of accessibility with regard to data curation, it is also important for all documentation to adhere to widely accepted standards such as DOI and DDI (see report section on data curation). These are practices that have not been adopted by survey websites such as ANES, GSS, and PSID.

A number of other best practices are well exemplified by the ICPSR website. These best practices include

- Optimizing information density
- Aligning elements consistently
- Using fluid layouts that maximize screen size and resolution
- Helping users search in addition to browsing

One of the biggest challenges facing survey organizations is determining how to allocate scarce resources. Survey staff often view every dollar spent on something other than data collection as money that could have been spent better. This perspective can be seen in the current state of the websites of the ANES, GSS, and PSID where usability, data curation, and transparency best practices are often ignored or poorly implemented. Survey website usability requires increased investment and attention in order to optimize all of these best practices. If survey organizations are not able to implement these best practices on their own then they should partner with other archives such as ICPSR to ensure that their data are made more available and usable to a broader array of potential users.

Lastly, best practices dictate conducting usability studies when making changes to websites. The current state of many survey websites seems to imply that little, if any, systematic usability research was conducted. Even small-scale studies can indicate important problems or areas where improvements could easily be made. On a similar note, these survey organizations have considerable expertise in survey data collection and it might be useful for them to consider applying that expertise toward identifying what their data users like and dislike about the current websites and features that could be implemented to improve the user experience.

In summary, there are many best practices for improving website usability (Bailey et al., 2011) but many of these practices and principles are not implemented by survey websites. This presents an opportunity to for future work by these survey websites on improving their structure and design. Improving website usability also provides a platform with which survey organizations can display other best practices in the areas of data curation and transparency.

Areas for future research:

- Identifying practical approaches for the websites of the ANES, GSS, and PSID to increase their usability
- Conducting usability research on existing websites to identify easily implemented changes to improve usability
- Conducting surveys to identify what features users find helpful or problematic

Reference and Further Reading

Bailey, R. W., Barnum, C., Bosley, J., Chapparo, B., Dumas, J., Ivory, M. Y., et al. (2011). *Research-Based Web Design and Usability Guidelines* (p. 292). Usability.Gov.

David L. Vannette is a PhD Candidate in the Department of Communication at Stanford University, Stanford, CA, USA. and Principal Research Scientist at Qualtrics LLC, Provo, UT, USA.

41

The Role of Transparency in Maintaining the Legitimacy and Credibility of Survey Research

Arthur Lupia

Many scientific researchers are motivated to produce credible scientific contributions. At the same time, many face pressures to produce and publish findings as quickly as possible. With such pressures comes the temptation to ignore critical assumptions and to view rigorous documentation and reporting as secondary tasks to be completed later. I contend that limited introspection and documentation on the part of researchers threatens the credibility and legitimacy of scientific research as a whole (also see Lupia 2008 and Lupia and Alter 2013).

Survey research is implicated in such problems. Every survey-based claim follows from arguments whose conclusions depend on the truth-values of important assumptions. Some of these assumptions are statistical and others relate to the design characteristics of the study. Limited introspection about these key methodological assumptions puts scholars at risk of promulgating false claims.

Indeed, in several disciplines, scholars improperly interpret or have been unable to replicate highly visible survey-based claims. Collectively, these failures undermine the aggregate credibility of survey research. There are a number of constructive steps that researchers and organizations can take to increase the credibility of many survey-based endeavors. Such efforts should focus on ways to improve research transparency, meaning documenting the

A. Lupia (✉)
University of Michigan, Ann Arbor, USA
e-mail: lupia@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_41

processes, decisions, and actions that convert labor and capital into survey data and then into survey-based knowledge claims.

Research transparency has two components: production transparency and analytic transparency. Production transparency implies providing information about the procedures that converts labor and capital into data points. Today, many surveys provide little to no information about these procedures, which include but are not limited to sample selection, respondent recruitment, question selection, question wording, response options, directions to interviewers, post-interview processing, and coding of open-ended responses. Many users, in turn, passively accept survey data as accurate rather than a product of practices that can produce biases and inaccuracies if misunderstood. When users cannot access information about how survey data was produced, their ability to make accurate survey-based claims can be severely inhibited.

Analytic transparency implies a full account of how a researcher drew conclusions from a given survey, clearly mapping the path from the data to specific empirical claims. There are many examples of published work in which such paths are neither public nor reproducible. By meeting a small number of transparency requirements at production and analytic stages, scholars can protect their own credibility and contribute to the credibility and perceived legitimacy of survey research more generally.

Analytic transparency means making every step in a research process, including errors, open to evaluation. Methods courses across the sciences emphasize the value of documenting research processes in sufficient detail that an outside researcher could replicate the finding. Beyond simple replicability, our goal is to provide fellow researchers and outside onlookers all possible information available to evaluate the value of the contribution of the research; not just the information that leads to judgment in one particular direction. In terms of best practices, this implies that researchers should consider relying on a system of "lab books" that document all decisions made regarding a research project and are then made available when the project is published.

Effective data citation and curation practices are critical components of efforts to increase availability of production and analytic materials and to increase incentives for transparency. If data and documentation are not made readily and available to users in easily accessible formats, the benefits of increased transparency are reduced. Survey organizations should document all decisions that have any influence on data or data quality and make this documentation easily accessible online using best practices for data curation and website usability.

Funding agencies such as the NSF as well as academic journals can also play important roles in increasing research transparency. By developing and implementing data sharing and documentation transparency requirements, funding agencies can boost incentives for transparency in many different kinds of research. Journals, in turn, can require authors to deposit data and documentation to trusted digital repositories as a condition of publication (see, e.g., DART 2015). If funders and journals followed a similar model, it is likely that scholars across many disciplines would adopt new transparency norms and practices, which would benefit all of the social sciences.

Public trust in, and the public value of, scientific research depend on the extent to which scientists are willing and able to share the procedures by which their knowledge claims were produced. Scientific integrity of this kind corresponds to a kind of utter honesty, bending over backwards to make certain that evaluations of the work are made with complete information. For scientific researchers, the sources of our distinctive legitimacy are rigor and transparency. The elevated social value of scientific research comes from the ability of others to evaluate the meaning of our findings. Increasing transparency is a critical step to the continued production of social science in the public interest.

Areas for future research:

- Identification and publication of a set of transparency standards for social science research funded by NSF
- Creation and implementation of institutional means to support greater transparency

References and Further Reading

Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors. (2015). Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors. *Political* Science *Research and Methods*, *3*(03), 421–421. http://doi.org/10.1017/psrm.2015.44

Lupia, A. (2008). Procedural transparency and the credibility of election surveys. *Electoral Studies*, *27*(4), 732–739. http://doi.org/10.1016/j.electstud.2008.04.011

Lupia, A., & Alter, G. (2013). Data Access and Research Transparency in the Quantitative Tradition. *PS: Political Science and Politics*, *47*(01), 54–59. http://doi.org/10.1017/S1049096513001728

Arthur Lupia is the Hal R. Varian Professor of Political Science at the University of Michigan and research professor at its Institute for Social Research. He examines how people learn about politics and policy and how to improve science communication. His books include *Uninformed: Why Citizens Know So Little About Politics and What We Can Do About It*.

He has been a Guggenheim Fellow, a Carnegie Fellow, is an American Association for the Advancement of Science Fellow, and is an elected member of the American Academy of Arts and Sciences. His awards include the National Academy of Sciences Award for Initiatives in Research and the American Association for Public Opinion's Innovators Award. He is Chair of the National Academy of Sciences Roundtable on the Application of the Social and Behavioral Science and is Chairman of the Board of Directors for the Center for Open Science.

42

Evidence-Based Survey Operations: Choosing and Mixing Modes

Michael Bosnjak

Survey methodology is inherently pragmatically oriented (e.g., Bosnjak and Danner 2015, p. 309; Goyder 1987, p. 11): how to sample and recruit respondents to reduce coverage and sampling errors, how to operationalize concepts to reduce measurement error, and how to minimize the differences between those who responded from those who did not on all variables of interest (i.e., nonresponse bias) are generic guiding questions in survey methodology (Dillman, Smyth and Christian 2014; Groves et al. 2011). Accordingly, survey methodology is aligned toward generating a body of knowledge to support survey operations in making decisions about how to design, prepare, implement, and post-process survey-based projects. In doing so, survey methodology is structurally similar to other disciplines being committed to generating the best empirical evidence and using it to guide actions. The very first and most notably among the scientific disciplines championing such a view are the health sciences: evidence-based medicine, defined as the "conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al. 1996, p. 71), has decisively contributed to replace anecdote, case study insights, and theoretical reasoning by focusing on the findings from systematically synthesized, high-quality experimental research. The achievements that evidence-based medicine has produced include, among

M. Bosnjak (✉)
ZPID – Leibniz Institute for Psychology Information, Trier, Germany
e-mail: michael@bosnjak.eu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_42

others, establishing the Cochrane Collaboration[1] to independently collate and summarize clinical experiments, setting methodological and publication standards,[2] building infrastructures for pre-registering experiments[3] and for developing and updating guidelines for clinical practice,[4] and developing knowledge resources and courses for teaching evidence-based medicine.[5] Since the early 1990s, other research disciplines have followed, resulting in scientific movements such as evidence-based education (Pring and Thomas 2004), evidence-based management (Rousseau 2012), evidence-based criminology (Farrington et al. 2003), and evidence-based software engineering (Dybä et al. 2005).

A key common characteristic of those committed to the evidence-based paradigm is the belief that empirical findings should be classified by their epistemological strength. Furthermore, the strongest types of empirical evidence, namely meta-analysis of randomized experiments, are taken to be the considered primary basis for decision-making. Meta-analysis can be described as a set of statistical methods for aggregating, summarizing, and drawing inferences from collections of thematically related studies. The key idea is to quantify the size, direction, and/or strength of an effect, and to cancel out sampling errors associated with individual studies. Consequently, meta-analytic findings are typically characterized by greater precision and validity than any individual study. In contrast to meta-analytic evidence being regarded as most conclusive, study types being classified as “mid-level quality” or lower should be treated with due caution for drawing conclusions or making decisions. Such mid-level evidence encompasses sporadic randomized experiments and quasi-experiments not yet synthesized in a quantitative fashion, and findings from study types not allowing for causal inference, such as narrative reviews, correlational studies, and case reports.

While survey methodology has produced a large number of (quasi-)experimental findings, the use of meta-analytic techniques to systematically summarize their outcomes is still underdeveloped: a bibliographic database search using Web of Science and Google Scholar performed in April 2016 has identified about 50 meta-analyses relevant for survey operations. This overall output matches the number of published meta-analysis in a top-tier psychology journal such as *Psychological Bulletin* within just two annual volumes.

[1] http://www.cochrane.org/

[2] http://prisma-statement.org/

[3] http://www.alltrials.net/

[4] http://www.guideline.gov/

[5] http://www.cebm.net/

Therefore, survey methodology as a discipline appears remarkably reluctant relative to other fields to apply existing quantitative tools to synthesize evidence systematically.

The overall aim of this chapter is to briefly sketch how the pursuit of an evidence-based mindset by focusing on meta-analytic evidence is instrumental in deriving best practice recommendations for survey operations. We focus on one perpetual decision call in survey operations, for which a considerable number of meta-analyses is available, namely how to choose or combine survey modes. We will first review the meta-analytic evidence available and then derive actionable recommendations for survey operations. Finally, we will sketch avenues for future meta-analytic research on mode effects and mixed-mode issues.

Mode Choice and Mode Combination: Meta-Analytic Findings

For much of the twentieth century, mixing survey modes was uncommon (Dillman and Messer 2010). With a goal of optimizing costs, response rates, and measurement quality, researchers increasingly began to combine multiple modes of data collection (DeLeeuw andToepoel, this volume; Groves et al. 2011, Chapter 5.4). For instance, by starting with the least expensive mode (e.g., Web), and then moving to more expensive ones to collect data from nonrespondents (e.g., mail, telephone, or even face-to-face interviews), costlier approaches are successfully applied to fewer cases, optimizing the overall expense for a survey project. In longitudinal surveys, starting with expensive face-to-face interviews as an attempt to maximize initial cooperation rates, and then moving to more reasonably priced Internet-based data collections, became usual practice in probability-based panel surveys (e.g., Blom et al. 2016; Bosnjak, Das and Lynn 2016). Moreover, the optimal survey mode may even depend on the type of survey questions asked (de Leeuw 2005; Dillman et al. 2014, Chapter 11): interviewer-administered surveys are assumed to be ideal to gain cooperation and to ask nonobtrusive questions. Then, to reduce item nonresponse and to yield valid answers for sensitive questions, follow-up surveys using self-administration seem most appropriate.

Remarkably, it remains unclear if the recommendations that have developed as best practice are actually supported by meta-analytic evidence. The first type of meta-analytic evidence relevant for deriving recommendations on the choice and combination of survey modes is on the synthesized effect within various survey modes in head-to-head mode comparison studies.

Here, primary studies varying the survey mode as the independent variable, and assessing its effect on measurement-related and/or representation-related dependent variables, are aggregated. The specific outcome variables considered here were, for instance, the average mode-specific degree of social desirability bias (e.g., De Leeuw and Van Der Zouwen 1988; Richman et al. 1999; Gnambs and Kaspar 2016), response rate differences between modes (e.g., Hox and De Leeuw 1994; Lozar Manfreda et al. 2008), and differences in terms of item nonresponse (de Leeuw 1992). The meta-analytic findings based on studies comparing the effects of mode will be summarized in the next two paragraphs.

The second type of meta-analytic evidence is on relating two distinct mixed-mode designs to the two pertinent quality outcome dimensions (mixed-mode effects meta-analyses). Within this category, the overall effect of mode preferences are of interest to survey methodologists, that is, meta-analytic findings focusing on studies summarizing the impact of allowing respondents to choose between two or more response modes offered concurrently (e.g., Medway and Fulton 2012). In addition, another noteworthy type of mixed-mode effect involves combining modes sequentially, either varied to follow-up on nonrespondents, or varied according to data collection occasions in multiple-wave surveys (e.g., Shih and Fan 2007). The key findings from these mixed-mode meta-analyses will be sketched in the third and fourth paragraphs to follow.

Mode Effects and Measurement-Related Survey Quality

Previous meta-analytic research on mode effects has primarily focused on the degree of social desirability as a measurement-related survey quality indicator. Depending on the operational definition of social desirability, three distinct eligibility criteria to identify studies about the impact of mode can be employed, each discovering unique patterns of findings:

A first approach focuses on studies having administered validated social desirability scales and instruments to detect misreporting tendencies within two or more modes concurrently. The two most recent meta-analyses based on such studies have found no substantial differences between Web-based and paper-based modes of survey administration (Gnambs and Kaspar 2016, based on 30 experimental studies), and computer-assisted and paper-and-pencil administration (Dodou and Winter 2014, based on 51 studies that included 62 independent samples), paralleling the findings from older meta-analyses (Dwight and Feigelson 2000; Tourangeau and Yan 2007; Richman et al. 1999). However, when comparing computerized self-administration with interviewer-administered modes (e.g., telephone or face-to-face surveys), social desirability bias scale estimates are about 0.5 standard deviations larger for interviewer administration (Richman et al. 1999).

In the second approach, socially desirable responding is viewed as deliberate over-reporting of favorable characteristics, such as, conscientiousness, and under-reporting of unfavorable traits, such as neuroticism. Gnambs and Kaspar (2016) have pooled 10 experimental mode comparisons between Web-based and paper-and-pencil administration of the Big Five personality traits, and 28 studies comparing modes on self-reported psychopathological conditions. No substantial differences were found between the two self-administered modes considered.

A third approach defining social desirability involves under-reporting of sensitive behaviors, such as illicit drug use, delinquency, sexuality, and victimization. A recent meta-analysis having employed this conceptualization of social desirability by Gnambs and Kaspar (2015) revealed that computer-assisted surveys resulted in prevalence rates of sensitive behaviors that were about 1.5 times higher than comparable reports obtained via paper-and-pencil questionnaires; for highly sensitive issues, such as sexuality and illicit drug use, this mode effect was even larger. A re-analysis of six papers on mode-specific differences for self-reported sensitive behaviors summarized in Tourangeau and Yan (2007, Table 3), including interactive voice responses and audio computer-assisted self-interviews, corroborated the finding that computer-assisted modes yield self-reported frequency measures of sensitive behaviors of about 0.2 standard deviations larger on average compared to paper-and-pencil administration. If self-administration modes (computerized or mail surveys) are compared with interviewer-administered modes, such as telephone surveys (de Leeuw 1992, p. 32), or face-to-face surveys (de Leeuw 1992, p. 31; Weisband and Kiesler 1996), reports of sensitive behaviors are consistently lower in interviewer-administered modes. Only two published papers report comparisons within the category of interviewer-administered modes (telephone and face-to-face; De Leeuw and Van Der Zouwen 1988; de Leeuw 1992, p. 28), and neither of those detected substantial differences in terms of reporting sensitive behaviors.

Measurement quality indicators other than social desirability have been only occasionally considered in past meta-analyses. De Leeuw (1992) found no substantial mode differences between mail, face-to-face, and telephone surveys for factual questions which could have been checked against public records. Ye et al. (2011) found, based on 18 experimental comparisons, that telephone respondents are more likely to select the most extreme positive option than respondents to Web, mail, or IVR (Interactive Voice Response) surveys but not respondents to face-to-face interviews. The differences found in choosing the most extreme option amounted between 6 and 11 percentage points on average.

Taken as a whole, the meta-analytic evidence suggests that within each of the two classes of mode: (1) self-administered modes (computerized modes, including Web-based surveys; paper-and-pencil administration, including mail surveys) and (2) interviewer-administered modes (face-to-face surveys; telephone surveys), there are no substantial differences in terms of social desirability bias. The propensity to report sensitive behaviors is the notable exception to this generalization. The highest prevalence rates for sensitive behaviors are typically estimated in self-administered computerized survey settings, followed by paper-and-pencil administration. Interviewer-administered modes typically yield the lowest prevalence rates for sensitive behaviors. Based on this evidence, survey researchers might consider using self-administration for topics being substantially affected by social desirability bias. For collecting data on sensitive behaviors, using computerized data collection approaches (e.g., Web surveys) appears to be the most suitable option.

For other, less frequently investigated measurement quality indicators, the few available meta-analytic findings point to comparable levels of validity for answers to factual questions between mail, face-to-face, and telephone modes. Moreover, the most positive category is chosen more often in interviewer-administered Computer-Assisted Telephone Interview (CATI) surveys compared to self-administered ones. Therefore, positivity biases are best counteracted by using self-administered modes.

Mode Effects and Representation-Related Survey Quality

Despite the fact that response rates are only loosely related to nonresponse bias (Groves and Peytcheva 2008), the majority of primary studies about the impact of mode on representation-related outcome variables have used response rates as outcomes. These primary studies have been summarized in six meta-analyses. The evidence from these studies indicates the following ranking: face-to-face surveys typically achieve the highest response rates (De Leeuw and Van Der Zouwen 1988; de Leeuw 1992; Hox and De Leeuw 1994), telephone surveys the next highest (De Leeuw and Van Der Zouwen 1988; de Leeuw 1992; Hox and De Leeuw 1994), followed by mail surveys (Hox and De Leeuw 1994), and lastly Web-based surveys (Lozar Manfreda et al. 2008; Shih and Fan 2007).

Specifically, Hox and De Leeuw (1994) found, based on 45 primary studies, average completion rates for face-to-face surveys amounting to about 70 percent, for telephone surveys 67 percent, and for mail surveys 61 percent. For Web surveys, Lozar Manfreda et al. (2007) estimated, based on 24 studies reporting 45 experimental mode comparisons, a lower response rate of about 11 percent on average compared to other survey modes. Also, 27 out of 45 experimental comparisons in Lozar Manfreda et al. (2008) involved Web versus mail comparisons. In both Hox and De Leeuw (1994) and Lozar Manfreda et al. (2008), the average effects were not homogeneous, that is, they were moderated by specific study characteristics. Most notably, in Hox and De Leeuw (1994), response to face-to-face and telephone surveys went down in the period covered (1947–1992), and the response to mail surveys went up slightly. In Lozar Manfreda et al. (2008), the difference between Web and other modes was moderated by the number of contacts: The more contact attempts to recruit nonrespondents, the larger the discrepancy between Web and other modes became. A recent study by Mercer et al. (2015) explicitly addressed the moderation effects between modes and incentives on response rates. Based on 55 experiments containing 178 experimental conditions, Mercer et al. (2015) found that while prepaid incentives increased response rates to face-to-face, mail, and telephone surveys, postpaid incentives did work for the two interviewer-administered modes only. Moreover, the dose-response relationships proved to be mode-dependent in the prepaid condition: the expected increase in response rate for a $1 prepaid incentive is 6 percentage points for mail surveys; promising the same amount over the phone yields an expected improvement of only 1 percentage point. In practice, survey researchers might consider these estimated mode-dependent incentive elasticities and decide whether or not the to-be expected increase in response rates by using prepaid incentives for the mode considered will most likely pay off.

Only two meta-analyses have addressed representation-related outcome variables other than response rates: De Leeuw and Van Der Zouwen (1988) and de Leeuw (1992) did not find mode difference for the amount of item nonresponse between mail, face-to-face, and telephone surveys.

Mixed-Mode Effects and Representation-Related Survey Quality

The three meta-analytic research summaries about mixed-mode effects (Mavletova and Couper 2015; Medway and Fulton 2012; Shih and Fan 2007) on representation-related survey quality indicators have exclusively compared self-administered modes and used response rates and breakoff rates as outcomes.

Medway and Fulton (2012) found, based on 19 experimental comparisons between either a mail survey or a survey in which participants were given the choice of responding by mail or by Web, that providing a concurrent Web option in mail surveys lowers response rates: doing so decreases the odds of response by 12.8 percent as compared to a mail-only survey.

Shih and Fan (2007) have synthesized both experimental and nonexperimental findings reported in 52 studies. From those studies that offered respondents options for either Web and mail surveys simultaneously, the overall response rate was 25 percent. In those studies where respondents were sent mail surveys first and offered the option for Web survey mode in follow-up reminders, the overall response rate was 42 percent. In those where respondents were sent Web surveys first and then offered the option for mail survey mode in follow-up reminders, the overall response rate was 47 percent. When optimizing response rates is an issue, the findings reported in Medway and Fulton (2012) and Shih and Fan (2007) suggest not to offer Web and mail modes concurrently. Instead, they should be combined serially.

A recent meta-analysis by Mavletova and Couper (2015) based on 39 independent samples suggests that allowing respondents to choose the visual format in mobile Web surveys (PC versus mobile optimized Web surveys) seems beneficial for decreasing breakoff rates. If respondents have an opportunity to select their preferred visual format, the odds of breakoff rates are decreased by OR=0.62 ($p < 0.05$) compared to the surveys in which respondents are initially assigned to a mobile Web survey mode. Furthermore, mobile-optimized Web surveys decrease the odds of breakoffs among mobile respondents by OR=0.71 compared to nonoptimized Web surveys.

Recommendations for Survey Operations and Avenues for Future Meta-Analytic Research

The meta-analytic findings described in this chapter support a number of recommendations for best practices for survey operations. To reduce social desirability bias and the tendency to endorse positive statements, (computer-assisted) self-administered modes should be used. To gain cooperation, interviewer-administered surveys should be considered. Because item non-response rates do not seem to differ substantially between modes, optimizing the completeness of survey data from actual respondents is not an issue of mode, but other factors, which are outside of the scope of this chapter.

Some of these recommendations are partly based on meta-analysis performed more than 20 years ago. In the meantime, numerous new primary studies emerged that have the potential to change the meta-analytic results, indicating a strong need for updated research syntheses. Such research updates, called cumulative meta-analyses (Borenstein et al. 2009, Chapter 42), decisively contribute to understanding the robustness of findings across time and would be a valuable addition to the survey methodology literature.

The meta-analytic evidence suggests that the following survey implementation approaches should be abandoned: (1) offering mode choices should be avoided for the initial invitation to participate in mail and Web surveys because this survey implementation strategy decreases response rates. (2) Promised incentives should be avoided in favor of using prepaid incentives, and the use of prepaid incentive value denominations should be tailored toward known mode-specific findings regarding their effectiveness. In other words, incentive response rate elasticities appear to be mode-specific, a relationship that deserves further research attention.

Surprisingly, despite the prevalence of documented mixed-mode surveys, no meta-analytic evidence seems to be available on combining self-administered and interviewer-administered data collection modes. Future quantitative summaries on such issues would be a valuable addition to this literature, helping to derive recommendations for how to combine modes in an optimal way. The nonexisting meta-analytic evidence regarding mixed-mode effects for interviewer-administered surveys is also striking. Filling these gaps in existing knowledge is a promising avenue for future research.

Moreover, there is a lack of meta-analytic studies on the impact of (mixing) modes on estimates of biases in terms of measurement and representation, not just on proxy variables or partial elements of bias, such as response rates (Groves and Peytcheva 2008).

Overall, given the few research syntheses in a core area of survey research such as mode effects, the use of meta-analyses and systematic reviews as a basis for deriving best practice recommendations for survey practice seems – compared to other areas such as the health and behavioral sciences – in an infancy stage. To achieve a similar level of professionalism in deriving evidence-based recommendations, survey methodology might consider establishing the pre-conditions for promoting high-quality research syntheses by

- Establishing a central portal to gather research synthesis, helping survey methodologists to make informed decisions about survey implementation choices (mode choice and mode combination effects being one of them)

- Setting methodological and publication standards for research syntheses in survey methodology
- Developing and updating guidelines for primary studies in terms of topics, that is, research areas for which there is no or insufficient evidence, and the correspondingly recommended research design (e.g., "ideal" research design templates to relate mode effects to specific survey errors and biases)
- Developing knowledge resources and courses for teaching evidence-based survey methodology

References and Further Reading

Blom, A.G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S., & Krieger, U. (2016). A comparison of four probability-based online and mixed-mode panels in Europe. *Social Science Computer Review, 34*(1), 8–25.

Borenstein, M., Hedges, L.V., Higgins, J.P.T, & Rothstein, H.R. (2009). *Introduction to Meta-Analysis*. Wiley, Chichester, UK.

Bosnjak, M., & Danner, D. (2015). Survey participation and response. *Psihologija*, 48(4), 307–310.

Bosnjak, M., Das, M., & Lynn, P. (2016). Methods for probability-based online and mixed-mode panels: Recent trends and future perspectives. *Social Science Computer Review*, 34(1), 3–7.

De Leeuw, E. D., & Van Der Zouwen, J. (1988). *Data quality in face to face interviews: A comparative meta-analysis. Telephone survey methodology*. Russell Sage Foundation, New York.

De Leeuw, E.D. (1992). *Data Quality in Mail, Telephone and Face to Face Surveys*. TT Publikaties, Plantage Daklaan 40, 1018CN Amsterdam.

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, 21(2), 233–255.

Dillman, D.A. & Messer, B.L. (2010). Mixed-mode survey. In J.D. Wright & P.V. Marsden (Eds.), *Handbook of Survey Research* (2nd edition) (pp. 551–574). San Diego, CA: Elsevier.

Dillman, D.A., Smyth, J.D., & Christian, L.M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Dodou, D., & De Winter, J.C.F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, 36, 487–495.

Dwight, S.A., & Feigelson, M.E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement*, 60(3), 340–360.

Dybä, T., Kitchenham, B.A., & Jorgensen, M. (2005). Evidence-based software engineering for practitioners. *Software, IEEE*, 22(1), 58–65.

Farrington, D.P., MacKenzie, D.L., Sherman, L.W., & Welsh, B.C. (Eds.). (2003). *Evidence-based crime prevention*. Routledge.

Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, 47(4), 1237–1259.

Gnambs, T., & Kaspar, K. (2016). Socially desirable responding in web-based questionnaires a meta-analytic review of the candor hypothesis. Assessment, Online first: http://dx.doi.org/10.1177/1073191115624547.

Goyder, J. (1987). *The silent minority: Nonrespondents on sample surveys*. Westview Press.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. A meta-analysis. *Public Opinion Quarterly*, 72(2), 167–189.

Groves, R.M., Fowler Jr, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561). John Wiley & Sons.

Hox, J. J., & De Leeuw, E. D. (1994). A comparison of nonresponse in mail. telephone, and face-to-face surveys. *Quality and Quantity*, 28(4), 329–344.

Lozar Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., Vehovar, V., & Berzelak, N. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *Journal of the Market Research Society*, 50(1), 79.

Mavletova, A. & Couper, M.P. (2015). A meta-analysis of breakoff rates in mobile web surveys. In D. Toninelli, R. Pinter & P. de Pedraza (Eds.), *Mobile research methods. Opportunities and challenges of mobile research methodologies* (pp. 81–98). Ubiquity Press, London, UK.

Medway, R. L., & Fulton, J. (2012). When more gets you less: a meta-analysis of the effect of concurrent web options on mail survey response rates. *Public Opinion Quarterly*, 76(4), 733–746.

Mercer, A., Caporaso, A., Cantor, D., & Townsend, R. (2015). How much gets you how much? Monetary incentives and response rates in household surveys. *Public Opinion Quarterly*, 79(1), 105–129.

Pring, R., & Thomas, G. (2004). *Evidence-based practice in education*. McGraw- Hill Education, UK.

Richman, W.L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, 84(5), 754–775.

Rousseau, D.M. (2012). *The Oxford handbook of evidence-based management*. Oxford University Press.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal*, 312(7023), 71–72.

Shih, T.H., & Fan, X. (2007). Response rates and mode preferences in web-mail mixed-mode surveys: A meta-analysis. *International Journal of Internet Science*, 2 (1), 59–82.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883.

Weisband, S., & Kiesler, S. (1996). Self disclosure on computer forms: Meta-analysis and implications. Proceedings of the ACM CHI 96 Human Factors in Computing Systems Conference April 14–18, 1996, Vancouver, Canada, pp. 3–10. http://www.sigchi.org/chi96/proceedings/papers/Weisband/sw_txt.htm

Ye, C., Fulton, J., & Tourangeau, R. (2011). More positive or more extreme? A meta-analysis of mode differences in response choice. *Public Opinion Quarterly*, 75(2), 349–365.

Michael Bosnjak is director of ZPID – Leibniz Institute for Psychology Information in Trier, Germany, and Professor of Psychology at the University of Trier. Before joining ZPID in July 2017, he was team leader for the area Survey Operations at GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany, and Full Professor for Evidence-Based Survey Methodology at the University of Mannheim, School of Social Sciences. Between 2013 and 2016, he was the founding team leader of the GESIS Panel, a probabilistic mixed-mode omnibus panel for the social sciences. His research interests include research synthesis methods, survey methodology, and consumer psychology.

43

Best Practices for Survey Research

David L. Vannette

In this section, we present an extensive series of recommendations for best practices based on the chapters in this volume and other sources.

Probability versus Non-Probability Sampling

- To make inference from a sample to a population requires assumptions. These assumptions are well understood only for samples drawn with known or calculable probabilities of selection. Thus, only such samples are widely accepted (Baker et al. 2013).

Response Rates

- Researchers should carefully document nonresponse using American Association for Public Opinion Research (AAPOR) categories and make efforts to understand the correlates of nonresponse so that users of the data are alerted to potential nonresponse bias (Groves and Couper 2012).

D.L. Vannette (✉)
Department of Communication, Stanford University, CA, USA
e-mail: dave.vannette@gmail.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_43

- Nonresponse bias is rarely notably related to nonresponse rate, so reducing nonresponse bias should usually be a more important goal than maximizing response rates (Groves and Peytcheva 2008).
- Refusal conversion is critical for increasing response rates and also may affect reductions in nonresponse bias (Groves and Heeringa 2006).
- Effective training of interviewers and other survey staff has been demonstrated to increase response rates (Conrad et al. 2012; Durrant et al. 2010).
- Careful design of survey materials influences response rates. These materials include invitations, self-administered questionnaires, and other items that respondents see (Vicente and Reis 2010).
- Incentives, particularly prepaid monetary incentives, can notably increase response rates (Singer et al. 1999).
- Multiple contact attempts may increase response rates substantially (Keeter et al. 2000).
- Advance letters notifying household members of their invitation to participate may increase response rates (Link and Mokdad 2005).
- To best understand and minimize nonresponse bias, response rates should be modeled as a stochastic process based on the association between response propensity and the characteristic being estimated (Brick 2013).

Data Collection Modes (e.g., Telephone, Web, Face-to-Face, etc.)

- Response rates and data quality are often highest with face-to-face surveys and lower in other modes, such as telephone interviewing, paper questionnaires, and Internet data collection (Hox and De Leeuw 1994).
  - To minimize non-coverage in telephone surveys, both cellphones and landlines should be called (Brick et al. 2007).
  - Collecting data via multiple modes is an effective way to reduce costs and increase response rates (De Leeuw 2005).
  - To maximize comparability between modes in a mixed-mode design, a unimode design for the questionnaire can be implemented, avoiding design features that are not replicable in every mode used (Dillman 2005).
  - To minimize total error in mixed-mode surveys, use a responsive design framework – taking advantage of every measurement error reduction affordance available in each mode, even if not replicable between modes (Groves and Heeringa 2006).

Survey Incentives: Paying Respondents to Participate

- To increase response rates, use prepaid incentives (Church 1993).
- In interviewer-mediated surveys, promised incentives typically have no effect on response rates (Singer 2002).
  - Additional incentives paid (differentially) to recruit people who initially refused to become respondents may increase response rates among these reluctant people and reduce nonresponse bias (Singer 2002; Singer et al. 1999).

Using Responses from Proxies (e.g., Household Members Other Than the Respondent)

- Proxy reports can be used to reduce costs in face-to-face interviewing (Boehm 1989).
- Proxy reports can best be used to measure observable information about the target respondent, rather than unobservable information such as attitudes (Cobb and Krosnick 2009).

Improving Question Design to Maximize Reliability and Validity

- To reduce satisficing behavior, minimize task difficulty and maximize respondent motivation (Krosnick 1999).
  - To reduce respondent frustration, questionnaire design should follow conversational norms to the greatest extent possible, for example, via adherence to the Gricean Maxims (Grice 1975).
- Optimal approaches to questionnaire design include

a. When asking questions with numeric answers or categorical questions with unknown universes of possible answers, ask open-ended questions rather than closed-ended questions.
b. When life forces respondents to make choices outside of the survey context, study those choices via ranking questions rather than by asking people to rate individual objects.

c. To ensure consistent interpretation of rating scales, label all options with words and not numbers, and ensure that the range of options covers all points on the underlying continuum.
d. Break bipolar rating scales into multiple questions via branching and offer unipolar rating scales without branching.
e. To optimally measure bipolar constructs, use seven-point response scales.
f. To optimally measure unipolar constructs, use five-point response scales.
g. Do not offer "don't know" response options and encourage respondents who volunteer a "don't know" response to instead offer substantive answers.
h. Rotate the order of response options on rating scales and categorical questions across respondents, except when doing so would violate conversational conventions about order (e.g., always put positive response options before negative options, order unipolar scales from most to least).

   - When selecting words, choose ones that are in frequent use in popular discourse, have few letters and syllables, have only one primary definition (instead of two different frequently used definitions), and are easy to pronounce.
   - Avoid asking people to remember the opinions they held at prior times – answers to such questions are often wrong (unless the research goal is to assess what people believe their opinions used to be).
   - Avoid asking people to explain why they thought or behaved in particular ways – answers to such questions are often wrong (unless the research goal is to assess why people think they thought or behaved in particular ways).

Perception of Visual Displays and Survey Navigation

- In visual displays, avoid abbreviations, notation, and jargon (Kosslyn 2007).
- In visual displays, present no more than four visual elements (Kosslyn 2007).
- When presenting sequences of visual displays, identify pieces of new information with distinctive colors or sizes (Kosslyn 2007).

  - Visual emphasis on certain words should be used sparingly and bold typefaces and colors should be used instead of using all upper

  case words, italic typefaces, or underlining to create emphasis (Kosslyn 2007).
  - Rely on the perceptual grouping laws of similarity and proximity to organize information presentation by grouping similar or related items together (Kosslyn 2007).
  - Make important elements different from surrounding elements by making the former larger, brighter, or more distinctively colored (Kosslyn 2007).

Pretesting Questionnaires

- Implement “cognitive pretesting” of all questions and change question wordings to eliminate misinterpretations or eliminate respondent confusion or lack of clarity of meaning (Willis 2005).
  - Conduct multiple iterative rounds of cognitive pretesting to be sure that questionnaire revision does not introduce more error (Willis 2006).
  - Behavior coding of interview administration can identify problematic questions that require revision.

Interviewer Deviations from the Standardized Survey Interviewing Script

- To minimize interviewer deviations from scripted behavior, extensive training and monitoring should be implemented (Schaeffer et al. 2013).
  - To identify problematic behavior by interviewers, monitoring should include audio recordings and verbatim transcription (Schaeffer et al. 2013).

Coding Open-Ended Survey Questions

- Multiple coders should code the same text independently, based on detailed written instructions, and the level of agreement between coders should be examined to confirm sufficient reliability. If reliability is insufficient, instructions may need to be revised, or coder training or supervision may need to be enhanced.

  - All materials used in coding should be made publicly available, including raw responses.

Confidentiality and Anonymity of Survey Participation and Data

- Surveys should generally promise anonymity or confidentiality to respondents, which may produce higher participation rates and to minimize intentional misreporting (Lelkes et al. 2012).

Respondent Attrition from Panel Surveys

- Repeated interviewing of the same respondents affords valuable analytic opportunities for studying change and causal processes (Schoeni et al. 2012).
- Substantial effort should be expended to minimize panel attrition (Olsen 2005).
- Respondents who fail to provide data at one wave of a panel may rejoin the panel subsequently, so efforts to re-recruit such respondents are worthwhile.

  - Substantial paid incentives can often minimize panel attrition (Creighton et al. 2007).
  - Using information about individual respondents, tailor their incentive offers to be maximally attractive while not being unnecessarily high (Olsen 2005).

Paradata: Data About Survey Processes and Contexts

- Collect as much information about the process and context of the survey data collection as possible (Kreuter 2013).

  - To evaluate operational aspects of web survey design features, collect timings for every item and page, which can be useful for analyzing cognitive processes (Heerwegh 2003).
  - Use call record data for interviewer administered surveys to study nonresponse bias (Durrant et al. 2011).

Using Interviewer Observations About Respondents

- To collect information about non-respondents, have interviewers record observable characteristics (age, gender, race, etc.) (Peytchev and Olson 2007).

  - Have interviewers report evaluations of respondent engagement, perceptions of respondent honesty, perceptions of respondent ability (West 2012).

Leave-Behind Measurement Supplements

- In-person interviewers can leave paper questionnaires with respondents, to be returned to investigators by mail, to collect supplementary measures (Harvey 2002).

Capturing Moment-by-Moment Data

- Experience sampling and ecological momentary assessment (EMA) allow for the collection of data in real time from respondents to supplement survey data (Stone and Shiffman 1994).
  - To maximize accuracy of data collected using experience sampling or EMA, use very short recall periods (Stone et al. 1999).
  - Time-use diaries can be used to collect information on activities in real time (Bolger et al. 2003).

Collecting Biological Data via Biomarker Measures

- To collect measures of health, use biomarkers (Schonlau et al. 2010).
- To minimize costs of biomarker collection, dried blood spots can be collected (McDade et al. 2007; Parker and Cubitt 1999).

  - To collect extremely rich data, DNA can be collected, though analysis costs are high (Schonlau et al. 2010).

Specialized Tools for Measuring Past Events

- To improve respondent recall of past events, especially with very long reference periods, use an event history calendar method to structure the interview (Belli 1998).
  - To maximize the accuracy of respondent recall when using event history calendars, interviewers should be trained cognitive retrieval strategies such as sequential retrieval, parallel retrieval, and top-down retrieval (Belli et al. 2007).

Linking Survey Data with Government Records

- Linking survey data to official records on the same individuals has promise, but matching processes may not be as effective as is needed.
  - Materials used in matching should be made public to analysts, so the effectiveness of the matching can be evaluated.
  - If survey reports differ from measurements of the same phenomena in official records, consider the possibility that the records are incorrect rather than the self-reports (Presser and Traugott 1990).

Preserving and Maximizing the Accessibility of Existing Data

- After a survey is conducted, the raw data and all instruments used to collect the data (e.g., the questionnaire, show cards, interviewer training manuals) should be preserved in electronic form and made available to the community of scholars (Ruggles et al. 2003).
  - Older surveys that exist on paper records only should be scanned and made electronically available in machine-actionable structured formats.

  - When related surveys have been conducted over time, the data from such surveys should be archived in a common, integrated format to facilitate comparison between surveys (Ruggles et al. 2003)
  - To move toward a sustainable model of data curation, survey data producers need to expand the use of persistent identifiers such as a Digital Object Identifier.[1]
  - Archiving of survey data and materials should be done using a common method, such as that prescribed by the Data Documentation Initiative.[2]

Improving Survey Website Usability

- Websites that provide survey data and documentation to users should be designed based on the results of usability research to assure that users can easily find the information they require. (Bailey et al. 2011)
  - Standardization of the format and content of such websites across survey projects would be desirable.

Research Transparency and the Credibility of Survey-Based Social Science

- To maintain the credibility of survey data and survey-based empirical claims, researchers should be as transparent about research practices as possible (Lupia 2008).
  - To enable data users to evaluate assumptions made during data collection, processing, and analyses, survey data producers should clearly and uniformly document and make public all aspects of data production, including
  - Sample selection
  - Respondent recruitment
  - Question selection
  - Question wording

[1] http://www.doi.org/

[2] http://www.ddialliance.org/

  - Response options
  - Directions to interviewers
  - Post-interview processing (including construction of weights)
  - Coding of open-ended responses

  (Lupia 2008)

- AAPOR-recommended standards for disclosure should be followed.[3]

References

Bailey, R. W., Barnum, C., Bosley, J., Chapparo, B., Dumas, J., Ivory, M. Y. et al. (2011). Research-Based Web Design and Usability Guidelines. (p. 292). *U.S. Dept. of Health and Human Services: U.S. General Services Administration.* Retrieved from Usability.Gov.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M. P., Couper, M. P., Dever, J. A., et al. (2013). Summary Report of the AAPOR Task Force on Non-Probability Sampling. *Journal of Survey Statistics and Methodology, 1*, 90–143. doi:10.1093/jssam/smt008

Belli, R. F. (1998). The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys. *Memory, 6*(4).

Belli, R. F., Smith, L. M., Andreski, P. M., & Agrawal, S. (2007). Methodological Comparisons between CATI Event History Calendar and Standardized Conventional Questionnaire Instruments. *Public Opinion Quarterly*, *71*(4), 603–622. doi:10.2307/25167583?ref=search-gateway:84a4cdb1103e0262c7f8b5723003d767

Boehm, L. M. (1989). Reliability of Proxy Response in the Current Population Survey. Presented at the Proceedings of the Survey Research Methods Section.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary Methods: Capturing Life as it is Lived. *Annual Review of Psychology*, *54*(1), 579–616. doi:10.1146/annurev.psych.54.101601.145030

Brick, J. M., Brick, P. D., Dipko, S., Presser, S., Tucker, C., & Yuan, Y. (2007). Cell Phone Survey Feasibility in The U.S.: Sampling and Calling Cell Numbers Versus Landline Numbers. *Public Opinion Quarterly*, *71*(1), 23–39. doi:10.1093/poq/nfl040

Brick, J. M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, *29*(3). doi:10.2478/jos-2013-0026

Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly*, *57*(1), 62. doi:10.1086/269355

[3] http://www.aapor.org/Disclosure_Standards1.htm#.U2kEOq1dXzQ

Cobb, C., & Krosnick, J. A. (2009). Experimental Test of the Accuracy of Proxy Reports Compared to Target Report with Third-Party Validity. Presented at the American Association for Public Opinion Research.

Conrad, F. G., Broome, J. S., Benki, J. R., Kreuter, F., Groves, R. M., Vannette, D. L., & McClain, C. (2012). Interviewer Speech and the Success of Survey Invitations. *Journal of the Royal Statistical Society: Series a (Statistics in Society), 176*(1), 191–210. doi:10.1111/j.1467-985X.2012.01064.x

Creighton, K. P., King, K. E., & Martin, E. A. (2007). *The Use of Monetary Incentives in Census Bureau Longitudinal Surveys* (No. Survey Methodology - #2007-2). *census.gov*. Washington, DC: U.S. Census Bureau.

De Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics, 21*(2), 233–255.

Dillman, D. A. (2005). Survey Mode as a Source of Instability in Responses across Surveys. *Field Methods*, *17*(1), 30–52. doi:10.1177/1525822X04269550

Durrant, G. B., Groves, R. M., Staetsky, L., & Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly*, *74*(1), 1–36. doi:10.1093/poq/nfp098

Durrant, G. B., D'Arrigo, J., & Steele, F. (2011). Using Paradata to Predict Best Times of Contact, Conditioning on Household and Interviewer Influences. *Journal of the Royal Statistical Society: Series a (Statistics in Society)*, *174*(4), 1029–1049. doi:10.1111/j.1467-985X.2011.00715.x

Grice, H. P. (1975). Logic and Conversation. In R. Stainton (Ed.), *Perspectives in the Philosophy of Language* (pp. 305–315). Ann Arbor: Broadview Press.

Groves, R. M., & Couper, M. P. (2012). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.

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

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, *72*(2), 167–189. doi:10.1093/poq/nfn011

Hainer, P., Hines, C., Martin, E. A., & Shapiro, G. (1988). Research on Improving Coverage in Household Surveys. *Proceedings of the Fourth Annual Research Conference* (pp. 513–539). Washington DC: Bureau of the Census. Retrieved from http://www.census.gov/srd/papers/pdf/rsm2006-07.pdf

Harvey, A. S. (2002). Guidelines for Time Use Data Collection and Analysis. In W. E. Pentland, A. S. Harvey, M. P. Lawton, & M. A. McColl (Eds.), *Time Use Research in the Social Sciences*. New York: Springer. http://doi.org/10.1007/0-306-47155-8_2

Heerwegh, D. (2003). Explaining Response Latencies and Changing Answers Using Client-Side Paradata from a Web Survey. *Social Science Computer Review*, *21*(3), 360–373. doi:10.1177/0894439303253985

Hox, J. J., & De Leeuw, E. D. (1994). A Comparison of Nonresponse in Mail, Telephone, and Face-to-Face Surveys. *Quality & Quantity, 28*(4). 329–344

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. *Public Opinion Quarterly*, *64*(2), 125–148. doi:10.1086/317759

Kosslyn, S. M. (2007). *Clear and to the Point: 8 Psychological Principles for Compelling PowerPoint Presentations*. New York: Oxford University Press.

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

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, *50*(1), 537–567.

Lelkes, Y., Krosnick, J. A., Marx, D. M., Judd, C. M., & Park, B. (2012). Complete Anonymity Compromises the Accuracy of Self-Reports. *Journal of Experimental Social Psychology*, *48*(6), 1291–1299. doi:10.1016/j.jesp.2012.07.002

Link, M. W., & Mokdad, A. H. (2005). Advance Letters as a Means of Improving Respondent Cooperation in Random Digit Dial Studies: A Multistate Experiment. *Public Opinion Quarterly*, *69*(4), 572–587. doi:10.2307/3521522?ref=search-gateway:6086db8a8168c3f8b01ad3b50083ece3

Lupia, A. (2008). Procedural Transparency and the Credibility of Election Surveys. *Electoral Studies*, *27*(732–739).

McDade, T. W., Williams, S., & Snodgrass, J. J. (2007). What a Drop Can Do: Dried Blood Spots as a Minimally Invasive Method for Integrating Biomarkers Into Population-Based Research. *Demography*, *44*(4), 899–925. doi:10.2307/30053125?ref=search-gateway:50cdd3c9e61d93d73634a76cf5c5fe3e

Olsen, R. J. (2005). The Problem of Respondent Attrition: Survey Methodology Is Key. *Monthly Labor Review, 128*(2), 63–70.

Parker, S. P., & Cubitt, W. D. (1999). The Use of the Dried Blood Spot Sample in Epidemiological Studies. *Journal of Clinical Pathology, 52*(9), 633–639.

Peytchev, A., & Olson, K. (2007). Using Interviewer Observations to Improve Nonresponse Adjustments: NES 2004. *Proc Surv Res Meth Sect Am Statist Ass*.

Presser, S., Traugott, M. W., & Traugott, S. (1990). Vote "Over" Reporting in Surveys: The Records or the Respondents? *Technical Report no. 39*. Ann Arbor: American National Election Studies. Retrieved from http://www.electionstudies.org/resources/papers/documents/nes010157.pdf

Ruggles, S., Sobek, M., King, M. L., Liebler, C., & Fitch, C. A. (2003). IPUMS Redesign. *Historical Methods: a Journal of Quantitative and Interdisciplinary History*, 36(1), 9–19. http://doi.org/10.1080/01615440309601210

Schaeffer, N. C., Garbarski, D., Freese, J., & Maynard, D. W. (2013). An Interactional Model of the Call for Survey Participation: Actions and Reactions in the Survey Recruitment Call. *Public Opinion Quarterly*, *77*(1), 323–351. doi:10.1093/poq/nft006

Schoeni, R. F., Stafford, F., McGonagle, K. A., & Andreski, P. (2012). Response Rates in National Panel Surveys. *The Annals of the American Academy of Political and Social Science*, *645*(1), 60–87. doi:10.1177/0002716212456363

Schonlau, M., Reuter, M., Schupp, J., Montag, C., Weber, B., Dohmen, T. et al. (2010). Collecting Genetic Samples in Population Wide (Panel) Surveys: Feasibility, Nonresponse and Selectivity. *Survey Research Methods*, *4*(2), 121–126.

Singer, E., Groves, R. M., & Corning, A. D. (1999). Differential Incentives: Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation. *Public Opinion Quarterly*, *63*(2), 251–260. doi:10.2307/2991257?ref=search-gateway:e350d899bd9b4af2cfe85f85b5a78443

Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. *Survey Nonresponse*.

Stone, A. A., Shiffman, S. S., & DeVries, M. W. (1999). Ecological momentary assessment. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 26–39). New York: Russell Sage Foundation.

Stone, A. A., & Shiffman, S. (1994). Ecological Momentary Assessment (EMA) in Behavorial Medicine. *Annals of Behavioral Medicine, 16*(3), 199–202.

Vicente, P., & Reis, E. (2010). Using Questionnaire Design to Fight Nonresponse Bias in Web Surveys. *Social Science Computer Review*, *28*(2), 251–267. doi:10.1177/0894439309340751

West, B. T. (2012). An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series a (Statistics in Society)*, *176*(1), 211–225. doi:10.1111/j.1467-985X.2012.01038.x

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

Willis, G. B. (2006). Review: Cognitive Interviewing as a Tool for Improving the Informed Consent Process. *Journal of Empirical Research on Human Research Ethics: An International Journal*, *1*(1), 9–24. doi:10.1525/jer.2006.1.1.9

David L. Vannette is a PhD Candidate in the Department of Communication at Stanford University, Stanford, CA, USA, and Principal Research Scientist at Qualtrics LLC, Provo, UT, USA.

Section 5

Detailed Chapters

In 2012, the National Science Foundation (NSF) commissioned a pair of conferences to focus on the "Future of Survey Research: Challenges and Opportunities." Many of the expert contributors to the previous sections of this volume also attended these conferences and presented on similar topics. The following section provides an in-depth view of the valuable material that was shared with the NSF in 2012. Each chapter in this section is based on material shared with the NSF by the author. The following chapters provide substantial added insights into the critically important topics that the NSF sought feedback on from this world-class panel of experts.

44

Reasons for Optimism About Survey Research

Jon A. Krosnick

Survey research is facing some really serious challenges. There are reasons to be concerned about the methods that are being used in many cases and the potential for improving them. One could take that as a downer, but before I get into the downer part I thought it would be nice to look at an upper part. LinChiat Chang and I have been attempting to answer the question, "How accurate are surveys?" As many people know, this is thought of as a difficult question to answer, because in order to assess accuracy of, for example, reports of cigarette smoking, you need to know the truth of cigarette smoking. But for some things, such as voter turnout from government records, we think we know something about the truth.

As you'll read later, we actually don't know as much about that as we think we know, but in order to have validation, you need some other measure, and if you could get some other measure against which to validate the survey, why would you bother doing the survey? So, there are rare occasions, you might think, when it's possible, but, actually, it turns out when you scour the literature it's not so rare, and there are four different methods that have been used to assess accuracy in this literature.

The first is to look at accuracy at the individual level because if a survey's aggregate statistic is to be accurate you might assume that it's got to be accurate at the individual level. So if somebody says, "I voted," then

J.A. Krosnick (✉)
Department of Communication, Stanford University, CA, USA
e-mail: krosnick@stanford.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_44

presumably we'd like to see records confirming that that person voted. So here we would look at the match between the respondent's claim that he or she voted with "objective," meaning quotes around objective, individual records of the same phenomenon.

The second method is matching one-time aggregate survey percentages, let's say the percent of people who voted, who say they voted, or some mean, the number of times people said they went to the doctor against some benchmark from non-survey data to look at how close that matches.

The third is to compute correlations where you can think of people as ordered from the people who go out to the movies most often to the people who go out to the movies the least often, and then you can somehow measure their frequency by, for example, looking at their credit card statements. Then you can correlate those two with each other as an indicator of accuracy. Last, we can look at trends over time, so as people's reports of their crime victimization go up and down, is it true that records of crime victimization also go up and down?

When we look at the literature on these methods, we found 555 comparisons of this first method, 399 comparisons of the second method, 168 comparisons involving the third method, and 6 instances of surveys tracking trends over time that allowed us to do this sort of thing. So, in total, we've got over 1,000 instances in which we can look at validation.

The types of phenomenon are all phenomena that you would think of as objective phenomena, so alcohol use, crime and deviance, and you can quickly look across, so, did you give money to a charity, did you evade taxes and get caught by the IRS, and things of that sort.

For the first method, there are 555 instances. Because there are lots of respondents, there are actually 520,000 opportunities for people to match or not match. We found an average of 85 percent of respondents exactly matching their record, so they said they got a colonoscopy, and the record shows they got a colonoscopy.

If you are worried about the distribution and worried about computing a mean, there are 88 percent with perfect agreement. I won't discuss all of those examples, but they're fun examples showing a very high agreement.

Method two has 399 examples. There are lots of different units of measurement, so often they are percentages, but sometimes they're centimeters of height or kilograms of weight or number of days when you did something or hours or number of teeth you have, so lots of different metrics. We found 8 percent perfect matches, 38 percent almost perfect matches within less than one unit of difference, and, obviously, you can imagine maybe one tooth is different from one day, but we found 73 percent very close matches on average. Those very close matches, five units of difference, seem quite sensible. Getting more precise than that may not be necessary.

If we look at method 3, these are the correlations of self-reports with secondary objective data across individuals, 168 such comparisons. One of the interesting things about the literature is that very different ways of computing correlations have been used in different publications, so Yule's coefficient versus the Uniclass correlation versus Cohen's kappa.

The results indicate that the average and median associations drop considerably from the top row to the bottom row, depending upon how those calculations are done. We've actually tried to figure out a way to work backwards from these publications to harmonize these and so far haven't been able to come up with a way to kind of standardize them and compare them directly, but for the most part we're seeing strong associations.

Lastly, are correlations of trends over time, so imagine again a graph of the rate of crime victimization from surveys versus government records. These are correlations of these trends over time,.96,.94,.91,.90,.77, and.73.

Now, these last two, you can see these are all very, very, very strong correlations. These last two are correlations of Survey of Consumers' reports of intentions to buy cars and houses with actual house and car purchases. So it's not exactly saying, "I bought a house." It's saying, "I intend to buy a house," and the trends are obviously very correspondent with each other, but they're not. That's a little bit different from the validity of the self-reports.

In this work, we are finding lots of correspondence, lots of accuracy, so things are not hopelessly broken at this point by any means, but the accuracy is much higher at the individual level than it is at the aggregate level. So this is one of the interesting questions, then. If people are honestly reporting, why do the aggregate statistics differ? I'll just tip my cards a little bit to suggest to you that I think it's about the process of aggregation. It's about, perhaps, biased non-response, as opposed to biased reporting.

We tend to think about the fact that surveys overestimate voter turnout as being the result of respondent lying. That is, respondents know turning out is socially desirable, and so people who didn't vote claim to have voted in order to look presentable, but, in fact, the accumulating literature is suggesting, no, actually, the individual reports may be remarkably accurate, and the problem is that people who participate in elections also over-participate in surveys. Now, that's no problem if you're doing the survey in order to study voters. You're getting them more efficiently than you might otherwise, but if you're trying to study voters and non-voters, you're losing those non-voters. But the key point here is that the inaccuracy at the aggregate level is not the result of inaccuracy at the individual level, necessarily. Respondents seem to be reporting objective phenomena remarkably well.

Jon A. Krosnick is the Frederic O. Glover Professor in Humanities and Social Sciences at Stanford University, Stanford, CA, USA, and a University Fellow at Resources for the Future. This work was supported by National Science Foundation Award [1256359].

45

Probability Versus Non-Probability Methods

Gary Langer

This chapter presents a high-level discussion of the importance and uses of probability sampling in comparison with alternative methodologies. Let's start with a disclaimer: I have no financial interest in any particular sampling methodology. My interest simply is to provide my clients with data that we can obtain as quickly and cost-effectively as their needs demand, but also only with data in which they, and I, can be highly confident. New methods are available. How much confidence we can place in them is the question we face.

Our focus, then, is on survey practices, empirical evaluation, and, ultimately, data quality. An ancient parable speaks to the fundamental importance of this discussion. It tells about about a man who goes to his rabbi seeking forgiveness for having spread false rumors about an acquaintance. The rabbi wordlessly leads his visitor outside to the yard. There he proceeds to rip up a down pillow, then stands silently for a moment until a wind kicks up out of nowhere and sends these tiny feathers swirling in all directions – across the yard, over the fence, through the trees and away.

"There go your falsehoods," the rabbi says to his visitor. "Go get them, and bring them back."

When we work with data, we are working with information that is uniquely powerful. It reflects a human imperative to come to grips with our

G. Langer (✉)
Langer Research Associates, New York, USA
e-mail: glanger@langerresearch.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_45

surroundings, to understand our communities, our societies, our nation and our world through the act of quantification. You don't have to read the Book of Numbers, or to know the meaning of the writing on the wall – "*mene, mene, tekel, uparsin*" – "numbered, numbered, weighed and divided" – to recognize the singular authority with which data speak. This reality requires us to be particularly prudent and careful in producing and delivering data to the world.

Good data are powerful and compelling. They lift us above anecdote, sustain precision and expand our knowledge, fundamentally enriching our understanding and informing our judgment. They're absolutely essential. Other data, by contrast, may be manufactured to promote a product or point of view. Intentional manipulation aside, they can be produced unreliably using suboptimal techniques or inadequate methodologies.

Some of these methods leave the house of inferential statistics. As such they lack a firm basis for the validity and reliability that we seek. The very low bar of entrance to these alternative methods brings in nonprofessional participants, untrained in sampling principles, questionnaire design, and data analysis. The result can be numbers and percentage signs that would seem to speak with authority, but that in fact can misinform or even be used to disinform our judgment.

We need to continue to subject established methods to rigorous and ongoing evaluation, but also to subject new methods to these same standards. To value theoreticism as much as empiricism – they are equally important – to consider fitness for purpose, and fundamentally to go with the facts, not the fashion.

A brief background on survey sampling may be of use. We start with the full population, a census survey. It can give you a very high level of accuracy, but it's prohibitively expensive and highly challenging to produce, as our research colleagues at the Census Bureau can confirm.

The first alternative is availability or "convenience" sampling, straw polls for example. They're quick and inexpensive. The risk is that they don't have a scientific basis or theoretical justification for generalizing to the full population. The pushback from elements of the research community is "*Hey, they seem to work – they're good enough*," or in today's much-abused phrase, they are "*fit for purpose*." That argument is not new; *The Literary Digest* did straw polls for decades, from 1916 to 1932, and correctly predicted the presidential election in each of them.

Then the 1936 election rolled around. The Literary Digest sent out 10 million cards to magazine subscribers, addresses from phone books and automobile registration lists. They got two-and-a-half million back and they found an easy win for the wrong candidate.

What happened? One, coverage bias: The sampling frame differed in important ways from the population of interest. Republicans were more likely to have been in the sample – skewing upscale, they were more apt to be able to afford subscriptions, telephones, and cars during the Great Depression. Next, we had systematic survey response, with Republicans more likely to participate, perhaps in order to express discontent with the incumbent Democratic president and his New Deal policies.

Given this failure, after 1936, availability sampling largely was replaced with quota sampling, which had correctly predicted FDR's win that year. Quota sampling attempts to build a miniature of the population. It identifies what are supposed to be key demographic variables; researchers then go out and seek to mirror those groups in their samples. That's said to produce representative estimates.

How did that work out? Not great, as Thomas Dewey could attest. There are a few reasons for the polling debacle of 1948. One was timing. Pollsters stopped collecting their data early, assuming things wouldn't change – when they did. Pollsters erroneously assumed that undecided voters would break as decided voters had; that's unsupported. Likely voters also may have been misidentified. But another problem involved sampling. Purposive selection appears to have allowed interviewers to choose more-educated and better-off respondents within their assigned quotas – pro-Dewey groups. This marks a key risk in purposive sampling – the risk of unintentional (or, indeed, intentional) systematic bias in the selection of individual survey respondents.

Even if quota sampling weren't entirely to blame, the Dewey-Truman fiasco highlighted inherent limitations in the method beyond those of respondent selection. One is that it's impossible to anticipate, let alone match, every potentially important demographic group. If a researcher wishes to match on gender, Hispanic ethnicity, race, education, age, region, and income within customary categories, that yields 9,600 cells raising the question of how many interviews can feasibly be done, and whether it's enough to fill those many cells. Further, as with availability sampling, quota sampling lacks theoretical justification for drawing inferences about the broader population.

The research community had this out in the early, formative days of the American Association for Public Opinion Research (AAPOR). Probability sampling – a random sample drawn with known probability of selection – won out. As Hanson and Hauser said in *Public Opinion Quarterly* in 1945, "an essential feature of reliable sampling is that each element of the population being sampled…has a chance of being included in the sample and, moreover, that that chance or probability is known."

The principle behind this thinking in fact goes back a little further, to the philosopher Marcus Tullius Cicero in 45 B.C. Kruskall and Mosteller quoted him in 1979 and I do so again here:

> Diagoras, surnamed the Atheist, once paid a visit to Samothrace, and a friend of his addressed him thus: "You believe that the gods have no interest in human welfare. Please observe these countless painted tablets; they show how many persons have withstood the rage of the tempest and safely reached the haven because they made vows to the gods." "Quite so,' Diagoras answered. 'But where are the tablets of those who suffered shipwreck and perished in the deep?"

Cicero's lesson on the perils of non-probability sampling have echoed throughout the ages, expressed in modern times, for example, by George Snedecor in the *Journal of Farm Economics* in 1939, Johnson and Jackson's *Modern Statistical Methods* in 1959, and Leslie Kish's *Survey Sampling* in 1965. There are many such cites. What they share in common is an expression of the fundamental theoretical principles of inferential statistics, which tell us clearly how and why probability sampling works.

Non-probability sampling has yet to enunciate any such operating principle. Yet that doesn't seem to deter its practitioners. Non-probability based Internet opt-in surveys are estimated to be a $5–$6 billion business in this country. A vast amount of market research has moved into Internet opt-in samples, and much else is done there as well. Let's explore what we know about this work.

My personal journey of discovery started 15 or so years ago as I set up the first standards and vetting operation for survey research by a national news organization, at ABC Network News. With the support of management, we put in place a system in which any survey research being considered for reporting would come to my group first; we would check it out, and either clear it or kill it.

The sort of things we saw include, for example, a report in my local paper of a poll, picked up from the *Sunday Times* of London. ABC was interested. We looked up the *Sunday Times* and indeed there it was, a poll of nearly 2,000 people by an outfit called YouGov. I was unfamiliar with them at the time this was 2003. We checked out their website: "Voice your opinion and get paid for it," it said. "Register now."

We looked further around the Internet and found a lot of photos of people waving their arms in the air, really excited about the money they are earning taking surveys on the Internet. In effect, it turns out, these are poll-taking clubs comprised of individuals who have signed up to click through online questionnaires in exchange for points redeemable for cash and gifts.

The missives also go out by e-mail. Here's one a friend of mine at Princeton received, and I suspect that every college student in America has seen it: "We have compiled and researched hundreds of research companies that are willing to pay you between $5 and $75 per hour simply to answer an online survey in the peacefulness of your own home." Or skip the cash: another we found raffled a car.

You might imagine the attraction of signing up for these things in a variety guises to increase your chances of getting selected to take surveys to win cash and gifts. Just as a test, a colleague of mine signed up for one of these online panels. He identified himself as a 32-year-old, Spanish-speaking, female, African-American physician, residing in Billings, Montana, and received surveys to fill out starting that very same week.

Controls are possible. If you've got to sign up with an address, and then you have to have your checks and gifts sent to that address, there could be some match there, an identity check. Rather, we find redemption pages with instructions that take a very different direction, suggesting that you can have your reward sent to a friend. Collecting prizes for a whole bunch of alternative personalities seems a lot easier.

Does this happen? Do people actually burn through many, many surveys if they're interested in racking up points redeemable for cash and gifts? One report found that among the 10 largest opt-in panels, 10 percent of participants accounted for 81 percent of survey responses, and indeed 1 percent of participants accounted for 24 percent of responses.

Questions are apparent. Who joins these poll-taking clubs? What verification and validation of respondent identities is undertaken? What logic and quality control checks are put in place? What weights are applied, on what theoretical basis, from what empirical source, and to what effect? What level of disclosure is provided, just for example in terms of the use of survey routers? What claims are made about the quality and qualities of the data, and how are they justified?

These are important questions if we are going to try to understand these data and try to come to some judgment about the approach. Such questions should be asked about all survey research – probability-based, and non-probability alike. We need to ask them so we can assess claims like this one, from Google Consumer Surveys, saying that its samples "produce results that are as accurate as probability-based panels." This claim is made in relation to a product in which you can buy the ability to ask one or two questions delivered as pop-ups to users of a search engine who are looking at something called "premium content." The demographic data that then is associated with the responses to these one or two questions apparently are imputed through analysis of users' IP addresses and previous page views.

Interestingly, you can go to Google and see who they think you are on the basis of your IP address and previous page views. A young woman on my staff was identified as a 55-year-old male with an interest in beauty and fitness. Another as a 65-year-old woman, double her actual age, with a previously unknown interest in motorcycles. And I was identified by the Google imputation as a senior citizen, thank you very much, in the Pacific Northwest, which I've visited exactly four times in my life, with an interest in space technology, which is news to me.

Another non-probability data provider lays claim to Bayesian "credibility intervals." This looks like no more or less than the typical formula to compute a margin of sampling error for a probability sample. Indeed, by whatever name, claims of a margin of sampling error associated with convenience samples are commonplace, without seeming justification.

What's the rationale for all of this? Per one provider's web page: "Traditional phone-based survey techniques suffer from deteriorating response rates and escalating costs." So, the pitch goes, we've got something else going instead. This is more a knockdown than a build-up argument: The old stuff costs too much and has got low response rates, so let's just throw in the towel on supportable inference.

But let's dissect this a little bit. We cannot achieve perfect probability; there are no 100 percent response-rate surveys. The Pew Research Center, as Scott Keeter reports in this volume, has observed a dramatic decline in its response rates, from 36 percent to 9 percent. (Some of us do better.)

Does this trend of declining response rates poison the well? Extensive research consistently has found that response rates in and of themselves are not a good indicator of data quality. There is theoretical support for this empirical result. If non-response to surveys itself is largely a random rather than a systematic phenomenon in terms of the variables of interest, then it does no substantive damage to inference. And there are a lot of carefully produced data to support that conclusion.

Response rates were one knockdown in the pitch I showed you. The other was that costs are escalating and it is cheaper to use an opt-in panel. No argument; some vendors have offered them at $4.00 per complete. That's a tenth or less of the cost of high-quality probability-sample research. The next step then is to ask, is it worth it? To know, we have to move to empirical testing of these data.

Yeager and his colleagues wrote an important paper in 2011 comparing seven opt-in online convenience sample surveys with two probability-sample surveys. The probability sample surveys were consistently highly accurate; the online surveys were always less accurate and less consistent in their level of accuracy. Results of the probability samples were less apt than the convenience samples to be significantly different from benchmarks, and the highest single error was dramatically different.

Opt-in online survey producers like to compare their results to election outcomes: the claim is that if you do a good pre-election estimate of a political contest, therefore you've got good data. I suggest that this is an entirely inappropriate basis for comparison. Pre-election polls are conducted among an unknown population; we don't know who's going to vote, so we resort to modeling and perhaps weighting for likely voters. That introduces judgment, often applied in an opaque fashion. In a good estimate, are we seeing reliable polling, or just good modeling?

Administrative benchmarks against which we can measure unmanipulated survey data offer a far more reasonable basis for comparison. In the Yeager et al. study, you can see the average absolute errors across opt-in panels compared to the probability samples, and the sizable inconsistencies across those non-probability sources. The average absolute errors were significantly different; the largest absolute error as well.

This paper also found little support for the claim that some non-probability panels are consistently more accurate than others – so you really don't know if you've landed on a good panel or not, or if you've got a good panel, whether it will be good next time. Weighting did not always improve the accuracy of the opt-in samples. There was no support for the idea that higher completion rates produce greater accuracy. And the probability samples were not just more accurate overall, but also more consistently accurate.

This study suggested that it's virtually impossible to anticipate whether an opt-in survey will be somewhat less accurate or substantially less accurate, or whether knowing that it is accurate on one benchmark can tell you whether or not it will be accurate on others. Yeager and his coauthors said this shouldn't be a surprise because there is no theory behind it. And they warned that you can cherry-pick results to create the appearance of reliability and validity in these data when in fact it may not be there on a systematic basis.

Other studies reach similar conclusions. The Advertising Research Foundation produced a study, "Foundations of Quality"; one element that was released found far more variation in estimates of smoking prevalence across opt-in online panels than in probability methods. As one commentator put it, “The results we get for any given study are highly dependent (and mostly unpredictable) on the panel we use. This is not good news.”

Mario Callegaro, at the 2012 national AAPOR conference, presented a review of 45 studies comparing the quality of data collected via the opt-in online method versus either benchmarks or other modes. Findings were similar to those we saw from Yeager, Krosnick. In summary:

- Panel estimates substantially deviated from benchmarks, and to a far greater degree than probability samples
- High variability in data from different opt-in panels
- High levels of multiple-panel membership (between 19 and 45 percent of respondents belong to 5+ panels)
- Substantial differences among low- and higher-membership respondents
- Weighting did not correct variations in the data

Opt-in Internet polls aren’t the only non-probability samples of concern. Mike Traugott, of the University of Michigan, gave a presentation at AAPOR in which he compared the accuracy of four low-cost data collection methods. One was Mechanical Turk, a panel comprised of people who sign up on the Internet to participate in studies in exchange for minimal compensation. Another was an IVR or robo-poll. The results indicated that the unweighted demographic outcomes were substantially different from benchmark data. Mechanical Turk, for example, is a good place to find Democrats, but it’s probably not a good place to find a good sample. And whether it was unweighted or weighted to standard demographic variables didn’t help.

Additionally, we have AAPOR’s report on online panels, produced in April 2010. It said that researchers should avoid non-probability online panels when one of the research objectives is to accurately estimate population values. This report talked about the underlying principles of theory. It talked about the non-ignorable differences in who joins these panels. It also said that the reporting of a margin of sampling error associated with an opt-in sample is misleading. In sum, AAPOR concluded: “There currently is no generally accepted theoretical basis from which to claim that survey results using samples from nonprobability online panels are projectable to the general population.” In 2013, AAPOR produced an additional report on non-probability sampling in general; while encouraging additional research and experimentation, it also noted the absence of a theoretical framework to support inference, among other challenges.

Apart from population values, AAPOR's report on opt-in Internet panels left room for this: Perhaps we can still use convenience samples to evaluate relationships among variables and trends over time. In 2010 Pasek and Krosnick found otherwise. They compared otherwise identical opt-in online and RDD surveys sponsored by the U.S. Census Bureau assessing intent to fill out the census, produced in order to inform a marketing campaign to encourage participation. Many of the findings of this research replicate what we saw previously with the Yeager et al. study, including that the telephone samples were more demographically representative than the opt-in Internet surveys, even after post-stratification. Pasek and Krosnick also reported significantly, substantively different attitudes and behaviors across the two data streams.

What was new was that they also reported instances in which the two data streams, as they put it, "told very different stories about change over time," as well as suggesting different predictors of intent to complete the census – the basic information you want and need if you're going to create a campaign to improve census compliance. Who is not going to complete the census? The authors indicate that the Census Bureau would have arrived at fundamentally different conclusions from these two different data sources.

So we have very different research conclusions from these two datasets. One is based on a probability sample, the other on a non-probability source. The notion that, regardless of estimated population values, you can use non-probability samples to draw conclusions about correlations in data is now in question. More research is needed. But the picture this research paints is not pretty.

Can non-probability samples be fixed? Bayesian analysis is all the buzz; the question is whether it's simply the magician's cloak. What variables are used, how are they derived, are we weighting to the dependent variable or to its strong expected correlates, and is that justified? Sample balancing is another suggestion – in effect time-traveling back to quota sampling. Can we take this non-probability sample get our 9,600 cells filled up and be good with it? Let's return to Gilbert et al. back in 1977: "The microcosm idea will rarely work in a complicated social problem because we always have additional variables that may have important consequences for the outcome."

Some say otherwise. A recent paper on political behavior, published in an academic journal, peer reviewed, based on non-probability sample YouGov/Polimetrix data, claimed to use sample-matching techniques to build representative samples with quality that "meets, and sometimes exceeds" that of probability sampling.

AAPOR'S task force on opt-in panels, as we have seen, says otherwise. And AAPOR has company. Among many others, Paul Biemer and Lars Lyberg have offered similar conclusions; to quote from their book

Introduction to Survey Quality: "Unfortunately, convenience samples are often used inappropriately as the basis for inference to some larger population." Further, they state, "…unlike random samples, purposive samples contain no information on how close the sample estimate is to the true value of the population parameter." These references are not intended to stall further research, but rather to encourage its sober practice. Close and continued evaluation of these methodologies is critical. So are full, forthright disclosure, and defensible claims about the data's validity and reliability.

We also need to evaluate research stemming from social media. This is a giant mass of data – truckloads of it, generated on a minute-by-minute basis. There is a lot of work underway trying to grapple with how to understand it. I appreciate, value, and encourage that work, but I also suggest that it should be held to reasonable research standards and scrutiny.

Consider the sampling challenges. We may want to assume that a tweet or a Facebook post represents one individual expressing his or her actual opinion on something once. In fact, users are not limited; some can post incessantly and others rarely. Users can have multiple accounts. Some accounts don't belong to individuals at all, but rather to companies, organizations or associations that use them, not to express individual attitudes, but to promote products or points of view. These posts can be driven by public relations campaigns including paid agents using automated bots. And this all exists within this giant data blob.

Parsing it out is highly challenging. Assuming users are people, their location often is not accurately captured, if at all. Users, of course, are self-selected, so again we're lacking a theoretical basis to assume that this is somehow representative of others. Recent research finds that among all online adults – and again there is still a substantial population that is not online at all – 15 percent are Twitter users at all, 8 percent daily. It's been estimated that fewer than 30 percent are Americans – a challenge if you're purporting to measure, say, U.S. election preferences – and of course there is a substantial age skew.

Let's skip over the sampling issues and say we're going to take this stuff anyway and figure out what it means. Determining the meaning of posts or tweets requires content analysis. Traditionally, we would do this through independent, human coders, with a codebook and a measure of inter-coder reliability. But that sort of approach is unrealistic given this volume of data. So researchers now generally use computerized analysis programs. They're lower cost and they're faster, but problems arise. Tweets are chock full of slang, irony, sarcasm, abbreviation, acronyms, emoticons, contextual meaning, hashtags – a highly complex dataset from which to derive meaning. There also is ambiguity in determining target words.

Say we are going to grab every tweet posted during the 2012 presidential debates that used the word "Obama." But what about tweets that used the word "the president" or "the pres" or "Barack" or "BO" or other phrases meant to refer to the president during the debate? We may miss those. And even if we figure out some way through this, we still don't have contextual data, demographic or other attitudinal data that will help us to reach intelligent research conclusions.

Kim et al. gave a presentation at AAPOR in 2012 in which they compared computerized content analysis systems to human coding in the ability to code Twitter comments as positive, negative, and neutral. The good news is that they weren't dreadful on neutral tweets in terms of matching automated coding to human coding. But the bad news is that on the positive and negative tweets, which we're likely most often interested in, they were dramatically different. This inconsistency is a cause for concern.

Rather than attempting to parse the content, some researchers are trying to focus just on the sheer quantity of posts about a certain topic. One study found that in the German national election, the proportion of tweets mentioning a political party in fact reflected its vote share. But it also found an enormous skew in the creation of these items – 4 percent of users accounted for 40 percent of the messages. This seems to be vulnerable to an orchestrated campaign to promote a political party. Similar analyses in the United States have differed; one, for example, found that Google Trends data did worse than chance at predicting election outcomes.

Twitter, again in 2012, posted something called the Twitter Political Index. Just what it was or how it worked was not disclosed. It was scaled on 0–100, but apparently not a percentage. You're Mitt Romney, you're Barack Obama, you get a number. And the numbers varied widely. On August 6th, Obama's number was 61. On August 7th, a day later, Obama's number was 35. Publicly released probability-sample surveys showed no volatility of this type. The meaning of this exercise, its purpose and its contribution to our understanding of the election were unclear, to say the least.

It's been suggested that Facebook user groups can be used for snowball sampling, to try to build a sample of hard-to-get respondents. A paper by Bhutta in 2011 reported on this approach to reach Catholics. The author reported that it was faster and cheaper than traditional methods, but there were vast differences in the Facebook snowball sample that she obtained versus General Social Survey data in terms of demographic characteristics of Catholics, including gender, race, education, and mass attendance. Bhutta concluded that the Facebook respondents could not possibly serve as a representative sample of the general Catholic population.

We need to touch, as well, on falsification. It was suggested in 2012 that as many as 30 percent of Obama's Twitter followers and 22 percent of Romney's were fabricated. Forbes magazine reported how individuals can purchase 25,000 fake Twitter followers for $247, and that these bots automatically will tweet snippets of text. Are we going to use this information to assess public attitudes?

What does the future hold? In probability sampling, we need continued studies of response-rate effects; whether it's 19 or 9 percent, it's not what it used to be. Efforts to maintain response rates can't hurt. Renewed focus on other data-quality issues is important including quality control, coverage, and questionnaire design, too often the forgotten stepchild of survey research. Equally important is the development of probability-based alternatives, including online and mobile-based administration and improved panel-management techniques.

In non-probability sampling, further study of the appropriate as well as inappropriate uses of convenience sample data are very much in order. Further evaluation of well-disclosed and emerging techniques will be important; this means enhanced disclosure and a more sober approach than we often see today.

In social media, a vast and growing source of data, we also need further evaluation of the appropriate use of this material. We can study, for example, whether and how social media relates to attitude formation and change in public opinion, perhaps as a compliment to reliable attitudinal measurement. The key in all cases is to establish the research question, evaluate the limitations of the tools available to answer that question, and then go forward with informed judgment, full disclosure, and honest assessment on the basis of empirical results and theoretical principles alike.

Gary Langer is a survey research practitioner. He is president of Langer Research Associates and former long-time director of polling at ABC Network News. Langer is a member of the Board of Directors of the Roper Center for Public Opinion Research, a trustee of the National Council of Public Polls and former president of the New York Chapter of the American Association for Public Opinion Research. His work has been recognized with 2 News Emmy awards, 10 Emmy nominations, and AAPOR's Policy Impact Award. Langer has written and lectured widely on the measurement and meaning of public opinion.

46

Address-Based and List-Based Sampling

Colm O'Muircheartaigh

Address-Based Sampling (ABS) has emerged as the dominant form of sample design for social surveys in the United States in the past 15 years; a commercial clone of the U.S. Postal Service (USPS) Delivery Sequence File (DSF) provides the basis for these samples. Previously, samples for face-to-face surveys were based on an area sample (using maps and census counts of households) and subsequent field listing and mapping of housing units. The availability of a geo-codable list of household addresses covering the great majority of U.S. households has substantially reduced the cost of sample design and implementation. This development has facilitated a contemporaneous shift toward multimode surveys.

What we try to in sample design is first to identify the whole set of elements, people, entities, or units for which we wish to make inferences through our survey – the survey population. We then select a subset of these units, the sample, for observation or measurement. To delineate the population we need a frame, which is the set of materials that comprises the whole of the population and permits us to select a sample. If an element is not included in the frame, then it cannot appear in the sample. And if the frame is not to some extent a mirror of the population, the sample will not represent the population appropriately.

C. O'Muircheartaigh (✉)
University of Chicago, Chicago, Illinois 60637, USA
e-mail: caomuirc@uchicago.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_46

Although we don't normally directly consider it a part of sampling, the choice of a mode for data collection is critical because the two are so intimately intertwined that it's really not possible to talk about sampling frames and sample design without thinking about mode. And in the current research environment, with the challenge of dramatically low response rates in some modes, we are considering more frequently using combinations of modes in a single survey.

Alternative Frames and Modes

Ideally, the sampling frame for a population should be an identical match to the population—but it never is. The degree to which the frame excludes elements of the target population is important, but so is the extent to which the frame includes elements that do not belong to the target population.

For high quality surveys of the general population, conducted by face-to-face interviewing, maps combined with population data from the U.S. Census formed the base material, and within areas selected using those materials direct listing of housing units provided access to households and individuals. The universe was defined by the maps and associated materials used in these surveys.

Centralized frames of telephone numbers became available in the 1970s, and telephone surveys dominated the field through the 1990s. For those surveys, the universe of all telephone numbers came to define the population frame; Random Digit Dialing (RDD) as a method exploited the capacity to construct a frame that included all telephone numbers, and provided a reasonably efficient way of selecting samples from that frame. A key problem with the all-possible-telephone-number universe is that it contains many, many numbers that are not actually telephone numbers, in the sense that there is no working telephone or household associated with them. Identifying which part of the frame comprises real numbers is a non-trivial task, and an increasingly difficult one, as the number of possible telephone numbers is increasing relative to the number of actual (live) phone numbers.

Telephone surveys provide another, rapidly changing, challenge. The growth in the number, and proportion, of cell-phone-only households in the U.S. has made the construction and interpretation of a sampling frame for telephone surveys a daunting prospect. The percentage of households with cell-phone-only telephone service rose above 50 percent in 2016 for the first time; there is an additional 15 percent or so of households where cellphones provide the dominant form of telephone contact. Landline phones, like addresses, are typically associated with physical locations and particular households. Furthermore, in the case of landlines, the telephone number was generally considered by all household members to be a telephone number for the whole household. Cell phones occupy a completely different conceptual space. To most people, their cell phone is theirs individually, and they do not consider others to be contactable by that number, nor do they consider themselves to be contactable by the cellphone numbers of others in the household. The sampling unit—and the probability of reaching a particular individual—is now ambiguous; this complex relationship between telephone number and individual makes telephone sampling, when we are dealing with both landlines and cellphones, considerably more difficult than it was before.

An alternative approach to obtaining good coverage of the population would be to use a list of residential addresses, but no such list was accessible for the U.S. Many developed countries, especially in northern Europe, use such lists routinely. Some are derived from compulsory registration systems and many were originally based on lists of electors; as these lists became less comprehensive due to increased immigration and lower levels of registration by eligible voters, attention shifted to the use of postal codes and postal delivery addresses.

In the U.S. direct marketing organizations (such as mail order companies) have for many years been using access to the U.S. Postal Service (USPS) address lists as a basis for mass mailings. It transpired that it was possible to get access to this list through commercial vendors. The most valuable feature of this list is that, though not perfect, it has very positive coverage characteristics: it contains relatively few nonexistent addresses and excludes relatively few actual addresses. The mere existence of such a list in the U.S. was not particularly useful to samplers. It was difficult to get access to the list for sampling purposes and it was not stored in a form that made sampling a feasible practical process.

Two developments transformed the situation in practice. First, the list became available in computerized form, permitting easy sorting and selection. Second, there was a revolution in mapping software that permitted the inexpensive geo-coding of large numbers of addresses (more than 100 million) and their location on maps at high resolution, so that the addresses on a particular street or block could be identified.

Address-based sampling (ABS), as the name implies, is an approach to sampling where basic element in the sampling frame is an address. For social, economic, or demographic surveys the address is a residential address. Each address signifies one (ideally) or more households; each household will contain members; and the survey population will comprise some set of elements within that hierarchical set. Whichever the target elements may be, the address frame will give us coverage of those elements.

The Computerized Delivery Sequence File (DSF)

The basic frame we use is the Delivery Sequence File (DSF) of the USPS. Mailing addresses in the frame are ordered following the organizational (regional and spatial) structure of the Postal Service, and the way it manages the delivery of mail down to the level of the individual mail carrier. Within the list for each mail carrier, the addresses are arranged in the order in which the mail carrier delivers the mail. We believe that almost all the addresses in the country are on this frame—perhaps 98 percent or 99 percent; it is possible to have your address removed from the frame, but it is a complicated administrative process, and very rarely activated. The DSF has two great strengths for samplers: (i) the addresses are in a standard format; and (ii) the quality of the frame is very high. These characteristics were what motivated me and my colleagues at NORC, and samplers at other major survey organizations, to examine the properties of the DSF as a sampling frame, beginning in 2002.

The DSF is updated by mail carriers using edit books while carrying out normal mail delivery. The mail carrier (i) identifies in the edit book addresses that no longer exist or are vacant, and deletes them from the list; and (ii) identifies and adds to the list new construction and other new addresses. The amendments in the edit book are delivered to the administrative staff, who (eventually) enter the new data in a central data bank. The overall frame is updated every month. The update incorporates additions and deletions, though it is not clear how quickly amendments get from the carrier to the central data base.

The reason I had confidence in the quality of the frame from the beginning is that it is in the mail carrier's interest to have an accurate frame; this issue of motivation is one that many other authors discuss throughout this volume. For the mail carrier it is important that all addresses that receive mail should be on the file, because the carrier's route will be constructed and evaluated on the basis of the number of active addresses on the frame. It is equally important that nonexistent addresses should not be on the file as otherwise the mail carrier will have to deal with a large volume of undeliverable mail. Direct marketers and campaigns of all kinds use the DSF as the basis for mass mailings, and mail for these nonexistent addresses will accumulate in the mail carrier's load but not be deliverable.

Hence there is a built-in incentive for the mail carrier to keep this frame as up to date as possible; this contrasts with the lack of incentive for listers of addresses for sampling frames for traditional area sampling. The incentive structure for the DSF seems ideal for the construction and maintenance of a frame that includes all eligible elements while excluding blanks and other ineligibles. The objectives of the postal service match well the purposes of the sampler.

A further *a priori* endorsement of the DSF came from the U.S. Bureau of the Census, which has used the DSF as the basis for its own Master Address File for many years. And since the early 2000s Vincent Iannachione and his colleagues at Research Triangle Institute (RTI), Colm O'Muircheartaigh and a team at the National Opinion Research Center (NORC) at the University of Chicago, and Michael Brick and his colleagues at Westat have been examining the quality of the frame, and have concluded that it is of high quality and improving operational practicality. Almost all addresses are in the DSF, but not all are in a form that permits us to find them in the field. We calculate now that somewhere above 90 percent of the household population of the U.S. lives in areas where the DSF is a frame with very high (and acceptable) coverage. The frame is constantly improving, as the standardization of addresses is being extended to almost all communities as part of the 9-1-1 (emergency service) conversion of nonstandard addresses. At NORC we used the DSF on a large scale for the first time in 2002 for our national sample design. We used the 2000 Census results as our area frame, and at that time there were areas accounting for 27 percent of the population where we concluded that the DSF was not adequate as a sampling frame of addresses. In our national sample redesign a decade later, based on the 2010 Census results, only 10 percent of the population lived in areas where the DSF was an inadequate sampling frame for addresses. This is an impressive improvement in effective coverage.

The DSF was not designed as a sampling frame, however, its two key features are (i) its coverage and (ii) its standard format. The standard format defined and used by the Postal Service has facilitated enhancement of the frame by producers of maps and other data structures. In particular, mapping software will identify the location of addresses extremely well. And once an address is geocoded it can then be linked to any other database that contains geocode information. Neighborhood and area characteristics can be linked to the address and used for stratification in the sample design. We can connect demographic characteristics, not just of the individual, but of the household, the census block, of the tract, and any other spatially identified information. We can use this not only for sample design, but also to augment the analysis of the survey data. In the future, this may even replace the need to collect directly a great deal of information that we now collect (expensively) from the respondent.

This is particularly relevant to the other chapters in this volume that deal with connecting databases to survey data. All of those external databases that are either spatially identified or linked to an address are available to augment both the sampling frame and the analysis. This contrasts with the situation in sampling for, and analyzing, telephone surveys where there is a persistent problem of not being able to identify a location for either cellphone numbers or for many landline numbers.

There are some drawbacks to the DSF that are worth noting, because sometimes, when faced with the overall quality measures—98 percent coverage, for example—researchers may think that the DSF is always going to work for them. There are situations where this is not the case. First of all, geocoding does not work perfectly. Even good mapping software will sometimes put things in the wrong place, and this may be important in surveys that are targeting particular locations or spatially defined populations. The more narrowly spatially defined your target population is, the more important it is that the geocoding should be accurate. The DSF itself is not geocoded, so it is the quality of the mapping software that determines its quality as a sampling frame.

Post Office (P.O.) Boxes identify the location of the Post Office facility in which the P.O. Box is located—not the location of the recipient. Though most P.O. Boxes are not residential, there are many that are. Those carry an identifying flag—OWGM (only way to get mail)—that indicates that this is where and how a particular residence receives its mail. The location of the residence cannot be determined from this information.

The DSF can also prove misleading if you are interested primarily in new construction, or if new construction is an important element in your population. Though the USPS endeavors to add new construction in a timely way, the length of the time lag between construction and appearance on the DSF is variable. Conversely, depending on the state of the housing market, the frame may also contain housing units that are not residences because, while they were added in anticipation of their being occupied, this has not yet occurred. This was a particularly acute problem at the time of the housing market collapse post-2008.

Drop points are also important to mention; they are of particular interest to those of us working on city surveys or doing research that includes particular cities. A drop point denotes an address on the frame (a central location) at which the mail carrier delivers the mail. The mail carrier does not deliver the mail to individual housing units; the mail is distributed to individual residences by a third party, or collected from the drop point by the residents. Drop points are typically apartment buildings (some large with concierge service, some small without individual mailboxes) or gated communities.

On the frame as a whole, only 2 percent of addresses are drop points; however, those drop points are not distributed uniformly throughout the frame but are concentrated in particular locations. They are clustered particularly in major cities: about 15 percent of Chicago's addresses are drop points; in New York perhaps 20 percent; and in Boston, 10 percent. Of greater concern for local surveys, drop points are also disproportionately concentrated in particular neighborhoods; some neighborhoods in Chicago have a drop point rate of 60 percent. Such non-household-specific addresses present an additional challenge when we wish to match the address to an external database. Because the address does not identify a household or a person, we cannot match household- or name-specific data from other sources to the address.

Multimode Surveys

The second great change in survey design in the last 10 years is that we are now much more comfortable with the idea of using multiple data collection modes in the same survey. There has been a great deal of work over the years on differences in responses to different modes, demonstrating the difficulty of creating equivalent stimuli across modes. However, the daunting challenges facing single-mode surveys—plummeting response rates for telephone surveys, soaring costs for face-to-face surveys—have led to the realization that combining modes in a single survey may provide at least a partial solution to these challenges. This trend was greatly assisted by the reemergence of mail surveys as a viable alternative to telephone surveys. Mail surveys had been in widespread use for both government and commercial surveys through the middle of the twentieth century, but were largely displaced in social research with the growth of telephone surveys in the 1970s and 1980s. However, a program of work by Don Dillman and his colleagues at Washington State University and innovative tests at the Centers for Disease Control and Prevention (CDC) on the Behavioral Risk Factor Surveillance System (BRFSS) surveys demonstrated that mail surveys could obtain response rates comparable to, or surpassing, those for telephone surveys. The American Community Survey had been using, since its inception, a sequential strategy for data collection, beginning with mail, moving to telephone where possible for mail nonrespondents, then subsampling remaining nonrespondents for follow-up with face-to-face interviewing. The US Census Bureau of course had access to its Master Address File (MAF), which is not available to outside researchers.

Access to the DSF offers a comparable platform to other major survey organizations. The DSF is the ideal sampling frame for mail surveys, being constructed and maintained explicitly as a mailing frame. The work on validating the use of the DSF for face-face surveys was helpful in establishing its coverage and quality as a frame for surveys. And the fact that both modes were using the same sampling frame encouraged researchers to consider combining modes in a single survey.

Sampling for multimode surveys would be relatively straightforward if all modes were had a common sampling frame; when this is not the case, a crosswalk between the sampling frame for the initial mode and the subsequent modes will be necessary. (We will assume that the survey will take a sequential approach to data collection, beginning with one mode and progressing to the others when the initial mode is not successful.) The DSF and ABS work well for mail and face-to-face survey, but present a problem when it comes to telephone sampling.

The addresses of nonrespondents can be matched to a telephone number database; a number of vendors specialize in this work. Match rates will typically be around 50-60 percent; the best match rates are found for single-family homes and for homes with older inhabitants. Even though these match rates are not wonderful, telephone data collection may augment the response rate significantly. The remaining nonrespondents are available for face-to-face follow-up, if resources allow.

There are of course drawbacks to using multiple modes in a single survey. There are researchers who have demonstrated the potential for substantial mode effects, where the data collected by different modes are not equivalent. While mode effects are serious and important, if the choice is between having respondents answer in different modes and not having them answer at all, I would choose the former. Indeed, if we are willing to accept data from a single-mode survey in each mode, it is hard to argue that the data from any mode should be excluded from our data set.

We need to start thinking carefully about the difference between respondent capture [getting the respondent to agree to take part in the survey) and data capture (getting the respondent to provide information about the survey topics). The strategy and approach required to get the respondent to cooperate may vary, but this does not necessarily imply that you will collect all the survey data from the respondent in that mode, whatever it may be. You could envisage a situation where, once the respondent had agreed to be interviewed, the data collection itself might be through self-completion, either on paper or on the web. Indeed many of the topics on which we now strive to collect data may soon require only the respondent's permission to access these data directly. For example, should electronic health records become the norm and respondents have access to them on the web or from an electronic health card, all that would be required would be to ask the respondent for permission to access the data. Should this become an important component of data collection, most of the concerns about inter-mode effects in those situations would disappear.

If you plan to use more than one mode in data collection, you must have compatible sample designs. For mail surveys and for telephone interviewing, the spatial distribution of the sample elements makes almost no difference to your costs. But for face-to-face interviewing, costs increase almost proportionally with distance between sampled elements, and the spatial concentration of the sample is crucial. This implies that the sample design will require significant clustering; indeed the form of the sample design will be dictated by the needs of face-to-face data collection.

So if you want to do face-to-face interviewing as part of a multimode survey, at least part of your sample design, and I emphasize this, because it doesn't mean this has to be true of your whole design, but at least a representative part of the design must contain sufficient clustering so that you can consider the possibility of following up a subsample of nonrespondents through face-to-face interviewing.

Conclusion

The future will lie in augmenting the DSF. First, learning to cope with, and then repair, the deficiencies in the frame. Second, in parallel, linking the DSF —the address frame—to the wide, and increasingly rich, set of databases that will make identification of potential respondents easier, and eventually providing a basis for targeted methods of approach and recruitment. Such an augmented frame will pave the way toward data collection that will consist of (i) obtaining permission to access data in many contexts and (ii) devoting most of the interaction time with respondents to ascertaining information that is not otherwise or elsewhere available.

In the meantime, our efforts should be concentrated in five areas: (i) reconciling and linking the address and telephone databases; (ii) improving our survey management systems so that cases (respondents) can be passed seamlessly from one mode to another, and back if necessary; (iii) incorporating other methods of data capture, such as the internet (through smart phone, tablet, or computer) into the survey management system; (iv) developing multimode-compatible instruments, so that we match the stimulus to the mode, and do not try to use the same form in different modes; and (v) developing the capacity to track in real time what is happening in the field; the rise of adaptive or responsive designs will indicate how we should navigate the process of data collection, but the software will need to be developed to enable us to implement these insights effectively.

We are making progress on all these fronts, and I am confident that the future will be no worse than the present.

Colm A. O'Muircheartaigh, professor at University of Chicago Harris School, served as dean of the Harris School from 2009 to 2014. His research encompasses survey sample design, measurement errors in surveys, cognitive aspects of question wording, and latent variable models for non-response. He is a Senior Fellow in the National Opinion Research Center (NORC), where he is responsible for the development of methodological innovations in sample design.

O'Muircheartaigh is co-principal investigator on NSF's Center for Advancing Research and Communication in Science, Technology, Engineering, and Mathematics (ARC-STEM) and on the National Institute on Aging's National Social Life Health and Aging Project (NSHAP). He is a member of the Committee on National Statistics of the National Academies (CNSTAT) and of the Federal Economic Statistics Advisory Committee (FESAC), and serves on the board of Chapin Hall Center for Children.

47

The Impact of Survey Non-response on Survey Accuracy

Scott Keeter

This volume presents a wonderful opportunity for those of us who live and die by surveys to try to understand the issues facing us and how can we do better. Since I am no longer an academic and I'm a practicing pollster – I'm going to talk to you largely from the perspective of someone who is doing political polling, particularly in the kind of environment that we live in these days.

It has not escaped anyone's attention that over the past couple of elections there has been a lot of criticism of the polls in the presidential races. The charge is that we are oversampling Democrats. We're getting too many liberals in our polls, and thus the leads that most of the polls have shown for Democrats in the elections constitutes either malpractice on the part of pollsters, non-response bias, or downright conspiracy to pump up the Democratic campaigns and depress Republicans.

So, this is the environment in which many of us are doing our work, and the question of non-response bias is something that's in front of us constantly and that we worry about. So, while I will spend a little time on the literature in this field and particularly some studies that I think have made wonderful contributions to our understanding of non-response bias, I'm also going to focus a good bit on a study that we at Pew Research Center did to try to get a better handle on the issue as it affects

S. Keeter (✉)
Pew Research Center, Washington, D.C., USA
e-mail: skeeter@pewresearch.org

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_47

people who do the kind of social and political surveys that we do, surveys that tend to have very low response rates, much lower than they used to be and much lower than the government surveys that have been the basis for a lot of our understanding. I was reminded in researching this chapter that the National Science Foundation had a conference about 40 years ago, which was described as addressing the question of whether non-response has reached a level or is growing at a rate that poses a threat to the continued use of surveys as a basic tool of survey research. So this is not a new question, of course, and it's one that concerns us, has concerned the field for decades.

Of course, any trained survey researcher who is on the upper half of the 50/49 age break knows that when we went through our training in survey research the idea that you could make reasonable inferences from surveys with response rates below 60 percent or 70 percent or certainly 50 percent was just kind of laughable. Today, many of the important data collections that are funded by the federal government have fallen below that level, and, of course, many of the surveys of the sort that I'm involved in are now down in the single digits with response rates.

However, there is some good news, especially in the political world, that our surveys have continued to perform well in terms of predicting how voters are actually going to vote. This is, of course, a perennial anxiety for us in the field. In any election, we have to make a final pre-election prediction for the presidential election result, and it will be a very visible success or failure if we're right or we're wrong. But, fortunately, polls have been right a lot more often than they've been wrong in forecasting election outcomes. If you think about the number of different elections that are being polled, there is a very sizeable body of evidence about this issue. You have state-level races, the Electoral College vote each state level, and the polling aggregators and the individuals that I think many of you, if you're interested in politics, are familiar with, Nate Silver, Mark Blumenthal, and others and the fact that they do keep score cards. Polls have done extremely well in forecasting the outcome and not just in terms of saying who's going to win or who's going to lose but getting within a point or two when you take the collective average of the polls of the margins.

So, there is a good record. There is evidence that non-response, as serious as it is, is not creating a bias, at least for that particular survey statistic. But that's part of the key and part of the big finding from the literature, and that is not that there is no non-response bias – there is non-response bias all over – but that it is very survey- and very statistic-specific. It tends to occur on some kinds of measures and not on others.

Let's just review the evidence about the rise of non-response just for a moment. Again, this is not anything new to anybody. The first piece of evidence I want to highlight is the non-response in the National Household Education Survey, which is a telephone survey. The level of non-response has been growing from around 40 percent back in 1996 to near 70 percent. On the basis of this trend and the problems that it was creating in terms of the data collection, the decision was made to experiment with an alternative method for data collection, a different mode using address-based sampling. Similar results can be seen in the National Immunization Survey. The rate of decrease is not as great there, but the trend is unmistakable. One reason, though, that this rate of decrease is not so great is the fact that this particular survey largely consists of a very, very short screener instrument that poses very little burden. This is also a survey on a health topic, and health topics tend to have lower levels of non-response, and so this particular data collection, because of these two features, at least, has not been affected as seriously as some others.

When you shift over to the world of face-to-face interviewing, there is evidence that more effort is involved and that there is potentially greater non-response in many of these data collections, but when you look at the National Science Foundation-funded General Social Survey you actually see a kind of plateau. There was a drop in response rates a decade or more ago, and then for the last several data collections and the it's been flat at about 70 percent for roughly 10 years. When you look, though, at the kind of random-digit dialing telephone surveys that we do at the Pew Research Center, that have to be conducted over a very short field period, the data tells a pretty serious tale of woe for us.

I examined some representative surveys from our data collection going back to 1997 and looked at the contact rate, the cooperation rate, and then the overall response rate using the American Association for Public Opinion Research's (AAPOR) Response Rate 3 calculation. Two things were immediately evident. First, we went from a 90 percent contact rate in 1997 for a 5-day survey down to 62 percent in 2012. The cooperation rate has fallen from 43 percent down to 14 percent, and that translates into response rates going from 36 percent down to 9 percent. The 9 percent number when we released this in 2012 made quite a splash. A number of people who are in the survey business, including some of our competitors, said, "Oh, our response rates are not that low." They said, "Our response rates are in the mid to high teens," and I have to laugh that, "Okay, you've got a bigger response rate than we do," but still by the standards that we used to hold this is a pretty scary number.

We could go on through lots of different studies demonstrating this pattern – the Survey of Consumer Attitudes, run by the University of Michigan, has had the same kind of trend line. So, what do we know about the consequences of declining response rates of growing non-response bias? What is the potential impact here? This literature is happily very vast, and it is compelling reading because there's lots of clever work that has been done using a variety of different techniques.

Probably the best way to summarize it comes from a *Public Opinion Quarterly* article by Bob Groves and Emilia Peytcheva (2008). Their meta-analysis examined the level of bias for individual measures across a large collection of surveys, distributed across different levels of survey non-response. A couple of things are immediately evident from their results. One is that there is no obvious trend in this in terms of seeing a bigger cluster at one end of the graph than the other, which suggests that taken as a whole the level of non-response is not associated with the level of bias. The second observation is that at any given level of non-response the level of bias varies considerably. The difficulty is in trying to figure out what kind of measures are apt to have the greatest likelihood of being biased by non-response. But the problem is even more complicated than that because the type of bias that can occur can be a result of any number of different factors, including topic salience, how the survey is introduced, what your population is that you're drawing from relative to the nature of the questions, the measures that are being employed. So it's very difficult to come up with any clear generalizations about when you can expect non-response bias to occur and how severe it's likely to be.

Now, despite this difficulty of being able to predict, we do have some expectations, and the Pew Research Center has been interested in this question for a long time. We began trying to get at the question of non-response bias back in a study that we did in 1997 that we published in 2000 in *Public Opinion Quarterly*, which looked at the impact of greater effort in a survey, essentially trying to do a survey by gold standard methods, as opposed to the 5-day data collection that is our standard practice. In that study, we came up with the conclusion, which was a bit surprising at the time, that the level of effort did not seem to make any difference in most of the survey estimates that we came up with. We did have a couple of estimates that seemed different in the high effort/higher response rate study and the low effort/lower response rate study.

One set of measures had to do with racial bias or animosity, that is, respondents holding attitudes that were negative toward racial minorities. The second one was the level of reported volunteer activity, particularly reports of high levels of volunteer activity. Those also appeared to have a bias. The bias wasn't large, it was only about five points, comparing the lower and the higher response rate methods, but it was clearly there. Upon further examination, it turned out that the bias that appeared to be present in terms of racial attitudes, that is, that the survey that got the higher response rate turning up more racial animosity, was compromised somewhat by some design features in the study. The interviewers who conducted most of the interviews in the extended field period portion who did the refusal conversions and so forth were overwhelmingly male and white.

The analysis of the data that was done largely by Stanley Presser and Bob Groves as partners on that study revealed that, in fact, what had happened was that there was a race-of-interviewer effect that accounted for much but not all of the supposed non-response bias in racial attitudes in that study, but on the whole the study was reassuring and somewhat surprising. Other studies at the same time, including some that Stanley Presser was involved in, showed that varying levels of effort in the Consumer Confidence Survey also did not produce differences in the estimates. So, there was some reassurance in that that at least within the ranges of survey response rates that we were talking about that we were not seeing significant differences in the accuracy of survey measures. That is not at all the same thing as saying that there was no non-response bias, but at least within the boundaries defined in the study it was a good finding.

We repeated the study in 2003 with basically the same results, and then we had a lengthy hiatus as we turned our attention to the issue of non-coverage in cell phones, but we returned to the non-response bias issue in 2012 to try to look at it again. Given what had happened in the literature over this period of time, it was clear to us that a comparison of a low-effort and a high-effort survey was not really going to be definitive in answering the kinds of questions that we had, in part because I think the model underlying the notion of comparing the higher and the lower effort surveys was wrong.

The idea of a "continuum of resistance" model just didn't seem to fit the data in our study or many of the other studies that have come in the meantime. We also knew that even the high-effort survey was not going to get a very respectable response rate. So, instead, we decided we would do a multifaceted investigation in which we did not only put greater effort into a rigorous survey, but we made a very careful effort to benchmark many of the findings in the study to high response rate U.S. government surveys. Most researchers do this when looking at demographic characteristics, but we went beyond it to a number of other measures, which I will cover in greater detail next.

Then, finally, taking advantage of the growth in the availability of voter and consumer databases that we have now that we really didn't have at the time that we did our previous studies, we purchased a couple of those databases, and we made an effort to match our sample, at least from the landline sample, to the records in these databases so that we would have a basis for judging whether respondents were different from nonrespondents.

So, first of all, what do we get out of the extra effort that was involved in doing the high-effort survey? The answer is in some cases you get a fair amount in terms of additional contacting, but you don't get all that much improvement in cooperation, and response rates are still very low. In terms of the contact rate, we found that if you give yourself enough time, you can actually make contact in almost as many households now as we could in our previous two surveys. We ended up being able to contact 86 percent in the landline frame and 84 percent in the cell frame with multiple calls stretched over a long period of time. We were able to get the cooperation rate up but not to particularly high levels and with the consequence that the response rates, while higher – we ended up with a 22 percent response rate in the high-effort survey this time compared with the 9 percent in the standard survey – that's still just one out of five households successfully interviewed. Additionally, there were substantial differences between response rates in the landline and the cell phone frames. The cell phones pose considerable difficulties in terms of getting cooperation, both in terms of getting people on the phone and getting information from them.

So, what did we find in the substantive analysis? First of all, looking at the comparisons between the standard survey's estimates and the government benchmarks, the results are quite mixed. For many of the estimates, our numbers are within the margin of sampling error of what the government statistics are. One place where we fell short is on Social Security payments. We got 32 percent in the standard survey, just 27 percent taken from the government survey. In terms of food stamps, we got more people reporting getting food stamps or nutrition assistance than in the government survey. We did our best to try to make sure that measurement error would not confound these, though I'm not sure that we succeeded on this particular measure because of the way in which those questions are asked.

We actually ended up doing okay in terms of voter registration, but I think that many researchers are very skeptical about the voter registration numbers that are in the Current Population Survey's post-election supplement, and so depending on how you calculated it, it's either 67 percent, or it's 75 percent. So while this result looks like a victory, it may not be.

The final three measures, though, are the ones that are most problematic. Contacting a public figure in the last year, we get an estimate that's three times what the Current Population Survey gets. In terms of volunteering, we get 55 percent compared with 27 percent in the Volunteering Supplement. In talking with neighbors in the past week, we get a 17-point inflation in the reported incidence. So these three measures of social connectedness exhibit very large bias. This is something that we did expect to see from work that Katharine Abraham, Sara Helms, and Stanley Presser had done. We also looked at the databases that we were able to incorporate into the design and compared the responders and the non-responders. We did a fair amount of work to try to validate the estimates. That is, when we got respondents and we had the data from the database and the data from the respondents, we found very high levels of accuracy, very much in line with the chapters by Jon Krosnick in this volume. That was true on party affiliation, and it was true on many of these other kinds of measures to the extent that you could line them up. The results are really quite positive. On a number of different measures of financial circumstance, including the extremes, the high wealth and high income and the low wealth and low income, the measures don't appear to have any particular systematic bias.

Also very gratifying to us was the fact that in terms of the split in party affiliation between the respondents that we see virtually no bias there at all. That's the result we would expect if we're able to correctly predict the outcome of elections, but it was gratifying to see that, in fact, it showed up. On another set of comparisons we do see some other issues. The people who responded voted in the database at a 54 percent rate, those who didn't at a 44 percent rate. This is evidence that there is a non-response bias here, not necessarily an over-reporting. In terms of party registration or party affiliation, the splits in the people who responded to the survey and those who didn't were virtually identical.

The topic that I want to conclude on is bias in volunteering estimates. There is a table from a 2009 American Journal of Sociology article that Katharine Abraham, Sara Helms, and Stanley Presser wrote on volunteering rates. This result was based upon an analysis of the Current Population Survey, a Volunteering Supplement, and then a subsequent analysis of a subset of respondents who were recruited to participate in the American Time Use Survey.

The results indicate that 29 percent of respondents reported that they volunteered in the Volunteering Supplement to the Current Population Survey. The respondents who actually took part in the Time Use Survey, the third line, they report volunteering at a 36 percent rate, and the non-respondents, the people that were not able to be successfully interviewed, report volunteering at a 20 percent rate. So, for context, the Volunteering Supplement respondents basically are producing about an 80 percent or an 81 percent response rate. For the Time Use Supplement respondents, you're talking about a 53 percent response rate. In the article, they make the strong case that the difference in the reported volunteering rates from people who did and didn't respond to the Time Use Survey is a function of non-response bias.

This is the largest known instance of nonresponse bias in surveys in my field. It's a very worrisome bias because I have been doing volunteer and civic engagement research for many years, and I think that much of what I've discovered is probably not right because of the bias. Certainly any trend estimates over time in an era of growing non-response are going to be suspect if this conclusion is correct.

There is a silver lining here, which I think is important to note, and that is that the correlates of volunteering were seemingly unaffected by this bias in the overall reported level. That is good news because it suggests that you can use a survey that has a lower response rate. You may have a massive inflation in the level of reported civic engagement, as we apparently do in our surveys, but it may be the case that any analysis that we do on what predicts it and what causes it and what its correlates are is not going to be biased, at least to the extent that what Stanley Presser and his colleagues found is true.

Let me conclude with just a couple of points about what we can do about this and maybe one other observation about the consequences here. Much of the focus of the research in the area of non-response bias has been on what is the bias? How sizeable is it? I think it's also important to note that something else is happening here. With these trends in non-response, all of us in the business are having to devote considerable effort to fighting against those trends. It is much harder to contact people today on the telephone than it used to be, even with the greater effort, the bigger calling regimens and the use of incentives and other things that we may or may not be doing. We are still losing the battle in terms of the non-response rate trend, but we'd be losing it a lot worse if we weren't doing that. But what this also means is that we're putting a lot of effort into trying to prop up our response rates and keep our non-response down. Because survey research is fundamentally a craft of trade-offs, those resources are coming out of other things that we could be doing that might reduce total survey error. So, we have to make a calculation of where our money is most effectively spent in the pursuit of good survey estimates. I mean, no doubt about it, 9 percent response rates for public polls creates an appearance problem, a credibility problem, but 30-something percent response rates in the National Household Education Survey also does that.

The question that we have to confront as researchers and that I'm hoping that we will be able to do more work on is what could we be doing with the money that we're using to prop up our response rates that might be more valuable to us, more pre-testing, bigger sample sizes, and the like? So, it's really critical that we have a better understanding of the circumstances under which non-response bias occurs so that we can most effectively use our resources, the kinds of work that has contributed to our understanding of non-response bias thus far needs to continue. In particular, we need to do more with the available databases matched in with our samples than we've done in the past. This is a resource that wasn't available to us 5 years ago or 10 years ago to nearly the degree that it is now.

The quality of these databases, the amount of information contained within is growing, and it does offer us the opportunity to do more with the assessment of bias than we've ever been able to do in the past, but not a lot has been done with it. That is an area of further investigation that is well worth supporting. I think some smaller scope work is still very useful. The idea of seeding your samples with households with known characteristics to observe their response propensities remains a very smart technique and one that oftentimes we don't think about in the course of trying to get the work done. I'm hopeful that the National Science Foundation and other funding agencies will be partners in the continuation of this work. I think for all of us in the practicing survey business we just need to think more about how our own studies can contribute to this body of knowledge.

Scott Keeter is senior survey advisor for the Pew Research Center in Washington, DC. He is a past president of the American Association for Public Opinion Research and the recipient of the AAPOR Award for Exceptionally Distinguished Achievement. Since 1980, he has been an election night analyst of exit polls for NBC News. Keeter's published work includes books and articles on public opinion, political participation and civic engagement, religion and politics, American elections, and survey methodology. A native of North Carolina, he attended Davidson College as an undergraduate and received a Ph.D. in political science from the University of North Carolina at Chapel Hill. He has taught at George Mason University, Rutgers University, and Virginia Commonwealth University.

48

Optimizing Response Rates

J. Michael Brick

Much of this chapter will reinforce things written by other authors in this volume, especially those by Eleanor Singer and Roger Tourangeau. This chapter is nominally about optimal response rates, but I don't know exactly is meant by the term "optimal response rate." I've thought about this several times in the past because it comes up all the time in request for proposals from the federal government. These documents often say they want optimal response rates and I am left scratching my head without some more detail.

There are two possible conceptualizations of optimal response rates that I want to focus on. The first is maximizing the overall response rate and the second is a little bit different. It is based on research that has been done since about 2000 – when Scott Keeter published a related article. The idea is that the response rate is not the thing we should be most worried about, it's the nonresponse bias that should concern us. Thus, the second conceptualization is about optimizing response rates to minimize nonresponse bias.

Starting with the first concept, where the goal is to maximize the overall response rate, in this case, what we're interested in are design features that will achieve that goal within a fixed cost. There are ways to maximize response rates if you have an unlimited budget, and unfortunately nobody is in that situation. All of my discussion will be under the assumption that there is an approximately fixed cost.

J.M. Brick (✉)
Westat, Rockville, MD, USA
e-mail: mikebrick@westat.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_48

There is a lot of literature about how design features affect response rates – hordes of factors are associated both with response rates and with cost. I couldn’t find exactly the list I wanted in the literature, so I compiled a list of key factors that I consider in the design phase. This list includes the usual suspects such as sponsorship, content salience, and mode. These are the sorts of the things we’re faced with that we can rarely change if we want to do a survey. The sponsor, the content, and the salience are all largely determined once the decision is made to conduct a survey and we can do little to modify them. In nearly all cases, a particular mode is implied by the nature of the survey. For example, the two-and-a-half-hour survey where the researcher wants a blood specimen study is normally a face-to-face survey. You don’t do that type of survey by phone too often. Roger Tourangeau put it nicely when he writes elsewhere in this volume that the mode choice is entwined with everything else. If the budget is $200,000, then we might say, “Oh, you didn’t mean blood, you meant saliva, right?” We sometimes can change the sponsors or they can change their contractor is I guess maybe the right way to put it.

Other factors include the things that we as survey designers have a lot more control over. Material design and length – if you’re talking about mail, this might include things like the size of the envelope. When thinking about web designs, there are issues such as deciding on one question per screen versus the multiple items per screen design – we could send out a mail questionnaire with one item per page but we just don’t do that for a host of other reasons. The material design is very important across modes. The field period is one of the things we always worry about – if we had more time to finish the work we could get a higher response rate.

One issue not discussed much is the number of contact attempts, and it’s a huge factor. If you do refusal conversion, how many conversion attempts are appropriate? Don Dillman made many good points in his original book in 1978 about the number of attempts and how important that is to getting a high response rate. The training of staff is also important, whether it’s interview staff or other types of staff. In the production of a mail survey or web survey, there are staff involved with material design and they need to be trained on the appropriate content and structure.

One constant tension is the allocation of resources. Even though we do know that many factors affect our response rates, it turns out that the relative importance of these factors varies tremendously across the surveys and the types of surveys. Household surveys, provider surveys, and business surveys are very different even if you just look at the incentive structure. The incentive you would use with a household survey that is a 30-minute interview is totally different from the incentive you would consider for a medical provider. It may also be very different for a teacher. Now consider the method of contact for these survey types and immediately you see they are not comparable when you are considering what to do in order to get high response rates. Refusal conversion procedures differ by whether the survey is cross-sectional or longitudinal and by the burden of the survey. All those things are highly important in terms of figuring out the appropriate allocation of resources – how much you want to spend on which type of factor.

Monetary incentives are a key factor and those have been discussed in Eleanor Singer's chapter. The incentive may have to be more than token if you're doing a health survey and respondents are asked to undergo a two-and-a-half hour physical exam. In that case, it is more appropriate to pay them for their transportation, maybe for giving specimens or those types of things. But once you've done that, you've changed the nature of the interaction. It may no longer be a survey in which you're encouraging people to respond because they're good citizens and they can help. You may have changed it into an economic exchange. Incentives can have an effect on the way people view what they are doing and that is an important point to keep in mind when designing the research.

The material design is probably the most underrated factor, and it is one of the most important factors in terms of response rates. If you send out a poorly designed survey, say one that starts with a household roster that takes a long time to complete, that is going to have a negative effect on the response rates. It affects response rates directly through the respondents and, also, if it's an interviewer-mediated survey, it will affect the interviewers and that will be passed on to the response rates.

We did an experiment in a RDD telephone survey back in the mid-1990s in which we had one group of interviewers administer a screener that lasted the typical 2–3 minutes to determine if there was a child in the household so we could interview a parent about the child. In the other arm, we wanted to think about using the screener data more generally to avoid throwing away all those households that were not eligible because they did not have a child. A 10-minute interview was developed and implemented for every household, covering items about library services and whether they went to the museum and similar items. The response rate difference was about 10 percentage points between those two interviews, and it was virtually all associated with the interviewer. The experiment was set up in different interviewing centers so there was not any cross-contamination and the interviewers had the same training, the same background, and even the same response rates from previous similar surveys. The only difference was the interviewers in the one arm knew that they had to get through this 10-minute (somewhat uninteresting) interview in order to get to the point where they selected whether the household had a child for an extended interview. This provided evidence that interviewer knowledge about and attitudes toward the survey design can have an effect on response rates. Other factors are also important, such as the interviewers' training and selection. Again that can be classified into material design in some respects.

Once again, the number of contacts and conversion attempts are critical; one of the things that cannot be overlooked in household surveys. If you don't go back often you won't get a high response rate. Also, we know that when you send a mail survey out more than once, it increases the response rate. When you call back and you do a refusal conversion attempt in telephone surveys, you get about a cooperation rate maybe about half of what you got the original time. That's important if you want to get an overall high response rate. While there is not a lot of evidence that advance letters make a dramatic difference, they do have some level of effect on response rates.

It is also important to understand that these are not independent factors. I think this is where the leverage salience theory is a nice visual of what's actually going on. The factors interact with each other. The approach or strategy should be to use this combination of factors and manipulations. So it's not a simple process of choosing them, and which ones go together well. If our goal is to get the overall maximum response rate, we know some things that work. One option is to cherry pick. Who do we choose to expend resources on? The most cost-effective approach is to get the ones that we can get to respond most easily. Who are those people? Well, they're older, they're female, they're homeowners, they're English speakers, they're upper-middle income – they're the same people we got to respond without doing much special.

The fact is that this is what we've done for many surveys over a long time, even if we have not thought about it that way. When we need to get the response rate up to 70 percent – nowadays, maybe 30 percent or 10 percent – what do we do? The answer is to get the easiest ones and it makes economic sense. If we emphasize response rates to interviews, which many researchers and organizations do, it is likely to have this effect whether we intend it to or not. I think we have to do better if we want to improve the quality of the estimates.

We don't need a whole lot more female homeowners who don't live in the central city and speak English. Why not? Because it doesn't change the nonresponse bias, which is really the goal that we should be aiming for; a high response rate is a proxy for it. Of course we are also trying to make sure the yield in terms of the number of interviews is sufficient, but what we really are trying to do is to reduce nonresponse bias. Getting more people who have a high propensity to respond isn't going to reduce nonresponse bias. So the new paradigm is to choose your strategy or optimize response rates to minimize the nonresponse bias.

A host of papers, I credit Scott Keeter's with starting the flood that we've talked about here, are pretty convincing that response rates alone are not predictive of nonresponse bias. When Groves came out in 2006 with his meta-analysis, I think that was the final blow. Almost every talk I've been to, at every conference, has had that graph up. It's the coup d'état.

Unfortunately, the message many seem to take from that is that response rates do not matter. The real message is that we've got a lot of big biases out there and it is not terribly related to the nonresponse rate. There can be really big biases. Some surveys examined in nonresponse bias studies have shown relative biases of 70 percent or 80 percent. Nonprobability samples can do that for you. So we have to think about nonresponse biases and not dismiss them. Simple overall measures of a survey do not tell us exactly when those biases are present. The biases can and do vary greatly within any given study. Other work has alluded to the deterministic and stochastic models of nonresponse where the deterministic focuses on differences between respondents and nonrespondents, and the stochastic focuses on the relationship between the response propensity and the characteristic being estimated. Some formulas or models exist that can tell you what the bias is if you happen to know the characteristics for the nonrespondents, but, of course, we don't know much about the nonrespondents in most cases.

The formula we all gravitate to nowadays is the stochastic one. It shows that the covariance between the response propensity and the characteristic is key. We spend a lot of our time thinking about the means in these equations, but the mean is not the whole picture. For a total, the evidence indicates that we are always underestimating the total. That should cause you to scratch your head and say what is that all about? What do you mean we're always underestimating the total? If you're trying to estimate a total and you have a nonresponse, don't you fix it? Who would ever use an unadjusted estimator, right? If you get a 60 percent response rate, you're going to adjust the weights by at least one over 0.6, aren't you? You get something like the post-stratified estimator. It gets a lot more complicated now. The issue is that the nonresponse bias is different depending upon the type of statistic being applied. The bias isn't always simply the difference between the mean of the respondents and nonrespondents. It depends on lots of different things. For an odds ratio, it doesn't depend on the difference at all. For a ratio, it depends upon a different function of the data. It's complicated.

The message out of all of this is that nonresponse is not simple. It's a function of the response propensities in the domain being estimated. So if you're trying to estimate a certain characteristic, the bias depends on how their response propensities differ from the other people. The type of estimate and the auxiliary data all play a role. We are faced with these situations all the time in probability samples. We try to fix up the weights for coverage and nonresponse and we use different auxiliary data to do this. But we always adjust using auxiliary data. The secret about probability sampling is that it seems to and we don't know exactly why and when.

Now the question is when most surveys are multipurpose and you have all different kinds of estimates being produced, how do you optimize for nonresponse bias? It is not simple. And why are we getting results like Scott Keeter got that it doesn't matter much. It should matter, shouldn't it? The response propensities have to be varying. So why aren't we seeing more large biases from some of the surveys? I think the main problem with optimizing for nonresponse biases is that we don't understand the reasons for bias very well. Surveys like those done at the Pew Research Center are good examples that, despite having pretty low response rates, there's not a major source causing the estimates to differ from a random sample. There's not a specific cause out there that is leading to the nonresponse. The household population being surveyed is not saying, "I'm not going to answer questions for Pew because I have a particular opinion or characteristic." For a bias to be generated, you have to have a direct cause or a correlate of a direct cause. Of course, you can get random biases but they are not reproducible. Some statistics seem to bounce up with big biases, but they're probably related to something that either the survey does or the respondent does.

Suppose I'm trying to estimate a proportion. I'm going to try to estimate how many people go fishing, or how many people volunteer – that's a good one. How many volunteer to do something altruistic? So if I'm trying to estimate that, what it depends upon is the difference in response propensities between those who volunteer and those who don't volunteer. It's a ratio between those two propensities. So here's a specific estimate – 45 percent of volunteers will respond to the survey. The other people, those who don't volunteer, we will assume 30 percent of them respond. There are some results from the American Time Use Survey that may be similar to these, where people who are altruistic and volunteer respond at a much higher rate to surveys in general. The response propensities for the volunteers and the nonvolunteers are very different – volunteers are 50 percent more likely to respond. That drives our bias for some specific statistics. If 25 percent of people were volunteers, we'd estimate 32 percent due to the higher propensity to respond for this subgroup. That's a 27 percent relative bias.

If the differences in response propensities aren't huge you are not going to have big relative biases. In order to get a big bias due to nonresponse you need a big difference in response propensities. You can get a big bias with a relatively small difference in response propensities if you're trying to do estimate a 1 percent statistic, but that is not the usual type of estimate except for things like unemployment. The data can sometimes show us what is happening but it doesn't tell us why. Furthermore, it doesn't tell us when we're going to expect big differences in response rates. That's what's causing us so much uncertainty. We don't understand direct causes and the variables that are highly correlated with those causes. Nora Cate Schaeffer has a paper with Jen Dykema that provides a great example of where there were two sources of biases. One set of biases was associated with whether you could contact the people, and another set of biases was associated with whether they refused or not. Their paper clearly shows the biases can partially balance each other in some cases, but not completely. When we try to fix these biases from different sources together, we sometimes make it worse. That is one of the consequences of weighting sometimes.

We do understand some things about the sources of bias and the direct causes. From a psychological point of view I would point to Roger Tourangeau's book on the *Psychology of Survey Response*. I think a lot of the books by Dillman, Stoop, and Groves and Couper talk about how survey operations are associated with nonresponse bias. Unfortunately, it's not a very comprehensive theory. I think one of the highlights for me in this whole sea of research on response rates was the first of the paper by Groves and his colleagues on trying to create nonresponse bias in estimates by manipulating design features. They surprisingly did a lousy job of creating nonresponse bias. They found big biases in one survey out of three or four despite the different things they tried. But if you can't produce it, you don't understand it. I think that is the essence – if we can't go out and produce a nonresponse bias, we don't understand it very well.

In the future, I think we need more comparative analysis of respondents and nonrespondents. Unfortunately, most of the research we do is sort of like that done when the *Challenger* space shuttle went down. Everybody looked at the data on the *Challenger* and they couldn't figure out what was the problem. They were looking at the flights that failed, and they graphed the failures by the outside temperatures and it was all just noise. When they finally included the flights that didn't fail, all of a sudden you saw the picture of the difference between the flights that failed and the flights that didn't fail. I think that we've been doing too much looking at the nonrespondents and saying, "Oh, I don't get it." We need to start looking at the nonrespondents and the respondents together to understand why they're different and how they're different. In a paper with Doug Williams, we went back to the same survey, 30 years apart. And the categories and relative magnitudes of the reasons for nonresponse didn't change. A big part of the problem is that you're asking the sampled person the analytic question, "why didn't you respond?" I don't think that this type of study is going to advance the field. We have to find more indirect questions rather than asking them to do the research for us and to tell us why they didn't respond. We also need to focus on the things we can manipulate as survey designers, realizing we can't change the respondents.

Here is a second suggestion – we need to test the theories in practice and the work that Groves started on generating nonresponse bias is really important. I don't know that anybody is doing that now. I think we need to prove we can produce nonresponse bias before we can say that we can fix it. If we do that, then I think we can go to the next step and say, "Can we actually change who responds when we do something differently?" Eleanor Singer's work is also very much on target when she writes that most of the research on incentives suggests that we can't see the difference in response and associate it with who responds only when we give them incentives. There are some things related to this in the research on topic saliency, where it does work sometimes. But there's very little other research that suggest when things work. A favorite topic is level of effort work such as that done on the Survey of Consumer Attitudes. This is something we can manipulate and stop fieldwork early. If we see no change in our estimates, why are we doing more? We need to know why before we go further down this road.

And we have a lot of interest now in responsive or adaptive designs, but I'm not sure why. If I don't know the cause of the bias, how are we supposed to fix it? How do I get this domain of low-income households to respond? The other thing I suggest we consider is the Taguchi Method. Start with low-cost things to try to get a higher response from the people that you want. Material design is one of the things we spend money on, but we ought to be spending more on that if it is going to improve the overall quality of the survey.

Being a statistician, I suggest we need to link the statistical adjustments we make at the end to the data collection procedures. I don't think we have a very good overall concept of doing that today. Even though I've spent much of my career making these adjustments, I've been doing it blindly in most case. Just like everyone else. Bob Groves and I had time when we did the conference circuit with our nonresponse bias short course for a recurring dinnertime discussion. Bob would say, "We have to do this to increase the response rate for specific groups to lower the bias." I'd reply, "Well, no, I can weight for that. If you tell me we are getting a lower response rate from low income then I can adjust the weights to reduce the bias for it." We never came to a full resolution with that discussion. It was always fun, but we don't know the answer. Can we fix it in data collection or do we have to wait to the end to fix it? Do either of the two methods really help?

The last point that I would like to make is on identifying methods that improve the quality for different types of statistics, and this is especially essential for general-purpose surveys. This is one of the things that probability sampling has apparently done successfully. Nonprobability sampling has not done it at all successfully thus far. We ought to unlock the secret to the success of probability samples. If we knew what we are doing that worked, and we think the methods actually do work pretty well, then I think we might be able to help with some of these other low-cost methods of data collection. This type of advance might enable these studies to get over the hump that prevents them from achieving more bias reduction for multipurpose estimates with their weighting adjustments.

In conclusion, we need to be clear about what we're optimizing. There are choices, and the choices may give different results. We need a program of research on these issues, not a particular study. We don't know enough to have one study to tell us the right answer, and the program that is developed needs to lead to theoretical as well as empirical results.

Dr. J. Michael Brick is a Vice President at Westat where he is co-Director of the Survey Methods Unit and Associate Director of the Statistical Staff. He has more than 40 years of experience in survey research, with special expertise in sample design and estimation for large surveys and in the investigation of nonsampling errors in surveys. Dr. Brick has published in numerous journals and is a Fellow of the American Statistical Association, and an elected member of the International Statistical Institute.

49

Data Collection Mode

Roger Tourangeau

Introduction

This chapter focuses on the modes of survey data collection that are currently used for a range of different types of surveys. In particular, I focus on the potential impacts that the choice of mode has on the data. For the first several decades of survey research, most surveys relied on just two modes. They were either done by mail or by face-to-face interviews. Different sectors of the industry started to add telephone as telephone coverage improved. Telephone surveys only become common in the federal statistical system in the 1970s. A paper by Massey and Thornberry shows how telephone coverage increased until the tipping point was reached in the early 1970s, where 90-plus percent of American households had telephones. But before then, the coverage was not so great, and telephone surveys were done less often.

And, over time, there's been an evolution of the different methodologies. Beginning with the traditional triad of modes of data collection (say, from the 1970s or so), consisting of face-to-face, mail, and telephone, there have been two waves of change. In the first wave, computers replaced paper as the basic medium on which surveys were conducted. With telephone surveys, in the mid-70s, survey researchers were switching to computer-assisted

R. Tourangeau (✉)
Rockville, Maryland, United States
e-mail: RogerTourangeau@westat.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_49

telephone interviewing (CATI), which becomes more and more popular, eventually driving out paper-based telephone surveys.

Similarly, as the computers got lighter and it was possible for interviewers to carry them around, survey researchers started doing computer-assisted personal interviewing (CAPI), and CAPI drove out paper face-to-face interviews. The mail process has been a little slower to change, and a lot of surveys continue to use paper mail surveys. Still, many researchers believed that the web would replace mail. All of these changes involve a shift from paper to the computer.

And there has been a second shift, in which the computer has replaced the interviewer as the collector of the data. The advantage of the computer is that it allows much more complicated questionnaires with automatic routing to the correct follow-up question, things that are very difficult or impossible to do with paper. But there are additional advantages if the respondent can interact directly with the computer; these are the reduction of interviewer effects and the improvement in the reporting of sensitive information.

There are a few different forms of computer-assisted self-interviewing. Sometimes the computer presents the questions as text, but more often there is both visual and auditory presentation of the questions. A computer-administered interview can also include images or video. Some researchers, such as Fred Conrad at the University of Michigan, have experimented with having virtual interviewers conduct interviews. The latest in this succession of changes, as large segments of the population came online, are web surveys. Web surveys are sometimes used as an adjunct to other modes of data collection. So, for example, in a mail survey respondents will sometimes be given the option of completing it on the web instead of doing it on paper.

Web surveys are particularly popular for marketing and business surveys, and as Gary Langer pointed out earlier in this volume, a lot of these surveys are based on non-probability web panels. Some researchers have seen web surveys as an attractive alternative to mail, where as others have seen it as a less expensive alternative to telephone surveys. Partly reflecting this split, there have been two design traditions that have already evolved in the web literature.

One tradition is to have web surveys follow many of the same conventions as mail surveys. Mail surveys are a mature and successful method; why not try to replicate that success on the web? Others see web surveys as an outgrowth of computer-administered surveys and they want to take advantage of all the things that computer administration can provide. In a dynamic web survey, researchers can have all the functionality of a computer-assisted survey, such as automatic routing. Automatic routing is only possible if the survey is "dynamic" – that is, information from the survey is sent back to the server computer, which then has the information needed to tailor the next question or questions.

The limit in the dynamic approach is to have a single question per screen, which is the way many interviewer-administered automated systems worked. If there is only one question on each screen, there's no question about what the respondent (or interviewer) is supposed to do.

To make web surveys more appealing, many web surveys add photographs, color, complicated background patterns, and other extraneous visual material. Such embellishments are likely to have unintended consequences (Tourangeau, Conrad, and Couper 2013).

Mode and Non-Obervation Error

Survey mode, construed narrowly, is the method of data collection, in which the respondent interacts with either a live interviewer or a computer and provides data. But in fact, modes are really a bundle of features. For example, a method of sampling and a frame usually go along with a particular method of data collection. Most CATI surveys, for example, start off as random-digit dial (RDD) samples, and that implies a certain sampling frame. Throughout most of the history of survey research, there has been a similar bundling of methods of data collection with frames and methods of sampling.

The bundle of features associated with a specific mode of data collection comes with a whole suite of measurement characteristics, including different forms of non-observation error or observation error. The distinction between these two basic forms of error appeared early on in the literature on total survey error and is useful here. Non-observation error comes about because the researchers do on observe every member of the population and those that are observed may or may not be representative of the full population. Researchers often distinguish forms of non observation error: sampling error, coverage error, and non response error.

Coverage is often implicated in the choice of a mode of data collection because modes are linked with particular sampling frames, but even when a different frame is used, the only ones who can participate in a telephone survey are people who have a telephone; similarly, people without Internet access and a computerare generally excluded from web surveys. There are inherent access issues with the different methods of data collection. The different modes have also traditionally been associated with differences in response rates. In earlier eras, face-to-face surveys had much higher response rates than telephone surveys, which had much higher response rates than mail surveys, but now that's changed a bit.

Regardless of mode, there is almost always random sampling error, but the method of data collection might be related to the level of sampling error. Face-to-face surveys generally make use of cluster sampling, and that drives up the standard errors of the estimates. With opt-in web panels, there are often selection biases from the use of non-probability samples.

Then, there are observation errors. The different methods of data collection can produce systematic differences in random measurement error. There are certainly differences across modes in the level of interviewer effects. In a mail survey, for example, there are no interviewer effects. There could also be mode-related differences in response order effects or social desirability. Apart from its effects on the error properties of surveys, the mode might affect the cost or the timeliness of the survey. One of the great appeals of web surveys is the speed with which they can be done.

How do researchers select a mode of data collection? A key consideration, often swamping all other considerations, is the survey's budget. Face-to-face surveys are generally much more expensive per case. To reduce cost, such surveys usually involve clustered designs. Clustering increases sampling variance, but reduces cost. Geographic clustering of cases is not usually necessary in telephone, mail, and web surveys.

Before list-assisted sampling, telephone samples were also clustered, but list-assisted sampling eliminated the need for that. With self-administered modes, such as mail or web, based on a list sample, distance doesn't affect the cost. To date, no method of sampling or frame has become dominant with web surveys. Some web surveys use list samples, such as students at a university; some use panels of volunteers; and some use samples of addresses, to which mail invitations to participate are sent.

Coverage error, one of the forms of non-observation error, is likely whenever units in the population are not included on the frame; whenever there are systematic differences between the units that are covered and those that are not, coverage error is the result. Web surveys generally exclude those without Internet access.

There is a similar issue with single-frame telephone surveys, that sample only landlines or, more recently, only cell telephones. The exclusion of a part of the population can lead to systematic errors. For example, sampling only landlines excludes the cell-only population, reducing the representation of young people. The elderly are less likely to be represented in a web survey than younger respondents and college graduates are likely to be overrepresented.

A study by Dever, Rafferty, and Valliant (2008) found substantive differences between those with Internet access and those without. They examined data from Michigan's Behavioral Risk Factor Surveillance System (BRFSS) survey, which is done by telephone. They compared respondents who indicated that they had access to the web with those without web access and found that there are some health differences. For example, those with Internet access were much more likely to rate themselves as having good or excellent health than those without access.

A similar study by Schonlau and his colleagues (Schonlau, van Soest, Kapteyn, and Couper 2009) compared respondents to the Health and Retirement Survey (HRS) who had web access with those that did not. The HRS sample consists mainly of people aged 50 and older and some 55 percent of all respondents reported high blood pressure. However, among those respondents who reported Internet access, the figure was considerably lower. So the coverage of the American population on the Internet is not complete, and there are both demographic and non-demographic differences (including differences in health) between the ones with and the ones without Internet access.

Tourangeau, Conrad, and Couper (2013) describe the different forms of non-observation error this way: suppose one has a volunteer sample, an opt-in web panel. These volunteers may not represent the entire Internet population, and that difference is a form of sampling bias. In addition, the respondents may not represent the volunteers; this is non-response bias. Finally, the Internet population may not match the actual target population; this is coverage bias. How much do the three forms of non-observation affect the survey estimates?

Some evidence comes from a study by Chang and Krosnick (2009), in which they presented weighted and unweighted estimates from an RDD sample, a survey using the Knowledge Network panel (which is a probability sample with minimal coverage problems since Knowledge Network provides computers and Internet access for those that do not have them), an opt-in web panel, and the Current Population Survey (CPS), which is a very high quality interview survey. The conclusion was that the unweighted estimates of the demographic items from the opt-in panel were far off from those from the CPS. The opt-in panel suffers from all three forms of non-observation error. It is not a probability sample; it excludes adults without Internet access; and it has low response rates.

A number of studies have examined whether weighting adjustments can compensate for non-observation errors (for a summary, see Chapter 2 in Tourangeau, Conrad, and Couper 2013). The answer seems to be that weighting generally removes only part of the non-observation bias. Thus, the key statistical consequences of non-observation error are inflated variance and bias.

Weighting to Reduce Non-Observation Biases

There are two components to the bias produced by non-observation error. The first reflects the exclusion of population members with no chance whatsoever to get into the survey. For example, people without Internet access, or those who would never join a panel who have zero probability of being included in a sample based on the panel. So, the first component depends on the proportion of population members with no chance of being and the survey and the difference between the zero-probability population members and those with a positive probability. This is the deterministic component of the bias.

The second component of the bias due to non-observation error is non-deterministic; it reflects the differential chances of inclusion for those population members with some, non-zero chance of inclusion in the sample. If these inclusion probabilities covary with the survey variable of interest, this produces a second component to the bias. It is possible that the two components of the non-observation error would wholly or partially offset each other.

In the Dever, Rafferty, and Valliant (2008) study, the researchers compare the original BRFSS sample with just the BRFSS Internet users. The study simulates the coverage bias from dropping all those with zero chance of appearing in a web version of the study. They examined 25 estimates using generalized regression estimation to reweight the data. Their goal was to see how well the reweighted data compensated for dropping the cases without Internet access.

On average, the sophisticated weighting method they used eliminated only about a quarter of the bias. Several other studies have done similar things (see Table 2.4 in Tourangeau et al. 2013).

Applying weights does not always improve the estimates. It can increase the bias and usually increases the variance of the estimates as well. Various weighting methods have been used to reduce bias due to non-observation. For example, raking, which just means adjusting the weights iteratively to agree with marginal totals from some external survey, usually the CPS or the American Community Survey (ACS).

Propensity scoring is another method researchers have tried. But the empirical summary seems to be that weighting usually helps, but it does not eliminate the bias. It may be worth doing, but it does not represent a complete cure.

Mode and Observation Error

In their analysis of mode differences, Groves and his colleagues (Groves et al. 2004) argue that there are five major properties of the mode of data collection that affect observation or measurement error. One key variable is how much interviewers are involved. The mere presence of the interviewer seems to have an impact on respondents' willingness to provide sensitive information about themselves. Even if an interviewer is involved, he or she may not interact directly with the respondent. For example, the key data source may be administrative records and the interaction between interviewer and respondent may be limited to obtaining the respondent's permission to access the records. A third characteristic of the mode is the level of privacy it affords. Many studies have shown that interviews are often conducted with other people present besides the interviewer and the respondent(see Mneimneh et al. 2015 for a recent example), and that has consequences for the answers respondents give. A fourth characteristic of mode that can affect observation error is how the information is communicated to the respondent, either by the oral channel, the visual channel, or both, and how the respondent communicates his or her answers. Finally, there is the technology used – paper or computer. In a similar model, Tourangeau and Smith (1996) argue that there were four key features that distinguish modes of data collection. Were the questions self-administered? Did the initial contact come by telephone? Is it computerized? Is the material presented to the respondent orally or visually? According to Tourangeau and Smith, these mode characteristics affect three psychological variables – how impersonal the data collection setting seems to respondents, whether survey seems legitimate to them, and how much cognitive burden it imposes.

Both the Groves et al. model and the one by Tourangeau and Smith indicate that computer assistance is an important feature of mode (and in fact many, if not most surveys are now computer assisted).

One benefit of computer assistance is reduced item nonresponse. A study by my colleagues and me (Tourangeau et al. 1997) compared four modes of data collection – paper and pencil interviews, paper self-administration, computer-assisted personal interview, and computer-assisted self-administration. On average, we got 98.6 percent of the data when the computer administered the item, whereas 3.8 percent was missing when the survey was done on paper. This was the same as the difference in missing data rates between self- and interviewer administration. There were two main effects here. Computerization and interviewer administration both produced more complete data. So one useful thing that interviewers do is getting the data – that is, reducing the rate of missing data.

There have also been a number of studies that have looked at the benefits of self-administration, and, in particular, audio computer-assisted self-interviewing (ACASI). This literature started out with a small volunteer sample in North Carolina, but the last couple of studies involved national area probability samples (Epstein et al. 2001; Turner et al. 1998). These studies converge on the conclusion that people report more sensitive or embarrassing items in ACASI. Ting Yan and I (Tourangeau and Yan 2007) did a meta-analysis that looked at the issue of whether different methods of self-administration differ among themselves. Only a few studies compare more than one method of self-administration. We found that the mean effect size for computerization was positive indicating an increase in reporting with computerization, but the difference was not significant.

Multimode Surveys

The proliferation in the number of methods of data collection means that researchers have a lot of options, and more and more surveys are also giving the respondent multiple options. There are several ways in which multiple modes have been used in surveys. One variation is used with cross-sectional surveys. Often, a cross-sectional survey starts with the cheapest mode, and reserves the more expensive modes for the non respondents to the less expensive mode. So, for example, the American Community Survey starts as a mail survey (recently, with a web option offered in the initial mailing) with telephone follow-up with the mail/web non respondents, concluding with face-to-face follow-up with a subsample of the telephone non respondents. The progression is from less to more expensive modes. Often, in a longitudinal survey, multiple modes are used, but in the opposite order. In the first wave, where a high response is very desirable, face-to-face may be used, and then in the follow-up rounds, cheaper methods are used.

Unimode versus Best Practices

The use of multiple modes of data collection raises the issue of whether to attempt to minimize measurement *differences* across mode or to minimize the overall *level* of error. Some researchers seem to think of mode differences as a form of measurement error so that it is always desirable to minimize them. They advocate a “unimode” design. The alternative view is that mode effects represent, in part, differential measurement error, as in this model:

$$wb_A + (1 - w)b_B$$

In the previous equation (from Tourangeau, Conrad, and Couper 2013), there is the average bias in mode *A* (b_A), the proportion (w) perhaps weighted of respondents who did the survey in that mode, and the remaining respondents completed the survey in mode B with some measurement effect represented by b_B. Based on this equation, how can one minimize the total measurement error in this situation? It is not clear that the best way to ensure minimal error is to equalize the mode effects in the two mode groups. The alternative point of view is that the best strategy is to minimize the error in both modes, thereby minimizing the weighted sum of the errors.

To see the practical implications of this theoretical dispute, suppose the two modes are telephone and face-to-face. In the telephone mode, show cards are impossible, but they can be used and improve the measurement in the face-to-face interviews. The unimode approach would seem to imply show cards should not be used in either mode; the best practices approach would imply that they should be used in the face-to-face interviews.

Still, the unimode approach may make the most sense when comparisons across groups are more important than overall estimates. For example, in a patient satisfaction survey that compares satisfaction ratings at two hospitals, it is important that the differences across the hospitals represent real differences in patient satisfaction, not mode differences. Minimizing mode differences may thus take priority over minimizing overall error in this situation.

Conclusions

I would like to conclude this chapter with some suggestions about topics that should, in my view, be priority areas for future research.

One such area involves improving reporting of potentially embarrassing information. Comparisons of survey reports with records data (e.g., Kreuter et al. 2008) indicate that self-administration *improves* reporting but does not eliminate underreporting. We need additional tools for reducing measurement error when the questions are sensitive and respondents are motivated to misreport.

A second priority area involves the impact of screen size on responses to web surveys. Tourangeau, Conrad, and Couper (2013) concluded that web surveys generally have good measurement properties, but that conclusion largely rested on studies in which respondents completed the survey on desktop or laptop computers. More and more web surveys are now being completed by respondents on their smartphones and tablets, which have much smaller screens. It is important for us to know what effect this switch will have on the level of measurment error.

Another important shift in surveys involves the switch from telephone data collection from RDD samples versus mail data collection with address-based samples. It remains unclear how this shift will change the overall level of non-observation and observation error. It is a priority to find out.

A final major challenge involves distinguishing the effects of non-observation errors from those of observation errors in mode studies. That is, mode affects both *who* responds and *how* they respond. There needs to be continued work to understand both effects and to develop models for separating them.

References and Further Reading

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*, 641–678.

Dever, J. A., Rafferty, A., & Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, *2*, 47–62.

Epstein, J.F., Barker, P.R., & Kroutil, L.A. (2001). Mode effects in self-reported mental health data. *Public Opinion Quarterly*, *65*, 529–549.

Kreuter, F., Presser, S. & Tourangeau, R. (2008). Social desirability bias in CATI, IVR and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72, 847–865.

Mneimneh, Z. M., Tourangeau, R., Pennell, B.-E., Heeringa, S. E., & Elliott, M. R. (2015). Cultural variations in the effect of interview privacy and the need for social conformity on reporting sensitive information. *Journal of Official Statistics*, *31*, 673–697.

O'Reilly, J., Hubbard, M., Lessler, J., Biemer, P., & Turner, C. (1994). Audio and video computer assisted self-interviewing: Preliminary tests of new technology for data collection. *Journal of Official Statistics*, *10*, 197–214.

Schonlau, M., van Soest, A., Kapetyn, A., & Couper, M. P. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods and Research*, *37*, 291–318.

Thornberry, O, & Massey, J. (1988). Trends in United States telephone coverage across time and subgroups. In R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, & J. Waksberg (Eds.), Telephone Survey Methodology (pp. 41–54). New York: John Wiley.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The Science of Web Surveys*. New York: Oxford University Press.

Tourangeau, R., Rasinski, K., Jobe, J., Smith, T., & Pratt, W. (1997). Sources of error in a survey of sexual behavior. *Journal of Official Statistics*, *13*, 341–365.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*, 275–304.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859–883.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, *280*, 867–873.

Roger Tourangeau is a Vice President in the Statistics Group and co-director of Westat's Survey Methods Group. Tourangeau is known for his research on survey methods, especially on different modes of data collection and on the cognitive processes underlying survey responses. He is the lead author of *The Psychology of Survey Response*, which received the 2006 AAPOR Book Award, and he was given the 2002 Helen Dinerman Award, the highest honor of the World Association for Public Opinion Research for his work on cognitive aspects of survey methodology. He is also the lead author of *The Science of Web Surveys*. Before coming to Westat, Dr. Tourangeau worked at NORC, the Gallup Organization, and the University of Michigan; while at the University of Michigan, he directed the Joint Program in Survey Methodology for nine years. He has a Ph.D. from Yale University and is Fellow of the American Statistical Association.

50

Survey Incentives

Eleanor Singer

Twenty years ago, there was a consensus that incentives shouldn't be used in surveys less than an hour in length and there was a great debate about whether response rates were in fact declining or not. Today there is no question that response rates are declining. I think that incentives are used probably in most of the large national surveys that are being done today. I know this about 20 years ago because there was a symposium convened by the Council of Professional Associations on Federal Statistics (COPAFS) in the fall of 1992 for the benefit of the Office of Management and Budget, which was a contemplating drafting guidelines with respect to incentives. Norman Bradburn was a member of that panel and so was Dick Kulka and other luminaries. It was a very small group.

Half of the recommendations for research that were made at the end of that report, I made independently when I was writing this chapter. I don't know what that indicates about progress or the lack of it, but I found it surprising. We know that incentives can increase response rates and we know that in the surveys that we are most concerned with here, most of the loss in response rates is due to refusals and it is refusals that incentives

E. Singer (✉)
14 Haverhill Court, Ann Arbor, MI 48105, USA
e-mail: elsinger@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_50

most impact. But we know very little still about the bias that is caused by this decline in response and increase in refusals.

I am going to try to do three things. I am going to begin by summarizing quickly what we think we know about the effect of incentives on various outcomes: Response rates, sample composition, response quality, response distributions. The findings I am citing come almost exclusively from randomized experiments, preferably large ones, but not always. I'm also going to suggest how we might think about using incentives to accomplish different kinds of goals other than increasing response rates and with a list of recommendations for research. Just to be clear what kind of surveys I am talking about, I am talking about large, usually national surveys done for purposes related to social research. They are often longitudinal but not always. They are usually sponsored by government statistical agencies or research organizations or done with government grants. The intent is to generalize results to some definite population, not necessarily a national one, but some definite population. I am not talking about market research or customer satisfaction surveys or polls with a very short completion time.

So let me start by thinking about, we're talking about why people respond to surveys. To begin with, I think all theories of action that I know about emphasize the role of incentives of some kind in motivating behavior. These have been not necessarily monetary incentives. And results from open-ended questions on surveys suggest there are three kinds of reasons why people respond to surveys. One kind is altruistic motives; they want to help society or the researcher. A second category is egoistic reasons. This is a category that monetary incentives fall into. But, there are other things like I want to learn something. Or some people say, and you may not believe this, they like to do surveys.

There are also a lot of reasons that are associated with an aspect of the survey itself, the topic often, but sometimes the sponsor, often the organization that is doing the research, if there is a reputation good or bad associated with it. In other words, both theory and practice confirm the importance of some kind of incentive in motivating people to respond to surveys. So, I am going to start with the effect on response rates. Prepaid incentives yield higher response rates than promised incentives or no incentives. Monetary incentives are better than gifts and response rates increase with increasing amounts of money, but not necessarily in a linear fashion. There have been two meta-analyses, Church's in 1993 and another one by Edwards and his colleagues in 2002, and with very few exceptions more recent experiments come up with exactly the same findings.

What about interviewer-mediated surveys? Again, the initial meta-analysis found results very similar to those in mail surveys, although the effects of incentives in interviewer-mediated surveys were generally smaller than in mail surveys. When you stop and think about it, that's going to be a finding that cuts across various aspects of this. That is, interviewers to some extent compensate for incentives.

There was a study by Cantor, O'Hare and O'Connor in 2008 of 23 random digit dialing (RDD) experiments and they found prepayment of $1.00–5.00 increased response rates from 2 to 12 percent points over no incentives. Larger incentives led to higher response rates, but at a decreasing rate. The effect of incentives had not declined over time. We found in 1995 the difference made by a $5.00 or $10.00 incentive was five percentage points. David Cantor found about the same size increase at a much later point in time, but the base had fallen in the meantime. There's a steady decline in the base rate; incentives continue to make a difference, but they don't compensate for the decline in the base. Prepaid incentives for refusal conversion, and this is an important finding, had about the same effect as those sent at initial contact, but at a lower cost. Promised incentives up to $25.00 didn't increase response rates, but larger incentives sometimes did. There are some exceptions, or not exceptions but specific findings for larger surveys, those done by Michael Brick and his colleagues for example, who found that the most economical design for large screening surveys for instance might involve prepaid refusal conversion payments only plus some sampling of refusals. An experiment by Brick found that a promised incentive of $10.00 produced a higher screener response rate than $5.00 or no incentive, but advance notification of the survey and the incentive had no effect. But more research needs to be done on notification via cell phones.

What about longitudinal studies? These are generalizations so there may be exceptions that I am not dealing with. But as in cross-sectional studies, incentives increase response rates usually by reducing refusals, but sometimes reducing noncontact. So a study by McGrath shows they help to reduce noncontact. Some studies suggest that the initial payment continues to motivate response in later waves. It's not always clear whether respondents think of these later waves as part of one study or whether they think of them as discrete surveys. Obviously, the timing between waves can contribute to that and some other things will as well.

Prepaid incentives in longitudinal studies appear to increase response among those who have previously refused, but not among those who have previously cooperated. That is based on this one rather small study, but it is consistent with some findings in other contexts and about other uses of incentives. It suggests that there may be a ceiling effect to the use of incentives and Annette Jäckle and Peter Lynn looked at the methodology panel of a large longitudinal survey, and they found that payments in multiple waves reduced attrition in all waves, but they do so proportionately among different subgroups. So they did not do anything about attrition bias. They sustained the response rate, they did not address bias. The effect of the incentives decreased across waves. It continued to have some effect, but at a lower rate and that is a pretty consistent finding also. And this is the only study I have come across that finds there was actually a reduction in item response or an increase in item nonresponse across waves. There was an increase in unit response, but a decrease in item response. That as I say is the only study to have reported a finding like that.

Moving on to the effect of incentives on response quality, most studies have measured it with a couple of indicators, one is item nonresponse and the other is the length of responses to open- ended questions, and other measures would clearly be desirable. There are two alternative hypotheses about the effect of incentives on response quality. One is that you paid me some money and I am going to do this damn survey, but I am not going to work very hard at it. Another is, you've given me all this money, or you've given me some money, and I have an obligation to do my best to answer it correctly.

David Cantor and his colleagues suggest that these hypotheses really need to be tested in a context that controls a whole bunch of factors. They haven't been, in fact there have been very few studies altogether. But those that have been done have generally found no effect. No effect can conceal conflicting effects and that is what that suggestion about controlling a lot of things comes from.

But, just in time for this chapter, Becca Medway, she is a Joint Program in Survey Methodology (JPSM) Ph.D. came up with some very good findings with respect to the effect of incentives on response quality as part of an experimental JPSM survey. She looked at a large pool of measures of quality. So, item nonresponse, length of open-ended responses, but also straight-lining, interview length, underreporting to filter questions, and so on. The survey vehicle was an experiment embedded in a JPSM practicum survey with an overall response rate of about 16 percent, half of whose sample members received a $5.00 incentive and half none. The results, interestingly enough, were that the response rate was about 22 percent with the incentive and 11 percent without the incentive so the usual 10–12 percent increase. The cost to complete, which is also not often reported, was lower with the incentive than without the incentive. We have other anecdotal reports on this and the fact that it reduces cost, reduces callback efforts, and so on. You rarely get the reports of the actual cost with and without.

Medway found a significant effect of the incentive on only two indicators, reduced item nonresponse, so that's a good thing, and less time to complete the survey, which might be a bad thing but since none of the other indicators of quality showed a significant difference, shorter length is also good. However, with controls on cognitive ability and on respondent conscientiousness, even those two effects disappeared so essentially this is a study, which shows no effect of incentives on response quality. She also looked at the interaction of a bunch of demographic characteristics with the incentive to see if some groups were more likely to show such effects than others and she found no significant interactions. I didn't calculate just how big the groups were and she may not have been powered enough to find the effects. The findings are in accord with other reports of this in the literature.

One question that occurred to me as I was reading her dissertation, and it is in line with what I said before about incentives and interviewers, is whether you would find these effects in interviewer-mediated surveys. Jäckle and Lynn found greater effects of incentives on both unit and item nonresponse in mail than in phone surveys, again suggesting that the incentives have a greater impact where there is no interviewer to somehow mediate between the survey and the respondent.

There are very few studies on the effects of incentives on sample composition; almost all that have been done show no effects but there are a few that do show such effects on specific characteristics. These are for the most part *post facto* findings. They are not, with one exception, the results of experiments. So a study by Berlin, which was a study of literacy, found that you had higher education, lower education with incentives than without because you attracted more low-socioeconomic status (SES) students into the sample. Dan Merkle and colleagues found more Democrats in their sample when they offered a pen with the news organization's logo on it. Mack et al. found lower SES inclusion, so did Martin and colleagues. Groves and colleagues found a higher proportion of people with lower civic duty in the portion of the sample that received incentives than in the portion of the sample that did not. That was an experimental finding, the others I cited were *ex post facto* interpretations.

However, specific attempts to bring into the sample groups who are less disposed to respond because of lower topic interest have received only qualified support. So those are again experiments – one by Groves, Presser and Dipko and the other by Groves and a team of thousands, half of whom are here probably. And again it was a question of seeding the sample with people who are assumed to be interested in a particular topic and then looking at the effect of incentives in reducing bias in the composition of the sample and potentially in the response distributions.

Essentially what they found was that the incentives were successful in inducing people who had less topic interest to participate in the survey, but not enough to have significant effects on the sample composition or the nonresponse bias. We've been looking at sample composition, the real concern is not so much with the composition of the sample because you can obviously use adjustment and weighing procedures of various kinds to compensate for that. What you are concerned about is are you getting different responses from those you would get if you were able to eliminate nonresponse bias?

There are both direct and indirect ways in which sample composition could affect response distributions. One of them is the direct effect of incentives on attitudes. There is only one study, as far as I could tell, that seems to have tested that effect. So, do incentives directly affect the distribution of responses? They found no effects. This was a study of people who had been mental patients and this was a survey a year later of their appraisal of the treatment they had received. Some got incentives and some did not and there was no effect of the incentive on the evaluation of the treatment facility. You would expect to find this effect in customer satisfaction surveys I think, but I don't know of any studies that have looked at this. So the results here too are both reassuring and disappointing. On the one hand you don't have to be afraid to use incentives, you're not going to bias the responses directly or as far as we know indirectly. But we don't know enough about how to use them in order to counteract whatever nonresponse bias might exist in the sample.

Okay, how about Internet surveys? Comparisons are difficult because the terminology is different but in those places where you can make comparable comparisons, the findings from other modes generally seem to apply to Internet modes as well. Money has been more effective than gifts, prepaid incentives are better than promised incentives. Much of the published experimental work has been done by a woman named Anna Göritz, who finds that incentives both increase the number of invitees who start the survey and the number who complete it. Lotteries are the most commonly used incentives. But specific tests of lotteries among other incentives almost invariably show that lotteries are no more effective in web surveys than in other kinds of surveys. Göritz therefore concludes that you don't need to use lotteries at all because you don't do better than without an incentive. But another conclusion might be that you ought to be using other incentives instead of lotteries. But in any case, those are the findings. Incentives didn't affect item nonresponse or sample composition in any of the studies that looked at these effects and that is about all I have to say on Internet surveys at this point.

It is important to discuss the issue of differential incentives, and by differential incentives I mean primarily refusal conversion payments. So there are two arguments in favor of differential incentives. First is that they're more economical than prepaid incentives, and secondly they are believed to be more effective in reducing bias. There is an argument that they are really not unfair to respondents because people who refuse find the survey more burdensome and therefore they are entitled to a differential incentive because it's a more burdensome task. On the other hand, if you turn it around – it isn't that people who find it more burdensome refuse, but people who have more civic duty are more likely to respond – then the unfairness argument gets turned on its head. You are rewarding people who are uncooperative instead of the ones who are nice enough to participate without the incentive.

It turns out that most respondents do think of refusal payments, or differential incentives, as unfair. If you ask them hypothetically, they think these things are unfair. At the same time, they say they would answer another survey by the same organization, even though they've just been told that the survey is using refusal conversion payments. And in fact, they do, if you go back a year later, although with an ostensibly different survey organization and they do respond to that survey irrespective of whether or not they got an incentive the last time or whether they said it was unfair to pay incentives to people who refused. My personal recommendation for best practice in this area is to pay a small incentive upfront to everybody as an acknowledgment of their effort and our gratitude for it and then to use differential incentives for refusal conversion in order to try to counter direct bias. And by the way, I don't know if there is a lot of research on whether refusal conversion payments actually do counteract bias.

There is no good evidence for how big an incentive should be. It's all over the place and there is no best practice that can be inferred from the empirical work done so far. There have been some Census Bureau surveys that have experimented with different sized incentives and they generally find that bigger ones are better, although sometimes there is no difference between $40.00 and $20.00. But we cannot generalize from this and I think this changes year to year with the value of the dollar and God knows what else.

Secondly, there is no evidence that incentives reduce response rate. I have sometimes heard that said, but there isn't any good evidence for that either. There may however be ceiling effects. As I have said before, they seem to have greater effects on people who are less inclined to respond and in surveys where the initial response rate is low. In other words, they have an impact on people where there is some resistance to doing the survey. Relatively few studies have examined the cost effectiveness of incentives and that's an area that is in need of some further research I think.

In terms of best practices and some recommendations, there are very few best practices. This chapter has presented a summary of most of the important findings from the literature, but those are not necessarily the best practices, and there is a difference between the two. On best practices, I would say really just four things and one of them goes counter to what everyone else is saying, so be it. I think we need to spend more money. I think we have to recognize that if the aim is to do good surveys in this environment, I think we have to spend more money and I don't mean on higher incentives, I just mean we need better interviewer training, we need more pretesting, more monitoring of the quality of the interviewing, and so on. These things cost money and they cost more money every year because that is the trend in pricing. There is a lot of talk about tradeoffs. I think you can do some tradeoffs, but I think in the end you have to be prepared to do more things and spend more money doing them in order to get quality data.

Secondly I think what we need is to get used to using more theory instead of basing practice on past practice. I don't discount past practice, but I think it is better to have some theoretical justification for what you are doing in addition. Because otherwise you are sort of flying by the seat of your pants. This may work this time and you don't know why it's working and the only way you can count on it working the next time is if everything is exactly the same as it was before. Usually we don't know what the factors are, all the relevant, important factors, so we need more theory.

And then pretest. I know that's always been a recommended best practice. Secondly, I think we need to pretest more things. I think we need to pretest the effectiveness of different combinations of introductory materials and the appeals in those materials. Because different populations have different combinations of motives depending on what the survey is. I think, especially in a large survey, you need to find out with a small quantitative pretest what works and what doesn't work. Usually we invent it, we make it up, we think this thing is going to be a good introduction, it is going to make people want to participate in this survey. I think we can't rely on our hunches anymore. I think we need to test these things including how much money to pay and how that will work.

Finally, I think we need to investigate respondents' and nonrespondents' perceptions of costs and benefits of survey participation. I think the goal of that research would be to develop empirically based efforts to improve the survey experience for respondents. Again, what is it they hate, what is it they like? Do they like anything? Yeah, some of them do. I think this kind of study we ought to invest in would be a benefit to us in the long run.

Here are some recommendations for research, and as I say they are not necessarily new but I made them up independently. First of all, we need more research on how to reduce nonresponse bias for the most important dependent variables in a survey. And since almost all or all prior studies have used prepaid incentives, one recommendation would be to focus research on targeted refusal conversions and see what they actually accomplish. I think that they probably do accomplish something.

Secondly, I'm suggesting using address-based sampling rather than RDD as the first step, not necessarily in order to do a mail survey, but as a first step in a telephone survey. I know that the match rate isn't any better than if you go the other way around and you probably end up with the same people, but suppose you use that address as a means of contacting the respondent and you invite the respondent with a prepaid incentive, okay you have an address. You invite the respondent to call you back or you invite them to give you a telephone number so you can call them. You don't send the questionnaire in the envelope, you send simply an advance letter with a prepaid incentive and you see what it fetches you.

What it gets you, at least hopefully, is the ability to target further payments to people who aren't willing to take that first step or who refuse at a later step, people whose characteristics differ from those of a listed sample. That's the crucial difference. Because when you simply use either prepaid incentives or refusal conversion payments to the people whose addresses you can get you essentially bring in to your sample more people with the same characteristics as those who responded initially. That is okay, it increases sample size, but it doesn't address the issue of nonresponse bias. This recommendation would be to see whether something like this can get at a different group of people who you can then target with specific conversion efforts.

Third, I think we should measure the long-term effect of incentives on the public's willingness to participate in research going forward by adding questions about expectations for incentives to a sample of existing cross-sectional surveys like the General Social Survey (GSS) or the consumer expenditure surveys. I cannot believe that some of the resistance to surveys and the declining response rates is not due to this. I am not saying all of it, but some of it is due to some expectation on the part of respondents along the lines of "if you're really serious about this, you ought to be paying me."

The existing evidence does not support that, but it has all been done on fairly short periods of time, not a long trend of changing practices. I think we also for the same reason ought to measure interviewer expectations about the use of incentives. Because there I think there is a much clearer reason to think it might have an effect on response rates. That is, to the extent that interviewers are aware that an organization is using incentives regularly, they may come to use that as a crutch. It would be useful to measure interviewer expectations over time and correlate those expectations with their response rates.

I also think that it would be useful to do some research into reasons for responding and not responding, the public's reasons for responding and not responding. Do motives change over time? I think it would be useful to see whether the motives for participating or not participating are changing over time and whether they differ by different demographic categories. So, are altruistic motives declining in general? It would also be interesting to look into this notion of ceiling effects on incentives, why aren't they simple additive? In other words, why doesn't offering an incentive to people increase, even if they're motivated to respond, why don't you get an additional boost from the incentive?

It's as if there is a certain limit. In the 1999 study by Bob Groves and Amy Corning and myself the finding was that the total response rate did not go up, it shifted. You had more people who hadn't responded before but not any more of the ones who had. It went up a little bit, but the people who were already motivated to respond by civic duty did not get an equal size boost from money. Why not?

Relatively few studies have evaluated the effect of incentives on quality of response and most have found no effects. I think we need research that varies the size of the incentive over a much wider range and other variables like topic and mode and we ought to be looking at things like reliability and validity of responses. Becca Medway did have one indicator of accuracy and she found no effect on that in her study. I think again, we need better evidence from different kinds of indicators.

Are incentives coercive? Our Institutional Review Boards (IRBs) often ask this question. There are a couple of studies that suggest they are not, neither in biomedical nor in social research, but again they have looked at this over a very narrow range of incentives, sizes, and over a narrow slice of risks. So that's an area I think that would be worth studying and wouldn't be that hard to do.

Then we need more research on cost effectiveness of incentives. If you like, look back at the COPAFS recommendations for research and look at those that haven't been included on this list and you will get some more ideas.

Eleanor Singer's research focused on the causes and consequences of survey nonresponse. She published widely on the role of incentives in stimulating response to surveys, especially among respondents for whom other motivating factors, such as interest in the topic or connection to the sponsor, are lacking. Her research also explored how concerns about confidentiality, and informed consent procedures more generally, affect survey response. The recipient of a Ph.D. in Sociology from Columbia University, at the time of writing she was Research Professor Emerita at the Survey Research Center, Institute for Social Research, University of Michigan, a past Editor of *Public Opinion Quarterly*, a past president of the American Association for Public Opinion Research, and the author or co-author of numerous books and articles on the survey research process.**Editor's note:** shortly prior to the publication of this book, Eleanor passed away. Her loss is felt widely in the survey research community.

51

Building Household Rosters Sensibly

Kathy Ashenfelter

What is a roster? Basically, it's a head count. Once an address is sampled, we then have to determine how many people are living (or staying) at that sample address. How many people are living at the sample address? People are counted according to a different set of residence rules, which differ for each Census and each survey. This may make sense depending on what you're trying to do, whether it's a snapshot like the decennial Census or an ongoing survey like the American Community Survey (ACS).

The questions are asked for each person who falls on the roster for who lives or stays at the address according to the rules. Under coverage and enumeration may be more problematic than surveys in the Census as we have found in the past. The U.S. Census Bureau conducts over 35 demographic surveys, and most of them do use some sort of rostering technique. I am going to focus on the big ones in this chapter, so the Decennial Census, the ACS, and the Survey of the Income Program Participation (SIPP) – the examples that I will use are from wave one of each survey as applicable. The roster of each wave is updated in the next and changes are recorded as Roger Tourangeau found, in the Current Population Survey (CPS), the supplement for demographics.

K. Ashenfelter (✉)
Senior Data Scientist for Cyber Analytic Services, Unisys Corporation,
Washington, DC, USA
e-mail: ktashenfelter@yahoo.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_51

This chapter focuses on the need for consistency in rostering. There are many different rostering rules across many different surveys, which may or may not make sense when you read or hear them outside of the context of the survey's branching pattern based on the responses that the respondent has already given. Adding to the confusion is that even within the context of the same survey, instructions throughout the instrument can seem to contradict themselves, language on computer vs. paper version can differ, interviewer differences can make an impact. Some developmental research to optimize a core roster procedure through conceptual and operational work would really make a big impact on the future of survey research.

There's a need for research directly comparing the outcomes for the different approaches. Do different rosters result from the same household? We've done a little bit of work in this in the Human Factors and Usability Research Laboratory and I can say that we've seen that the answer is 'yes' on a small scale, but it would be ideal to conduct some sample-based research and see what the outcome would be.

So some of the research opportunities I present here are inconsistencies with respect to the date or dates used to decide whom should be listed as a resident. And consistencies with result to whether the respondent should take into account the purpose of the visit, why they're there. In some of these surveys that are interview-based and administered, some residence rules are seen only by the interviewers – as instructions – and not by the respondents themselves. It might be helpful if the respondents knew some of these rules that applied to them.

So for the Decennial Census, these examples are from Telephone Questionnaire Assistance (TQA), or the verbally based version of the Census that respondents would complete if they called the U.S. Census Bureau for help completing the survey. One option for completing the Census in these situations is to answer the Census questions over the phone. Some instructions say April 1, 2010, such as count the people living in this house, apartment or mobile home on April 1, 2010. But then the language sort of slips and becomes less specific. Another example instructs the respondent to indicate foster children if they were staying at the address on or around April 1.

There are also instances where we talk about people who visited for a long period of time around April 1. There, the interview says something like, now thinking of all the people you just mentioned in April, so the whole month of April, were you or was anyone else living in college housing? For example, college students and armed forces personnel should be listed where they live and sleep most of the time. And "most of the time" isn't defined so, how long is that? The paper form of the 2011 ACS says "…include everyone who is living or staying here for more than two months. Include yourself if you are living here for more than two months, including anyone else staying here who does not have anywhere else to stay, even if they are here for less than two months." So how are they defining who should be in that household is changing. Then the Computer-Assisted Personal Interview (CAPI) for the 2011 ACS, this in-person interview says, "I am going to be asking some questions about everyone who is living or staying at this address." Then it changes, asking, "Are any of these people living away now for more than two months, like a college student or someone living in the military? Is there anyone else staying here? Do any of these people have some other place where they usually stay?" What does "usually stay" mean? It could mean different things to different people.

The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) also has similar issues here. It asks, "What are the names of all persons living or staying here?", and there is no time frame given for that at all. The flash card that interviewers use says, "Start with the name of the person or one of the persons who owns or rents this home." What happens if that person doesn't live there? A later instruction says, "List all persons who are staying in the sample unit at the time of the interview." So here, there is a time of reference that is very inconsistent. In wave one, there is an instruction that says, "Now I'll make a list of all the people who live or stay here at this address." And then there's a verification, "You live and sleep here most of the time," so the concept of sleep is introduced as relevant. It also asks, "During a typical week over the last month or so, how many nights did this person stay here overnight or was there no usual pattern?" So there's not only stay here overnight, there's a usual pattern. You can see that there are several concepts that are undefined within the same question.

As we progress, we come to the point where the purpose of the person's visit becomes relevant at some points in this instrument, or in other words, why is this person there at the sample address? Some instructions provide detailed information about whether someone should or should not be included according to the rules and purpose of the visit. Although most of the surveys do not include this level of detail do not, some of them do sometimes include it. So in Census TQA you get, for example, "Do not include babies when the purpose they are at the address is for daycare or vacation." Other instructions say to include the person if they are "looking for a place to stay." So people who are looking for a place to stay, i.e. crashing on your couch.

The SIPP, wave one, does this, too. It asks, "Does this person usually live here but is away travelling for work, on vacation, or in the hospital?" And then sometimes, we miss people when it is not totally clear where they live or stay. For example, the SIPP asks, "Just to make sure, have I missed anyone else who is staying here until they find a place to live?" And as I mentioned earlier, some instructions are only seen by the interviewer. So for this question on the ACS CAPI the question is, "Are any of these people away now for more than two months, like a college student or someone living in the military?" There are additional applicable categories included within an instruction to the interviewer, "Do not select children in boarding school or summer camp" and can all only seen by the interviewer so. Another instruction that contains key information is "Select children and shared custody who are not currently staying at the sample address regardless of the length of stay." Here's one, "Do not select children and shared custody who are currently staying at the sample address regardless of where they usually live or stay." These are important pieces of information that the respondent never sees while completing the ACS.

Whether they're staying there or not, and I am especially interested in this pattern after seeing participants do this in the U.S. Census Bureau's Human Factors and Usability lab, respondents tend to count their children whether they are there on the day of the ACS or not. Some more research on this situation, and on custody situations in particular, would be beneficial to the survey methodology community.

Here are some more instructions that respondents do not see during an ACS CAPI interview, "Select commuter workers who stay at the sample address to be closer to work" and "Do not select commuter workers who stay in some residence somewhere else to be closer to work when their family residence is the sample address." These would be good to know, especially if you are one of these people in a commuter situation and the rules are pertinent to your response. These are not necessarily problems that are entirely negative, as complex residence rules present rich opportunities for research. Each and every one of these cases of semantic or referential ambiguity could be its own line of empirical research, like research with branching and perhaps some further work on rostering probes. A goal in the long term would be reducing the amount of understanding of the rules that respondents have to do, maybe by answering a series of simpler questions. Also, households can be complex. Research on living situations and response tendencies has been done, but definitely warrants further work.

More research is needed on self-response rostering and mobile device rostering both in the self-response and interviewer-administered modes. As the modes for survey data collection and for building household rosters are changing, we need to do stay current and continue looking at how these rosters are changing. Basically, it looks like no matter what set of rules you have, if you have the average household, you're probably going to get the same result to how you word these things. So the average household is for the Census 2010, the median age was 37.2, the average household size was 2.58 people, among the nations occupied housing units, 65.1 percent were earned compared with 34.9 that were rented. So that is your typical household, and if all of them were like that, determining who lives at each address would be easy and this chapter would not be necessary.

However, there are a lot of complex households where these rules are sometimes very difficult to apply and understand. So large households, extended families, people with tenuous attachment to the household, roommates, roomers and borders, and hired hands are some categories of residents who are included in some surveys. People in boarding school, college students, commuter workers, babies, children in shared custody, people temporarily living away from their primary residence, and sometimes concealment of family members for different reasons, fear of loss of welfare benefits, deportation, arrest, etc. Immigrant households and homelessness are also some challenges faced with respect to enumeration.

But now we have new enumeration issues coming out. Now we have what are known as "couch surfers" and they are very, very common on Craigslist.com. Just doing a preliminary search while putting this presentation together, I found three illustrative examples on Craigslist: "For $200, I am seeking a couch to sleep on for four weeks only." Another one read, "I need a couch to sleep on for three nights a week." Yet another stated, "Couch surfer, I need a couch to sleep on!" When it comes to residence rules, what do you do with these people, where are they, where do they usually stay? I don't know. I want to know.

How do you determine if something is someone's usual residence? Based on the large corpus of work from Eleanor Gerber, Roger Tourangeau, Lori Schwede, Jenny Childs, and the Living Situation Survey team, there is a kind of list of ideas that people associate with usual residence. "So when considering Joe, does he contribute money for rent, food, bills or anything else? How frequently does Joe sleep here? Does Joe use this address to receive mail or phone messages? Does Joe usually eat here? Do you consider Joe a member of the household? Does Joe have another place or places where he stays? Does Joe perform chores like cleaning? Does Joe have a say in household rules? Does Joe usually live or stay here?"

Now, you might think, "Well, this doesn't really apply to too many people," but some of these categories represent there a lot of people. The 1994 CPS results indicated that 18.4 million children were living with a single parent, and that roughly two-thirds of those parents were separated or divorced. So, there are a whole lot of children in shared custody. There are about 100,000 kids in boarding school, and although there probably aren't a lot of commuters second residences compared to the average household, we don't really know. So I ask the Journey to Work Branch at the U.S. Census Bureau a while ago where to find these data on commuter residences and they said that they don't collect it, but that it is a great idea for a research project. So, there's another future research area we could investigate.

Much of what we know about capturing data for these hard to reach cases comes from the research associated with conducting the Living Situation Survey, which was a survey that was done in 1993. There are 13 roster probes resulted with 999 households. It was done by a number of my colleagues at the U.S. Census Bureau, the Research Triangle Institute, and Roger Tourangeau. In looking at the results of this survey, Schwede and Ellis, in 1994, did a log linear analysis of the categories of that long list I mentioned of various roles that a person can play within a household as a way to evaluate household attachment. The best-fitting model for the data showed that the two attachments that best predict respondent's assessment of household membership were helping with chores, such as cleaning or watching children, and having a say in making rules.

In our current work in the Center for Survey Measurement at the Census Bureau, we're looking at extended roster probes. The idea here is to remove the burden of interpreting residence roles, giving a list of instances, when considering the residency status of an individual, in which the respondent should include him or her and when he or she should not be included. We just ask, for instance, "Was this person away in the military, was it for more than two months, was someone away in boarding school, was it more than two months, or was someone away at another address for work?" The respondent does not really need to know what we're getting at – from an institutional or research perspective – by asking if that person lives there or not and there is no need to put the onus on the respondent to have to think about the residence rules. Although a series of simpler rostering probes may chunk the questions into easier-to-understand concepts, this approach introduces a different challenge in that it introduces a whole lot more questions, which can make the survey take longer for the respondent to complete. We are still assessing how this might impact the response rate and the results of the survey.

In 2005, a book written by my Census Bureau colleagues, which was entitled, *Complex Ethnic Households in America*, was published. In 2012, we at the Census Bureau conducted the National Census Test with two different roster coverage paths, and one key manipulation being that one has more roster probes than the other. The results are still "To Be Announced" because the analysis is going on right now. Also, Patti Goerman and Jenny Childs are working on a fantastic project that includes alternative roster paths within NCS.

What remains to be done? A whole lot. A study equivalent to the Living Situation Survey has not been conducted recently or carried out using moderate technology. So, we need to determine whether the same categories of attachment are predictive of the subjective household membership or whether the categories are changing and evolving along with the population of the United States. Some questions to ponder are, "Do these couch surfers, which I defined earlier, have a say in making household rules?" "What is a typical couch surfer's typical role within a household? How would you define the attachment there?"

Another question related to concept of roster complexity is, "What is the best way to ensure that interviewers are reading the questions as worded?" So we can design the questions and make sure that they are as easy as possible for respondents to comprehend and answer. We can come up with a core set of roster questions, which we haven't done yet, but then we need to also make sure is it the best way that interviewers read these as worded. In other words, are the questions easy to read off of a piece of paper, a computer or a mobile device? Are they easy for the interviewer to pronounce and for him or her to clearly and consistently annunciate? When the questions are read aloud to respondents, do they actually make unambiguous sense? If so, we might see an improvement in the quality of survey data and an improvement in initial accuracy of household rosters.

The National Research Council (NRC) in 2006 suggested areas of research, such as whether Census respondents find a pure *de facto* rule or a traditional de jure rule easier to follow. *De facto* is a Latin term that translates to "concerning fact" and in a roster context, describes a rule whereby people would be counted wherever they usually sleep or where they physically were located at the time the roster was created. *De jure* means "concerning law" and describes a residence rule where people would be counted at their legal or primary residence as defined by that rule. Furthermore, is one type of rule or the other easier to follow with specific reference to large or complex households? In other words, is it easier to apply and follow rules of residence based on who is at a given household on a specific day or whether or not one legally resides at a given address? Further still, it should be determined whether

Census Bureau standards of the concept of where people “live or sleep most of the time” are consistent with the general population’s most widely held notions of usual residence. So, does what the Census Bureau thinks of usual residence map one-to-one with what most people think is a usual residence? This specific line of research has not been thoroughly conducted to date, so there is still a lot of room for investigation in this area.

Ultimately, we need to determine which basic residence rules make the most sense and are easiest to understand for respondents. Most of the research on these topics was conducted before the Census Bureau started the process of moving all of our surveys to a self-report Internet mode as well as incorporating the use of the most widely used types of mobile devices into our data collection procedures. We now have a variety of new technologies to help us try to address these issues, so we should definitely capitalize on the increased availability and methodological acceptance of technological tools to maintain a current and relevant understanding of the structure of the American household.

As I mentioned earlier, the main areas of research that should be conducted in the near future include these topics: Do different approaches to rostering yield different results in terms of accuracy, response rate, and overall data quality? There should be some research to empirically look at how these different approaches really do turn out. Also, does the application of different types of residence rules yield significantly different results in terms of which people get included in a roster, and what are the long- and short-term implications for the survey and its data if there is a difference? And if there isn’t a difference? Additional, we need to determine what set of questions should be included in an optimized set of core residence rules that both makes sense to respondents and leads to an accurate count? Once there is a consensus on what the core set of questions is composed of, researchers and data collection agencies would have a baseline from which they would be able to design more topic-specific questions for each type of survey or Census that needs to be conducted.

In summary, the goal of an expanded and continuous investigation of the most effective and accurate way to build a household roster involves answering a question that seems rather straightforward. What’s the best way to ask whether a place is someone’s usual residence to determine whether or not that person should be counted? However, as discussed in this talk, there are a myriad of factors to consider when designing such a question so that it is ideally constructed for both interviewers and respondents and leads to an accurately constructed list of residents for each household.

Addendum by Roger Tourangeau: I have three things I’ve done that are relevant to this topic that I would like to contribute to this chapter. The first was I was involved in this project for the Census Bureau to look at questions for the SIPP. We did an experiment where we looked at different approaches to the roster questions and the short summary of what we found were some problems with under-reporting of individuals that seemed like deliberate concealment. When we went to an anonymous roster, we found a lot more black males. As a caveat, it was a purposive sample of blocks that were thought to be likely to be problematic. But I think the results are still interesting with further follow-up.

The second thing I did, in part based on that and Elizabeth Martin's follow-up work to the Living Situation Survey, was a series of experiments. We did some studies where we tried different ways of explaining the residence rules and we did a series of experiments. It was myself and Fred Conrad and some students and it was published in the *Journal of Official Statistics* (JOS). What we found is people don't read the explanations and rules very carefully. In fact, I think we thought that there were two major obstacles to the approach that the Census form takes. First of all, people think they know what the concept of a usual residence is. They think they know who lives in their house and that they don't need a lot of instructions about how to answer the question. And so we found respondents doing very superficial reading of the instructions. We found it difficult, even when we altered the definition, in major ways to get respondents to change the number of people that are reported.

We also gave respondents a bunch of vignettes and the rules and definitions we gave them didn't seem to affect their answers to the vignettes, as to whether they should count this person or shouldn't count this person. This is one situation, and I think there are a lot of situations in surveys, and we described this problem. There is an official concept that the survey is trying to get at and then there is sort of a similar, but not quite completely identical, everyday version of the concept. So, like this rule about how you're supposed to count college students away from a home as not living there. There you are taking the survey, they're your children and you're paying their tuition, and you want to count them! It is very, very hard to overcome that conviction on the part of respondents. So, there are both cognitive and motivational reasons why it's difficult to persuade people to follow the rules.

The third thing I was involved in was when I was on the 2006 NRC panel and I thought that was a font of great ideas, not all of them mine, about how to do a better job. I do think it's really surprising. We looked at all the Census forms that are in English around the world. We looked at what they do in New Zealand, what they do in Ireland, what they do in the UK and so on. It's almost a 50/50 split, at least in the countries we looked at, between a de facto and a de jure rule. It just seems like the de jure rule is way easier to implement. So, you as the interviewer ask, "Who was here at midnight of last night?" This is a much clearer question than, "Who usually lives here" and runs into a lot fewer ideological and emotional issues.

Some of the forms do a two-phase thing where they start with the de facto question and then they'll ask, "Do any of these people normally live somewhere else?" And then they collect the address. Presumably, that leads to duplication, although I am not sure how far people go with this. It does permit the possibility of applying either rule. The final point that I want to describe about my experience on the NRC panel was it seemed to be a non-starter to a lot of people to have a reference date on the Census form other than the current day. It is complicated and it's hard to explain because if somebody moved into to an address on March 30th, the inclination most people have is you want to count them where they were on April 1st. Even though that wasn't their usual residence during that already-past period of time, where did they live most of the time during the past three months? This approach would again be a simpler when compared to the simplicity of the de facto rule.

It is a lot simpler and an approach to the simplicity of the de facto rule. But, people just feel like it's a non-starter. And you can't do that because of the care where the respondent just moved, and remember that the Census Bureau is allocating congressional seats. There is so much to be politically careful about concerning the Census residence rules that you can never reform really the Census rule because of congressional opposition and so on. It does seem like a reference period like they have in the ACS. Maybe a new and improved version would make the task more determined and easier for people.

Kathleen Targowski Ashenfelter earned her PhD in Quantitative Psychology, with a focus on Dynamical Systems Modeling, from the University of Notre Dame in 2007. She also holds a master's degree in psycholinguistics. During her tenure as the principal researcher for the Human Factors and Usability Research Group at the U.S. Census Bureau, her empirical research focused on identifying potential methods for improving the efficiency and accuracy of conducting residential enumeration, reporting the standard error associated with U.S. Census Bureau statistics, and enriching the user/respondent's experience with U.S. Census Bureau Web sites, data products, and surveys.

Kathy is currently a senior data scientist on the Cyber Analytic Services for Unisys Corporation. She applies advanced behavioral algorithms and dynamical systems models to large structured and unstructured datasets (from private industry as well as trusted government sources like the U.S. Census Bureau) for clients in both the private and the public sector.

52

Proxy Reporting

Curtiss Cobb

"*What kind of work do you do at your job, that is what is your occupation?*" Asking respondents about their occupations in survey research is a common practice; after all, occupation is a fundamentally important demographic characteristic that relates to numerous other attitudes and social phenomena. Among survey researchers, there are usually high expectations of accuracy when asking people to self-report their occupations. The question is very easy for most people to answer about themselves; one's occupation is generally an important part of his/her identity, a component of the self encountered on a near daily basis.

But what about asking survey respondents to report on the occupation of someone other than themselves? How well can respondents answer the question: "*What kind of work does he or she do at his or her job, that is what is his or her occupation?*" When a survey respondent is asked to report on information about someone else, it is known as proxy reporting and is a widespread practice in survey research. This chapter explores a history of research on how accurate are proxy reports relative to self-reports, and a study funded by the National Science Foundation that explores how the design of the survey questions contribute to the accuracy of proxy reports.

Proxy reporting is common in survey research and is used to collect a wide variety of information about other people in a faster and cheaper way with

C. Cobb (✉)
Menlo Park, CA, USA
e-mail: curtisscobb@gmail.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_52

higher cooperation rates. For example, almost all household surveys rely on a single household member to inform about everyone in the household rather than trying to track down and interview everyone in the household separately. Proxy reports account for more than 50% of interviews conducted by the U.S. Census Bureau, and the Current Population Survey is estimated to be 17% cheaper than it would be otherwise because it relies on household proxies to provide information about date of birth, military service, education, race/ethnicity, voting history, and many other topics for everyone living in the household (Boehm 1989). Thousands of demographic and social scientific studies have been published using datasets that make use of proxy reporting, often without taking that factor into account when conducting statistical analyses.

The common methodological assumption is that while selfreports are generally ideal, proxy reports can provide data that are as accurate or nearly as accurate as self-reports, particularly for demographic types of information. It is even believed there are times when proxy reports can be more accurate than self-reports, such as when a parent reports on a child, when health impairments prevent self-reports or the information being asked about might be stigmatizing to the target individual.

In practice, a substantial body of research has found that agreement between proxy reports and self-reports can vary substantially (Looker 1989; Menon et al. 1995; Moore 1988; Mosely & Wolinsky 1986). One laboratory test saw agreement rates on employment-related questions range from 67% for whether the individual worked overtime in the previous week to 92% for whether an individual is paid hourly or salary-wise (Boehm 1989). Disagreement between self-reports and proxy reports have been related to the type of information being asked, the importance of the information to the proxy, and the relationship between the proxy respondent and the target (Bickert et al. 1990; Kojetin & Jerstad 1997; Lee et al. 2004; Mathiowetz & Groves 1985; and Moore 1988).

However, a closer examination of research on proxy reports indicates that the design of many studies limit how informative their findings are for assessing proxy reporting accuracy. Specifically, few studies have had reliable direct assessments of phenomena of interest with which to measure the accuracy of survey reports—whether they are from proxies or targets. Rather, researchers usually assume that self-reports are accurate and have used them as benchmarks against which to compare proxy accuracy. It is possible that proxy reports might sometimes be more accurate than self-reports, and it is possible that self-reports are not especially accurate. As a result, Jeffrey Moore's observation in his 1988 meta-analysis still holds true–methodological shortcomings of much of the research means well-designed studies of proxy reporting are rare and the range of topics covered is limited. Until more data are gathered, researchers must remain tentative about the quality of proxy responses.

Assessing Accuracy

A large body of research spanning ninety-three (93) studies over more than sixty years has sought to compare the accuracy of proxy reports and self-reports. At first glance, many of these past studies appear to be useful for documenting the conditions and ways in which targets and proxies vary in reporting accuracy. However, upon close inspection, the designs of many of these studies make it difficult to draw any generalizable conclusions from them with confidence.

To evaluate these 93 studies, it is useful to begin by acknowledging the features of a study that would allow it to be informative about the accuracyof proxy reports and target self-reports:

(1) Target and proxy respondents should constitute representative samples of the same population. (*25 studies failed*)
(2) Both target respondents and proxy respondents should be interviewed. (*17 studies failed*)
(3) The questions asked of the targets and proxies should be identical, and they should be asked in identical contexts; that is, the number, content, sequence, and formats of the questions being asked should all be the same, and they should be administered in the same mode. (*21 studies failed*)
(4) An independent, external measure of the attribute being assessed should be used to assess the accuracy of the target and proxy reports. (*76 studies failed*)

Using the above approach, two research designs can be implemented. To assess accuracy at the aggregate level, target respondents can be randomly assigned to be informed about by himself/herself or by a proxy respondent with the aggregate response distributions compared to the ground-truth external measures. Alternatively, accuracy at the individual level can be assessed by collecting information from both the target and his/her proxy and compared to ground-truth external measures. An additional strength of the individual-level analysis is the ability to assess the relationship between reporting errors made by both targets and proxies.

Only 6 out of the 93 studies examined have the necessary features to evaluate proxy response accuracy, and they all involve reports of medical events. Cobb et al. (1956) found that self-reports of arthritis or rheumatism diagnoses were accurate for 60% of target respondents, while the accuracy of proxies within the same household was only slightly lesser at 56%. Thompson and Tauber (1957) found that both self-reports and proxy reports from a "responsible adult of the same household" were equally accurate 73% of the time at reporting diagnoses of heart disease. One study found that targets and same-household adult proxies were similarly accurate at reporting number of physician visits (55% target vs. 54% proxy). Balamuth (1965) reported similar results for hospitalizations when proxies were spouses (88% targets vs. 87% proxies). Taken together, these four studies suggest that reporting accuracy can vary considerably from study to study (from a low of 54%, which is about equal to chance, to a high of 88%), but also that target reporting and proxy reporting were consistently nearly equally accurate.

Two of the six studies that met the criteria for assessing proxy accuracy yielded different results–one where proxies were substantially less accurate than targets and one where targets were less accurate than proxies. Magaziner et al. (1997) compared proxy reports to self-reports and direct observations made by medical professionals on targets' ability to perform daily physical and instrumental activities during recovery from a broken hip. Each proxy was identified as the person most knowledgeable about the target's health and general abilities. Self-reports yielded ratings of abilties that were significantly different from direct observations for five of thirteen activities. Proxy reports performed substantially worse; they were different from direct observations on ten of the thirteen activities. Moreover, proxy reports were different by a larger magnitude in each activity from direct observations than self-reports.

Cannell and Fowler (1963) reported that parents were more accurate in reporting the number of times their minor children visited a doctor than were the children themselves. Of course, this finding seems attributable to the young age of the targets; however, it still demonstrates there are situations when proxies may be more accurate than targets.

While the other 87 studies reviewed on proxy reporting do not meet the necessary conditions to evaluate reporting accuracy, they are informative of the social and cognitive processes that proxies engage in to report information on others and may provide clues on the scenarios when proxy reports can be relied upon to be accurate. For example:

- Proxies and targets likely use different estimation and recall strategies to report on information. Proxies are more likely to rely on estimates and anchor their answers to their own behaviors and attitudes when information is missing. Targets are more likely to use recall to retrieve information necessary to deliver a report.
- The closer a proxy is to a target and the more time spent together increases agreement. Time together increases the proxy's awareness of situations in the target's life and also increases the similarity in the cognitive strategies used by the target and proxy to answer questions (Bickart et al. 1990; Mathiowetz & Groves 1985; Moore 1988).
- Knowledge of and exposure to the question topic increases agreement between proxy and target. For example, when children reach an age where they think about earning income themselves and still live at home, they are more accurate at reporting parental income than before they are familiar with money or after they leave the home (Amato & Ochiltree 1987).
- Proxy reports of stable traits and characteristics are more likely to agree with self-reports (Kojetin & Jerstad 1997; Lee et al. 2004).

Proxy Accuracy, Education and Question Format

A study conducted by Cobb, Krosnick and Pearson (forthcoming) looked outside the field of medicine to explore the accuracy of proxy reporting about targets' field of study for the bachelors degree. Alumni from Stanford University and their parents were interviewed using a web survey that asked about field of study using both an open-ended and closed-ended question. Respondents were randomly assigned whether to receive a 10-category or 7-category closed-ended question to answer. Answers were evaluated against ground–truth information obtained from official Stanford University records.

Background: The Stanford study was funded by the National Science Foundation to inform the design of a new question to be placed on the American Community Survey (ACS) conducted by the U.S. Census Bureau. The intent of the question was to be used to help with sampling for research–targeted individuals with educational backgrounds in STEM fields (science, technology, engineering and mathematics). The ACS is a household survey that relies on proxy responses to collect information on everyone in the household from a single informant.

Question 1: Are proxies similary accurate at reporting the field of bachelors study of targets?

It may seem obvious that self-reports of fields of degrees will be accurate and that they will be more accurate than proxy reports. For targets, selecting and completing a major is an important milestone that was re-enforced over four years of study. Proxies are further from the experience and may rely on current information to reason what was the college major, such as current area of employment. But this assumption may be incorrect for a number of reasons. Targets may seek to clarify uncommon, unconventional or interdisciplinary majors. There may be social desirability to expand credentials similar to what influences overstating qualifications on a resume.

Question 2: Are proxies and targets more accurate at reporting fields of bachelors study using an open-ended or closed-ended question?

There is a mixed history of research comparing the accuracy of open-ended vs. closed-ended questions that use external ground-truth benchmarks for assessment, even among self-reports. Some prior research found no difference when questions were about facts, whereas Whipple (1909) found that 92% of responses to open ends and 75% of responses to closed-ended questions were accurate. Burton and Blair (1991), Cady (1924), and Burtt (1931) all reported similar findings suggesting greater accuracy for open-ended questions. In contrast, Dent (1991) found that adults' answers to closed questions were more accurate than open questions, and Hooper (1971) found that children were also more accurate when answering closed questions. For proxies in particular, an open-ended question may yield more accurate results because it avoids mapping answers onto categories or a closed-ended question may be more accurate because it jogs the memory or does not rely on such a precise answer.

Question 3: Do more categories or fewer categories yield more accurate answers for closed-ended questions from proxies and targets?

We also examined two different closed-ended questions, one offering seven response options and the other offering ten. Using fewer categories may increase difficulty because categories are overly broad and vague. A longer list of more specific fields of study might make it easier for a respondent–target or proxy–to find a category that is a close match. On the other hand, an overly specific list may require the respondent to understand fine distinctions in categories which may make the task more difficult. An overly specified list may also lead to more over-or underclassification.

Lastly, the study design allowed us to investigate whether reports of information become more or less accurate with time.

Methods: A web survey invitation was sent to a random sample of Stanford University alumni who graduated in 1970 or later, as well as to one or both of their parents via email. A total of 4,021 alumni and 3,653 parents were invited to complete the questionnaire, with 2,013 interviews completed by alumni (50% response rate) and 1,893 completed by parents (47% response rate). Respondents were instructed to answer the survey by themselves and not to ask for help from anyone else or to look up the information being requested from another source.

Alumni and parents were asked both a closed-ended question with either seven or ten response options and followed by an open-ended question. Theclosed-ended question for alumni asked:

> "*Thinking only about the bachelor's degree(s) you received from Stanford, which of the following broad categories best represents the field(s) of your major(s)? Did you get a bachelor's degree in…*"

The closed-ended question for parents was similar but referencing the name of their alumnus child. Respondents were instructed to provide an affirmative or negative response (yes/no) for each category offered as a response option. The 7-category response options includes: biological, agricultural, physical or related sciences; health, nursing, or medical fields; engineering, computer, mathematical or related sciences; psychology, economics or other social sciences except history; history, arts or humanities; business, communication or education; and some other fields. The 10-category response options were: biological, agricultural or related sciences; health, nursing or medical fields; computer, mathematical or related sciences; engineering; physical or related sciences including earth sciences; psychology, economics or other social sciences except history; business or communication; education; history, arts or humanities; and some other fields.

The open-ended question for alumni was:

> "*In what specific field(s) did you receive a bachelor's degree?*"

Parents received a similar question referencing the alumni child's name. Data was collected between March 9, 2008 and April 6, 2008.

Results. Alumni were accurate at reporting their bachelor's field of study using the open-ended question 96.7% of the time. An additional 2.8% were partially correct by either mentioning one or more additional incorrect fields or omitting one or more fields. Only 0.5% of alumni were completely wrong. Parents were also impressively accurate, but their accuracy was significantly lower than alumni: 88.7% fully correct, 4.7% partially correct and 6.6% completely incorrect ($\chi 2(3) = 120.54$, $p<0.001$). Parents whose children did not complete the questionnaire were no less accurate than parents whose children did complete the questionnaire.

Proxy accuracy was related to alumni accuracy. Parents were completely correct 90.1% of the time when their children were also completely correct, while parents were correct 65.5% of the time when children were partially incorrect, and incorrect 100% of the time when children were completelyincorrect ($\chi 2(1) = 17.91$, $p<0.001$).

Accuracy was substantially lower in response to the closed-ended questions compared to the open-ended questions. Alumni answered completely correctly only 64% of the time and parents were correct only 60.3% of the time. Answers to the 7-category closed-ended question were about 10 percentage points more accurate than answers to the 10-category question for both alumni and parents (69.8% vs 58.3%, $\chi 2(3) = 31.72$, $p<0.001$; 64.8% vs 56.2%, $\chi 2(3) = 16.22$, $p<0.001$).

When errors occurred, alumni were more likely to give partially correct responses and parents were more likely to be completely wrong. The type of partial errors was also different between the two groups. When alumni were partially wrong, they over-reported majors, while parents under-reported majors when they were partially wrong. As with open ends, there is a strong relationship between alumni being able to accurately report and parents being able to do so.

Surprisingly, time since graduation had a positive relationship with accuracy for alumni. Those whose graduate was more recent were also more likely to provide incorrect information–often over-reporting majors. There was no relationship between the year of graduation for alumni and parents' reporting accuracy. This contradicts the notion that longer recall intervals always yield less accurate reporting.

Parents with advanced degrees were also more accurate than parents without an advanced degree–whether this is do to with cognitive ability orfamiliarity of majors/higher education systems is unclear.

Conclusion

Given how common proxy responses are within survey research, it is fundamentally important that survey researchers understand how accurate proxies may be at informing on others. Previous studies and the education study presented here suggest that proxies responses are nearly as accurate as self-reports, and often contain the same errors as self-reports. This generally corroborates Magaziner et al.'s (1997) conjecture that agreement between proxy reports and target reports is greater than agreement between proxy reports and external records of the same phenomena.

Furthermore, accuracy is contingent on more than just the topic of the information being collected or the proxy's relationship to the target; it also depends on the form of the question being asked. The difference in accuracy of approximately 30 percentage points between open vs. closed questions is considerably large.

Lastly, the dearth of research on the accuracy of proxy reports relative to their importance and use in the field of survey research presents an opportunity for the discipline to build in this area of knowledge. There are not yet enough studies to understand proxy reporting across so many domains. When new questions are being developed and tested, proxy respondents should be included in the process.

References and Further Reading

Amato, P. R., & Ochiltree, G. (1987). Interviewing Children about Their Families: A Note on Data Quality. *Journal of Marriage and the Family*, 49(3), 669–675.

Balamuth, E. (1965). Health Interview Responses Compared with Medical Records.

Bickart, B. A., Blair, J., Menon, G., & Sudman, S. (1990). Cognitive Aspects of Proxy Reporting of Behavior. *NA-Advances in Consumer Research*, *17*. 198-206.

Boehm, L. M. (1989). Reliability of Proxy Response in the Current Population Survey. In Proceedings of the Survey Research Methods Section, American Statistical Association.

Burton, S., Blair, E. (1991). Task Conditions, Response Formulation Processes and Response Accuracy for Behavioral Frequency Questions in Surveys. *Public Opinion Quarterly*, *55*, 50–79.

Burtt, H. (1931). *Legal Psychology*. New York: Prentice Hall.

Cady, H. M. (1924). On the Psychology of Testimony. *American Journal of Psychology*, *35*, 110–112.

Cannell, C. F., & Fowler, F. J. (1963). Comparison of a Self-Enumerative Procedure and a Personal Interview: A Validity Study. *Public Opinion Quarterly*, *27*(2), 250–264.

Cobb, C., Krosnick, J. A., & Pearson, J. (forthcoming). The Accuracy of Self-Reports and Proxy Reports in Surveys.

Cobb, S., Thompson, D. J., Rosenbaum, J., Warren, J. E., & Merchant, W. R. (1956). On the Measurement of Prevalence of Arthritis and

Rheumatism from Interview Data. *Journal of Chronic Diseases*, *3*(2), 134–139.

Dent, H. R. (1991). Experimental Studies of Interviewing Child Witnesses. In J. Doris (Ed.), *The Suggestibility of Children's Recollections*. Washington, D.C.: American Psychological Association.

Hooper, S. R. (1971). Communicative Development and Children's Responses to Questions. *Speech Monographs*, *38*, 1–9.

Kojetin, B., Jerstad, S. (1997). The Quality of Proxy Reports on the Consumer Expenditure Survey.

Lee, S., Mathoiwetz, N., Tourangeau, R. (2004). Perceptions of Disability: The Effect of Self- and Proxy Response. *Journal of Official Statistics*, *20*(4), 671–686.

Looker, E. D. (1989). Accuracy of Proxy Reports of Parental Status Characteristics. *Sociology of Education*, *62*, 257–276.

Magaziner, J., Zimmerman, S. I., Gruber-Baldini, A. L., Hebel, J. R., & Fox, K. M. (1997). Proxy Reporting in Five Areas of Functional Status

Comparison with Self-Reports and Observations of Performance. *American Journal of Epidemiology*, *146*(5), 418–428.

Mathiowetz, N., Groves, R. (1985). The Effects of Respondent Rules on Health Survey Reports. *American Journal of Public Health*, *75*, 633–639.

Menon, G., Bickart, B., Sudman, S., & Blair, J. (1995). How Well Do You Know Your Partner? Strategies for Formulating Proxy-Reports and Their Effects on Convergence to Self-Reports. *Journal of Marketing Research*, *32*, 75–84.

Moore, J. C. (1988). Miscellanea, Self/Proxy Response Status and Survey Response Quality, A Review of the Literature. *Journal of Official Statistics*, *4*(2),155.

Mosely, R. R., Wolinsky, F. D., (1986). The Use of Proxies in Health Surveys: Substantive and Policy Implications. *Medical Care*, *24*(6), 496–510.

Thompson, D. J., & Tauber, J. (1957). Household Survey, Individual Interview, and Clinical Examination to Determine Prevalence of Heart

Disease. *American Journal of Public Health and the Nations Health*, *47*(9), 1131–1140.

Curtiss Cobb leads the Population and Survey Sciences Team at Facebook. His research focuses on cross-cultural survey methods, web surveys, technology adoption patterns and evolving attitudinal trends related to people's online "presence". Prior to Facebook, Curtiss was Senior Director of Survey Methodology at GfK and consulted on survey studies for clients such as the Associated Press, Pew Research Center, CDC, U.S. State Department and numerous academic studies. Curtiss received his BA from the University of Southern California and has an MA in Quantitative Methods for Social Sciences from Columbia University. He holds an MA and PhD in Sociology from Stanford University.

53

Questionnaire Design

Jon A. Krosnick

We know a lot about best practices with regard to questionnaire design but there is also a lot we don't know and lots of future research that is needed. Many of the other contributors to this volume are experts in all of these issues and those who know this literature in detail know that everything I will include in this chapter has controversy behind it. There are a limited number of studies on each of these points, we need more, but we need to design questionnaires today and tomorrow and the next day before those studies are done. So taking a look at the literature that exists allows us to reach some conclusions about good ways of designing questionnaires as best we can based on current knowledge, but we need more work in lots of ways.

One of the ways I know that we need more work, especially with the help of NSF is this question, which I found on the Xerox machine recycle bin at my office at Ohio State some years ago. It began with a statement:

"I am calm and relaxed"
"How accurate is this statement as a description of you?"

0	1	2	3	4	5	6	7	8	9	10
Not at all accurate									Extremely accurate	Don't know

J.A. Krosnick (✉)
Department of Communication, Stanford University, CA, USA
e-mail: krosnick@stanford.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_53

As I am standing there waiting at the Xerox I saw that this was in somebody's questionnaire. When I saw this, I realized how much help the field needs. At least according to my reading of the literature, just about everything that can be suboptimal about this question is suboptimal about this question, the length of the rating scale, the approach to labeling the points, the general approach here, the use of words. And the truth of the matter is, how would this person who is in my office building know that these were bad decisions, where do you look to see that? We haven't had a single consolidated resource that provides the kind of guide I think we need.

In this chapter I am just going to run through, very briefly, the high points of some of what I take the literature to suggest about a series of practical decisions. We'll start with a general set of introductory remarks. I'll cover open versus closed questions for measuring some phenomenon. I'll cover the design of rating scales, how many points to put on the scale and how to label them. I'll cover acquiescence response bias, the impact of the order of answer choices, whether to offer a don't know option or not, briefly about question order effects, attitude recall questions, asking why questions and I am going to end with a longer discussion of question wording. I have come to the view that question wording is where the crises lies at the moment, that's one area where we severely need more work.

Just by way of introduction, I bring to this literature three goals in evaluating questions and trying to identify the best ways to ask questions. The first is, all other things equal, I would like to minimize administration difficulty. That is, I would like to ask a question that can asked and answered as quickly as possible and I would like respondents to make the fewest completion errors possible. So if we ask them to pick a point on a rating scale on a paper questionnaire we don't want them circling two or three points saying, "I am somewhere in this range, but I don't know where." And lastly all other things equal, we would like respondents to say they enjoyed answering the question and they weren't very frustrated by it. But, all else is not equal. When push comes to shove, I am happy to compromise all of these things, I am willing to go longer and have respondents be frustrated if I can maximize the reliability and validity of the measurements.

In most cases, we have not seen that kind of tension, the literature seems to suggest that what goes quickly and easy for a respondent also produces the most accurate data. The goal I would say underlining all the work I am going to tell you is when people answer a question, let's say we're trying to measure an attitude, ideally we would like the answer to be driven exclusively by the attitude itself. But we know from this literature that there's contamination from at least three other sources. Sometimes other constructs from the attitude we care about influence the judgment. If we ask, for example, do you approve or disapprove of the Affordable Care Act passed in 2010 that President Obama proposed? By mentioning him, we get evaluations of him contaminated in the answer in addition to evaluations of the Act. We would like as little method bias as possible. For example, a question could say, "you do like the Affordable Care Act, don't you?" We would like as little of that sort of leading as possible in questions and we would like as little random measurement as possible. So we are trying to minimize these three and to maximize the degree to whatever point people pick on, let's say a rating scale is a function of the attitude itself and distorted less by these other factors. The two theoretical perspectives that I have found helped me most in understanding when question wording matters, question structure matters and how are satisfying and conversational norms and conventions.

There is a set of ideas that Charlie Cannell and Roger Tourangeau and others are known for having brought into the literature. What they said is that if you think about the optimal way that a respondent would answer a question in order to provide an accurate answer, one of the things that he or she has to do is understand the intent of the question. What I mean here is if I ask "do you know what time it is?" You know the right answer is not yes. You know my intent is to actually find out the time. So the respondents will naturally see past the wording of a question to what they presume the researcher's intent is. And once they understand the intent, they need to search memory for information, integrate whatever information they come up with into a summary judgment, and then express that summary judgment somehow into answering the question.

If respondents do all of this, that's optimal, but our work suggests that often people don't and then instead they satisfice, they settle for shortcuts and there are two ways to do it. One is to do the two middle stages, searching memory and integrating information, superficially rather than effortfully. So rather than searching for all information I can settle for the first piece of information that comes to mind rather than integrating in a fully balanced way I could integrate with some bias. That is what we call weak satisficing. Or, if I have completely given up on this thing, I am completing the questionnaire but I don't want to think anymore. I will perhaps understand the question, but then skip all memory search and integration, just answer.

Now, when we first proposed these ideas back in the 1980s, I was a grad student at the University of Michigan Institute for Social Research and the idea that survey respondents would do any of this was unimaginable to me. It was just not a part of the way our field thought about this process very prominently. But, over time, I've seen more and more evidence to suggest that respondents do actually sometimes do this and in the process they're looking for an apparently plausible answer that would be easy to justify without thinking to get the interview over with. In order to understand and master this, we've proposed that it might happen as a result of three classes of factors, people who find it difficult to think might be inclined to shortcut. People who might be motivated to think might be inclined to shortcut and when we give people a difficult task, they might be inclined to shortcut. That's the satisficing perspective.

The conversational norms and conventions perspective says the following, that a questionnaire is of course a script for a conversation and yet respondents don't realize that the rules of this conversation are different from the rules of normal conversation, particularly if there is an interviewing sitting in their living room, they think this is a normal human speaking normal English with them in a normal conversation. And, respondents assume that the same rules apply so they're assuming speakers usually follow the rules and that listeners assume speakers are following the rules. And yet, if questions violate those rules, respondents can be mislead or confused because they misinterpret and this is the real reality of it, that our questions routinely violate the rules of everyday conversation.

And interestingly, the more effort the respondent devotes to thinking about the questions, the more mislead or confused it appears that he or she will be. These are just a few of the violations that are rampant in the work that we do. In a normal conversation, if I said to you, "How are you doing today?" You said, "Good". I wouldn't say, "How are you doing today?" after that. After having asked that once, I wouldn't ask it again. And yet in surveys, we routinely use multiple questions to measure the same construct, self-esteem, battery of 15 or more items.

So, what would a respondent do in a situation like that? They say if they ask you 15 questions that are all measuring the same thing, they would say okay, why would you do that? But, if we don't tell them that, it will be natural for them to think in everyday conversation no normal person would do that so these must not be 15 questions asking the same thing, they must be 15 questions designed to ask different things. So respondents would struggle to find ways to interpret them so as to produce those gaps.

Another one is, all information provided is relevant and necessary in a conversation. If I am going to tell you some information and ask you to make a judgment in everyday conversation I would only tell you information I think is relevant. In a survey, we could describe a DVD player that consumers might consider buying, and we the researchers want to provide lots of information and say you can use some, you cannot use some, we don't have an opinion, we're just giving you some information you might want to use. We don't typically explain that so people might feel pressed to use all of the information we provide, on and on.

There are a series of ways in which we routinely violate conversational conventions and those ideas have helped us to understand I think why different question formats and wordings make a difference in how to optimize. I am going to walk you through some various areas of research and indicate what I think are insights into best practices beginning with open versus closed questions. In particular, we're going to focus on two different types of questions. So this is what I will call a categorical, you might call it a nominal question, what do you think is the most important problem facing this country today? We talked a little bit about that yesterday. That question has an unbounded universe of possible answers, right? People can answer, however, they wish. And you can imagine asking it in this open-ended form or in a close-ended form as some survey organizations do.

Another second category is numeric questions. For example, "over the course of your life, for how many years would you say you've smoked cigarettes"? So this is bounded at zero, but it is unbounded on the high end. You can imagine asking this in an open-ended way like this or you can imagine offering to help respondents by simplifying the task and offering a series of ranges instead, zero to five, six to ten, and so on. So, you say to respondents, we don't need an exact number, don't agonize, just give us something in a range.

The use of open-ended questions has declined over time, Howard Schuman did an analysis where he was looking at the prevalence of open-ended questions in a series of what you might think of as commercial surveys and academic surveys between the 1930s and the 1980s. There is a decline in the use of open-ended questions over this time period. In some sense you might think of this as natural selection that maybe this is the survival of the fittest, in fact this is the verdict coming through that open questions are not worth the trouble. In fact, I don't think that's what the literature supports for those two types of questions.

First of all, it is true that in experimental comparisons, open-ended questions take, on average, about twice as long to answer as closed questions and respondents prefer closed questions. On practical grounds, it might appear that closed questions are preferable. In studies of reliability, open questions prove to be more reliable than closed questions and in lots of different studies of validity, open questions prove to be superior to close questions across the board using these various different methods of assessing validity.

This is the remarkable thing I think about this literature and there are actually behind it a series of insights about why. A series of studies have looked at potential problems, limitations with open questions, maybe why things would go wrong. One concern was articulation ability. Maybe there are some people who just aren't so good at talking. They can pick a choice, but actually expressing their point of view verbally spontaneously might be difficult, that appears to have no empirical support at all.

A second concern was that open questions might be particularly susceptible to salience effects. If I ask you what's the most important problem facing the country and you happened to have seen a news story about crime on television last night maybe that enhances the likelihood that you would retrieve crime as a potential problem to answer with whereas, with a closed question on a list, maybe that salience effect may be minimized. That also appears to have no support. In fact, salience appeared to affect open and closed questions about equally.

Lastly, there is concern about frame of reference effects. That is if I ask you what's the most important problem facing the country, you need to understand what counts as a problem, what is an acceptable answer? That can be ambiguous with open questions whereas with closed questions if you offer a set of choices, what is an acceptable choice is made explicit. There is evidence that in some cases, the open-ended question stem is ambiguous enough that the frame of reference is not established. I would say that's not necessarily an inherent problem with open questions, but it is an inherent problem with some open questions that might better be solved in other ways.

With regard to closed questions, it turns out that a series of concerns have been articulated, all of which do have empirical support. One support is non-attitudes, that by offering people options, people select choices without actually having substance behind them. Second, with regard to the numeric options, if I am offering ranges of zero to five hours, six to ten hours, and so on, the way I choose those ranges sends a signal to a respondent about what an acceptable and normal answer would be. The answers in the middle of the range are what people assume to be what a normal person would pick so there is some gravitation toward the middle. As a result, better not to offer them.

Lastly, there is the idea that if I say to you, what's the most important problem facing the country, is it the federal budget deficit, crime, inflation, unemployment, or something else, it's the other option. That will cover us. In an unbound universe of potential problems, as long as I put the other option there, then that will solve the incomplete nature of the list. That turns out to be a serious problem. In fact, it appears from studies dating back 70 years or more that offering that other option does almost nothing. People almost never select it and they think what you're asking instead of saying what's the most important problem facing the country, they're asking which of the following is the most important problem facing the country? If you insist on picking something else you can, but we prefer that you don't so that seems not to work.

There's one last argument I think in support of this perspective, is that closed questions make respondents do more work than open questions. The first thing is if you say to them, what is the most important problem facing the country? They have to first answer the open question in their own mind, right? They have to say "well, I think its unemployment." Then they have to look at the list and see which answer choice maps on to unemployment. Similarly, if you say "how many years did you smoke cigarettes?" The first thing they've got to do is answer that open question and then choose an answer choice to express it. It's not that the close question is somehow simpler, it's actually more work for respondents. It's less work and more likely to be accurate if we ask respondents to answer the relevant open ended question and it looks like they produce more reliable and valid results than having the respondents do our work.

Let me just illustrate for you. The NSF asked us to evaluate a question for them, a question that Curtis Cobb also covered in his chapters in this volume. The question asks people what field their Bachelor's degree is in. There are seven choices of the closed question. Just looking at the list makes your head hurt. If I say to you, what did you major in and you say math, that's an easy judgment to make. Now you have to look at this list and find a place to go. Let's see, related sciences, math is related, no, let me keep going. Health, no, not that, engineering and computing, oh, here's math, great. So I've finally found me. You see how there's a bunch of work that gets done and there are words like or related sciences where you have to make a judgment, what is a related science?

The conclusion that I take this literature to suggest is that we should ask open-ended questions when you can't be sure of the universe of possible answer to a categorical question and the other specify option does not work. The only way to be sure that we know the universe is to pre-test, ask open questions from the population we care about, build the big list, offer it to people and we could do that, but it's so much work we might as well as the open-ended question in the real survey. And lastly, if we're looking for a number we should just ask for the number. That is what I take that literature to suggest and I don't know if that's really widely recognized and followed.

With regard to the number of points on a rating scale, there is lots of variation in surveys, including in the American National Elections Studies (ANES), everything from a two-point scale, do you approve or disapprove of the president's job performance, up to 101-point scales that take lots of texts to explain a feeling thermometer to people where they're providing ratings on a scale from zero to 100. Theoretically, there are a variety of principles you might imagine bringing to bear here to try to guess what the best length of a rating scale be. In order to understand as much as we can about respondents and to make the process of mapping your feelings on to a ratings scale easier, maybe more points is better. However, if you offer too many points on a rating scale, you can imagine respondents might get confused. What's the difference between 75 and 79 for example? And, we were concerned when we started our work on satisficing that offering a middle alternative on a ratings scale might be an invitation to satisfice, like, dislike, neither like nor dislike. Maybe that's a way to grab an option without thinking much that would undermine the process. On the other hand, maybe people need that option in order to accurately report neutrality.

We could make a prediction based on theory alone, increasing precision of ratings, validity and reliability as lengthen a scale to some length and after that length we've gained information while some people understand the scale, but as it gets longer and longer, the scale becomes ambiguous, is seeking more refined opinions than people actually have to offer and we end up getting less and less data quality.

In fact, the evidence I think is quite clear when I look at it at the moment that there are more completion errors for longer scales. The longer the scales get, the more people say I am in a range, but I can't tell you where exactly in this range I am, which signals that long scales have gotten too long. The longer the scale is, the longer it takes people to place themselves on the scale so it's a potential waste of time. And, when we look at reliability and four different indicators of validity, it looks like bipolar scales from like to dislike with neither like nor dislike in the middle might produce the most reliable and valid answers when they're seven points long and that unipolar scales with a zero point at the end, let's say from not at all important to extremely important are best with five points on the rating scale.

My favorite studies in this region are studies of what I call natural discrimination where you offer people a line and you ask them to put an X on the line to indicate wherever they are on this dimension without offering any number of points. Then what you can do is divide the line up into different numbers of segments to see which divisions produce the most reliable and predictive measurements. This has been done for bipolar scales and it reinforces the conclusion that seven points are good.

The evidence in terms of response speed as rating scales get longer, from 2 to 11 points, for example, indicates that the speed goes up. Although, it's sort of interesting that five-point scales take significantly less time than their neighbors do, that's a hint that it is preferable to have mid-points. This is even more direct evidence, when people are asked how difficult it is to use the scale, you can see that three points, seven points, and nine points, these are significantly less difficult than the scales of other lengths. So these are reinforcing that notion that in this case seven points is the optimal way to measure bipolar scales and that not only is it less difficult for people but that it produces more valid and reliable results.

It turns out my concerns about mid-points were unfounded. Offering mid-points is a good thing, the majority of people who select them belong there because they're neutral and they are not doing so to satisfice. The last thing in this arena that we have found is that it's helpful to branch bipolar dimensions. This is one of the first branching questions that I paid attention to from the ANES: "generally speaking do you consider yourself to be a Republican, a Democrat, an independent or what?" and Republicans and Democrats are asked if they're strong or not very strong. The independents are asked if they lean one way or another. You can produce seven-point scale like this. If you do it in those two steps up here it goes more quickly and it produces more valid and reliable results rather than presenting all seven of these, which people have to slog through and place themselves on.

There are two ways to branch. One is if you have a precise mid-point, so for example if you say, should defense spending be increased, decreased, or kept the same? If the mid-point is precise, it turns out that asking people, do you lean one way or another actually adds noise. In order to produce the seven-point scale, the people who place themselves at a precise mid-point like this belong there. We shouldn't branch the mid-point, instead we should branch the end points into three categories. So the people who say increase should be asked, do you want it increased a little, a moderate amount, or a great deal? But, there are other questions that offer a fuzzy mid-point, keep it about the same. And in cases like that, the fuzzy mid-point grabs some people who are truly off that mid-point. So, branching those mid-point people whether they lean toward more or less is a good way of splitting them up and the end point people should be branched into two categories instead.

With regard to verbal labels of scale points, you can imagine presenting let's say a five-point scale with numbers on all points and words only on the ends or we could put words and numbers on all of the points or we could get rid of the numbers and just have words on each of the points. As you think about selecting those labels, you might imagine a series of goals are worth pursuing. One is that we would like respondents to find it easy to interpret the meaning of all the scale points, we don't want them struggling. After they're done interpreting them, we would like them to say that the meanings are clear. That is, we don't want them to go through a process whereby they say, okay I've done the work I need to interpret and I'm still confused about what it means.

Third, we want all respondents to interpret the meanings of the scale points identically. That is, we don't want different people interpreting the scale points differently from each other. Fourth, we would like the labels to differentiate respondents as much as possible and as validity as possible. And lastly, we would like the resulting scale to include points that correspond to all points on the underlying continuum. You wouldn't want to make the assumption, let's say that of course people like pizza so we would say, how much do you like pizza, a great deal or a moderate amount? And leave out parts of that dimension. That's all fine in principle; the question is what do the data show in practice? What we have seen is this, numbers alone seem intentionally ambiguous and longer scales seem potentially more ambiguous, sorry actually still in the theoretical part of this rather than the data part. That there has been concern in the literature that labeling only the end points might attract people to those end points if the labels clarify the meanings of those points more so than other points. But, if you pick vague labels you might cause problems and if you pick labels that are overly specific, maybe people can't find the place on the scale where they belong. So some optimal degree of vagueness might be desirable.

In terms of respondent effort, maybe labels are a pain because they create more work for respondents to interpret. On the other hand, because labeled scales require reading and interpreting they require literacy as well. So you can see why there might be advantages and disadvantages either way.

Lastly, some people say it's difficult to administer numbered scales over the phone, words are better because people get confused about the numbers. The recommendations that follow from this literature need to be separated into two parts, dimensions that have no natural metric like liking importance and dimensions that do have a natural metric. The literature I think is quite clear, in terms of evaluating the quality of data, respondents like scales with more verbal labels, reliability is higher for scales with more verbal labels and validity is higher in various ways for scales with more verbal labels. It looks to me as if when people see a ratings scale without verbal labels on all the points, the first thing they have to do is figure out what the meanings of those points are with words. So, if we do that work for them, it makes their job easier and more reliable and more valid.

Furthermore, more widespread end points so instead of like and dislike, like a great deal and dislike a great deal, cover the dimension more fully and respondents appear to presume that we mean for ratings scales to have equal spacing among the points. You no doubt heard from people talking about customer satisfaction surveys, oh, everybody likes airlines all the time. So we wouldn't want to have a ratings scale that is balanced all across the dimension. What we need to say is did you love it a great deal, did you love it a moderate amount, did you like it or dislike it? Three or four positive answer choices and one negative would be sufficient. In fact, respondents presume that we mean for the points to be equally spaced across the continuum so better to do that rather than fighting it. How do we do that? Well, there are a series of studies like this method for example, giving people a word or phrase such as, I like it a moderate amount, and ask people to draw a little pie slice to say if the whole pie was gray, let's say that's liking it as much as you can. If the whole pie was white, that's disliking it as much as you can. How much does the phrase, like it a little mean? So somebody draws a little pie slice like this and then you calculate the area of the circle to provide a percentage. Why one would go through all this effort is not exactly clear because you can simply ask people to put a number from, let's say zero to 100 on the word or phrase, which other studies have done.

There are lots of studies like this. We have gathered up all the numbers, standardized them and built a series of tables. These are the studies across the top in one little slice of these big set of tables. These are the numbers that come out of the studies for phrases like absolutely perfect or superior or terrific or wonderful. When we average across the rows we produce this column and sort the column so you can then choose labels to go on as many points as you wish. If you want, let's say, five points on a rating scale that is unipolar, you can pick five that you like that are equally spaced. In terms of validity, comparing the verbal and numeric scales described earlier, Excellent gets an average of 94, Very good has 81, Good has 70, Fair has 51, and Poor has 21. This is a very common quality scale and surveys. You will notice that the gaps between these numbers are big, big, bigger, biggest. We can also look at he percent of people who place themselves in each of these categories who died during the 20 years after they rated their health in one of these ways. These are the death rates, Excellent = 1 percent, Very good = 3 percent, Good = 5 percent, Fair = 8 percent and Poor = 13 percent, small, small, bigger, biggest. This to me is evidence of validity in the sense that if we had taken this five-point rating scale and done what many people do, that is code it five, four, three, two, one and try to predict death rates, we would have gotten some positive significant coefficient. But, by capturing these gaps in meaning, in the numbers if we coded it instead 94, 81, 70, 51, 21 and predicted death rates, we would do significantly better. So, that's just a little bit of evidence to suggest that I think these numbers do have validity.

Now, with regard to dimensions that have a natural metric, so for example I could ask you how often do you go to the movies, very often, often, sometimes, rarely or never? I'm not going to take the time to go through all of the evidence on this so forgive me for just skipping this. What I can tell you is what this literature says is using those kinds of what you might think of as vague quantifiers actually cause many more problems than they solve. If what you want is a number just ask for the number. Other than saying, how often do you go to the movies, you ask people in the last month how many times did you go to the movies; you can get around that problem.

One illustration of how our instincts about labels have gone wrong can be seen very commonly in surveys with a scale that goes: Very often, Often, Sometimes, Rarely, or Never. This corresponds with numeric ratings of 88, 78, 20, 5, and 0 so nowhere near what you might think of as equal spacing across that dimension. But if you have a survey that has a rating scale like this, you can analyze it using numbers like this and I think do better than you might otherwise do.

Here's the conventional wisdom about question wording in research design textbooks: First of all, use simple, direct, and comprehensible words. Don't use jargon, be specific in your question, avoid ambiguous words, avoid double barreled questions that ask two things at the same time, and avoid negations if you can avoid the word not. Avoid leading questions, and include filter questions. The common-sense version of using filter questions is don't ask people what brand of car they have if they might not have a car. Be sure that questions read smoothly aloud, avoid emotionally charged words, avoid prestige names. When you look at textbooks and research design, it's almost like the later ones copied the earlier ones because they are remarkably consistent in this kind of advice.

The point I want to make is all of the aforementioned advices are fine, but there are really a series of challenges that we face in wording questions where we don't have nearly enough empirical guidance for. I want to just do a few exercises with you in my remaining space. In a survey that we did where we were asking people about how did they feel about various potential changes to the civil justice system? We found that on average, 75 percent of respondents found each change either very or somewhat acceptable. Imagine for a second, how would that number change if we instead asked how much they support each of those changes rather than how acceptable they find them. It sounds like a stronger word, right? Maybe even requires action. Maybe 75 percent wouldn't protest it, they would find it acceptable, but a smaller number would support it and in fact that is what we saw, 45 percent of people on average said they supported those very same changes in another experimental group.

But here's the challenge, what if we asked people, do you strongly or somewhat favor these options? What's that number going to be? How many people want to say closer to 75, favor sounds acceptable? How many want to say closer to support? You have to vote. Who says somewhere in the middle? The answer is we don't really know. If it was obvious to you, you would have voted. The answer is, it's almost exactly the same as acceptable. Would we have guessed that, obviously not. When we make this choice of wording, this is the problem. There is a New York Times Op Ed essay that reported the following information: more than nine in ten Americans would support special school programs for lower-class children beginning at age 8. These classes would be designed to motivate the kids to stay in school and to extricate themselves from poverty. That's the description of a result of a survey question using the word support. Here is what the question actually asks, "do you favor or oppose starting special school programs with young, underclass children when they are age eight design to increase their motivation to stay in school and to arose hope within them that they can lift themselves out of their miserable life situation?"

This is a bit of an extreme example, but the point is we're tempted to rephrase in ways to make the reading maybe more interesting. I think the point is changing words, you might imagine, can make the difference. Theoretically, I think I could suggest the following principals to you, might be good for selecting words. First, we want to mention only the construct that we want to measure in a question, this is what I will call univocality. So avoid some of the things that earlier advice suggests to avoid. Second, we want meaning uniformity, we want every question to mean the same thing to everybody and lastly we want economy of words, those principles seem very straightforward. I'm not going to review this evidence in detail, but there are studies that people haven't really looked at recently that very much to suggest that when you ask people to say back to the researcher in a cognitive interview what a question means from highly used questions that the rate at which people clearly misinterpret the question is sometimes quite distressing.

That process of designing questions that respondents understand doesn't go all that well and there are lots of studies suggesting, and again I won't review these in detail, that these are all different ways of showing the more ambiguously worded a question is, the more ambiguous the meaning of the words are, the less reliable and valid the data are. These are a series of principals that one can imagine might be good for selective wording. First of all, we can use a dictionary to select words that have only one meaning, or a primary meaning. That might lead us to avoid, for example, a sanction, that has two different and opposite meanings. Second, there are studies of familiarity of words, how often words are used in print and oral communication. We might lean toward using words that are familiar to people. We might lean toward words that are simple, have fewer syllables, sentences that have fewer words, questions that have lower readability scores that require less education in order to understand them. Most researchers do not design a questionnaire and compute the readability score on it to make sure that lower education people can understand it. You might want to avoid homonyms in a telephone survey, right? So, the word fair is often used in evaluation questions, is it a good idea to use a word that sounds like another word? Is it a good idea to use words in a written questionnaire lead or lead that are spelled the same, but you have to use the context to figure out what they mean. Obviously, a lot of pretesting seems like a good idea. These kinds of principles are all useful to think about as guidelines for potentially improving data quality, but we don't know. There's actually, I don't believe, a body of research even begun to explore the impact of the kinds of common sense decisions I just suggested to you to see if they in fact improve data quality.

I'd like to conclude this chapter with a series of studies that have shown how remarkably important question wording is, not only in terms of reducing respondent burden but how much research we really need to do in order to guide these decisions. For example, imagine these two questions: "during the last month have you seen a movie?" Or, "during the last month have you seen one or more movies?" Do you think that might make a difference? The percent less is not significantly different, but yet our instinct was maybe it would be. So we need to think carefully about these kinds of things. How about this, "during the last month how many times have you seen a movie?" or "during the last month what's the total number of times you have seen a movie?" Should that make a difference? It's a lot more words in the second case. Do they really mean the same thing? No they don't, they are different. How are they different?

Okay, the first one could be different movies where the second one could be how many times would you re-watch the same movie? Could be, so it does make a difference, there is not a significant difference between these two. That turns out not to matter whether we're asking about seeing movies or going out to movies. Another example, a direct question, "how many times have you seen a movie or gone out to a movie?" Verses a filter question, "in the last month have you seen a movie?" and if so," how many times?" So that, it turns out, it does make a difference that when we ask how many times have you gone out or see a movie, you get a larger number, significantly larger than if you asked the filter question first, in the last month did you see a movie and if so, how many? The argument that people make is that the reason why this difference happens is that the direct question how many times have you seen a movie in the last month seems to imply that of course you have seen a movie and pushes people in the affirmative direction. Whereas asking have you seen a movie as a filter doesn't do that pushing as much.

To test that hypothesis, you can do an experiment like this. Some people could be asked, "how many times have you seen a movie?" and the other two question forms take away that suggestion. "Some people see movies often, other people see movies occasionally and still other people never see movies, during the last month, how many times have you seen a movie?" Clearly. we're suggesting it's okay for people to not have seen a movie, or this last one, "how many times, if any, have you seen a movie?" Both these are designed to accomplish with this question what this filter is thought to have accomplished. When we compare answers to the three, we should see fewer movies being reported for the second two than the first one. Except there are no differences.

What we've learned there I think is yes, different forms of questions may produce different estimates, but it's not that explanation that is addressed here. I could walk you through lots more experiments like this and go through this exercise of asking you to guess whether wording would make a difference or not and sometimes we'd be right, sometimes we wouldn't be right, but there are lots of decisions like this that need to be made. Unfortunately, the reason we need to make these decisions carefully and the reason I call this a crisis at the moment is because of patterns of data like this.

I am going to end with this, the percent of Americans who said they believed that Barack Obama was born in the U.S. in surveys between April 2010 and April 2011 varied substantially across survey organizations. A variety of different organizations asked this question, some with numbers as big as 77 percent or more getting the right answer, some as low as 58 percent or fewer getting the right answer. Clearly, lots of differences between these organizations. Similarly, there were big differences between them in terms of the size of the partisan gap, the gap between Republicans and

Democrats. The difference between Republicans and Democrats was 40 percentage points or more in some cases and as low as 20 percentage points or less in others.

In my estimation, this is a series of high level, well-paid, well-trained, very experienced professionals designing questions, trying to measure exactly the same thing and getting very different results. Let me show you the wordings that they used in these questions. One was, "do you think Barack Obama was born in the United States or not?" It seems pretty straightforward. Another, "was Barack Obama born in the United States or was he born in another country?" Now, the second version is more explicit than the first and notice the word think is in the first wording, but not in the second. It is possible now to combine these by asking "do you think Barack Obama was born in the United States or do you think he was born in another country?" Now we've combined the two. Or, "do you think Barack Obama was definitely born in the United States, probably born in the United States, probably born in another country or definitely born in another country?"

If you put together the definitely and the probably on the two sides, you get notably different results than you get up on the other versions of the question, but there are even longer questions. "As you may know, some people have suggested that President Obama was not born in the United States, do you think that Obama (notice, no president there) was not born in the U.S., was born in the U.S. or it's not clear whether Obama was born in the U.S. or not?" So now we're adding into this "not clear" thing. We're not referring to him as President Obama. Notice back here its Barack Obama on this whole screen. It says the word "Barack," does that maybe do something? Also, why would you begin by suggesting what some people think and not what other people think? The folks that wrote this question said the reason they did it was to cover themselves. In other words, they and the news media believe this is a non-issue and we're embarrassed to even be asking this question. So they wanted to explain to certain respondents why they thought it was sensible to ask it. Okay, maybe that sounds reasonable, but is it a survey question a place to explain why you're asking a question?

Yet another one, "according to the Constitution, American Presidents must be natural born citizens. Some people say Barack Obama was not born in the United States, but was born in another country. Do you think Barack Obama was born in the United States or do you think he was born in another country?" My ` is this, if you and I were sitting at a desk and looked at any one of these questions, without seeing the rest of them, we might think it's reasonable to ask this, there's a basis, there's a justification for it. But the fact is, we as a profession are not guided by a single set of principles to lead us to a single optimal wording and as a result we publish results that look like these and I don't think this is good for us. This is one of the many places where I am hopeful that NSF will be able to consider supporting more research to help us to begin to grapple with language and to help us figure out how to get on the same page. The idea is not to say what's good is to have lots of researchers out there asking lots of different questions producing lots of different results and eventually we'll figure out, you know, it's probably better to get organized.

Let me just summarize quickly what I hope you get from this very, very brief and much too superficial, skimming across of this literature. First, lots and lots of studies are done that help us to inform the issues of questionnaire design and yet there is much more work to be done, especially with regard to language. I think we understand a lot about structure, but we understand much less about language. There is also an issue of dissemination. NSF of course has commitment not only to making discoveries but also to disseminating those discoveries through educational efforts and outreach. I think there is a real opportunity here because there are lots of people who don't know it's a bad idea to offer a 'don't know' response option. There are lots of people who don't know the order of answer choices and close questions matter and how to handle that. There are lots and lots of people who think it's fine to ask agree/disagree questions, so a major educational outreach effort to disseminate the findings of this literature would help.

Jon A. Krosnick is the Frederic O. Glover Professor in Humanities and Social Sciences at Stanford University, Stanford, CA, USA and a University Fellow at Resources for the Future. This work was supported by National Science Foundation Award [1256359].

54

Cognitive Evaluation of Survey Instruments

Gordon Willis

Cognitive interviewing – which its practitioners hope is a science rather than simply art – is the focus of this chapter. One of the issues we have to deal with is how much of what we do in the cognitive laboratory is idiosyncratic, unstable, and unreliable. I think there's a lot of research that could be done here. To put things into context, I see what we do as fitting within a total survey error paradigm. So, everybody has their own favorite source of error, which to some extent different chapters in this volume have talked about. We are concerned with coverage error, sampling error, non-response error, measurement error due to the interviewer, measurement error due to the respondents, post-survey error with analysis, and so on.

The form of error that I deal with is the component of measurement error referred to as response error, which is presumably controllable through questionnaire design. The notion is that this is worthy of research, and that small changes in questionnaire design/wording/format can make a big difference. As an example of the kind of research that I think should actually be done, when I worked at the National Center for Health Statistics Cognitive Lab, we tested survey questions intended to measure behaviors such asstrenuous physical activity. One such question was: "On a typical day, how much time do you spend doing strenuous physical activity, such as lifting, pushing, and pulling?"

G. Willis (✉)
National Cancer Institute, National Institutes of Health, Bethesda, USA
e-mail: willisg@mail.nih.gov

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_54

In our Cognitive Lab, it became clear very quickly that there could be a problem, because we had people say something like “Three hours,” However, if we then probed them by saying “Okay, tell me about your typical day,” we would find out that they do nothing more strenuous each day than reloading the photocopy machine. This looked like it was a demand effect, classically. That is, no one wants to say, “None”? because then you look like a couch potato.

So, the alternative we tried instead was to use a filtered approach. We just added a yes/no question – a basic resolution to a common problem. That is: “On a typical day, do you spend any time doing strenous physical activity…?” If the answer is “No,” we’re done. And it looked like respondents were happy to say, for example, “No, I work in an office.” If they say, “Yes,” then we branch to the second question and add the follow-up question about how many hours of strenuous activity. However, at that point, there was a senior member of the establishment who objected that we were adding another question–he didn’t want to do that. And, he asked: Does it really make a big difference? I thought it did, and I got a chance to test that out empirically. That is, rather than saying, “We know it’s true because it came out of the Cognitive Lab,” we could instead see if our effects occur in actual ecologically valid environment. Here’s what happens when we use both versions. There’s no filter for the original, which we said was going to produce relatively few people reporting no physical activity. And the filtered version, with the additional questions should raise the proportion of people who report no activity. This is essentially what we found, and the finding is large and significant and in the predicted direction, within a field pretest. So, on that basis, we think that we got huge orders of magnitude of response error. So, I argue that this is the type and extent of response error that we’re dealing with, with cognitive testing.

As a way of background, concerning what cognitive testing actually exists of, we’re developing, evaluating, and testing either a questionnaire or other survey-related material, such as a consent form, or introductory letter. They’re all amenable to cognitive testing, as it’s a flexible method. We recruit people differently than we do for regular surveys: We pay them, because we want their undivided attention, and we conduct one-on-one interviews, typically in a cognitive lab, but sometimes we just go to where we can find people. The basic approach involves verbal probing approaches. We either probe, or ask the person to think-aloud. The most common model for cognitive testing is to first find problems – and then fix them, by making repairs. That seems obvious, but these are different skills that are involved in those two parts, and each one, perhaps, could be amenable to different types of research.

Cognitive testing in any form seems to be best done as an iterative type process, where we do small rounds and make changes to questionnaires in between rounds. It doesn't seem to be necessary to do a lot of testing before we change, although that does bring up some interesting issues related to sample size, which I will get to. Concerning the role of theory – What has held up really well, I think, over the years is this basic yet enduring model by Roger Tourangeau, which asserts that people have to understand the question – be able to retrieve the information (do they have a chance of even knowing the answer) and decide to tell the truth. And then finally, we have to ask for responses in the way that they're going to give them to us. I believe that we've widened this view over the last 10–20 years to incorporate more sociodemographic, sociocultural type of issues.

We operationalize cognitive testing through asking probes like, "What does the term dental sealant mean to you,". Or, by paraphrasing: "Can you repeat the question in your own words," and see what you get back. Or, asking how confident people are in their responses. For example, how do they recall something that has a very long reference period? We devise specific probes, and my favorite, perhaps, when I can't think of anything else, is, "Can you tell me more about that?"

So, with that as a very quick background, there are a number of unresolved scientific and methodological issues that we have to grapple with. The first is basic: "Is cognitive testing reliable and valid?" And that in effect relates to the issue of: "Overall, is this an effective thing to be spending our money and our time and our effort on, and hiring staff? The second issue, looking ahead, is: "Is cognitive testing useful for 'the survey of the future,' whatever that's going to encompass".

Further, there are several government laboratories that are devoted to cognitive testing, and at private contractors as well: There are labs at NCHS, at the Bureau of Labor Statistics, Census Bureau, and so on, which is a good development. However, we don't really know, for one, whether these independent labs, testing the same questionnaire, would come to the same conclusions. That is, how idiosyncratic is this all? There have been a few studies on this, with mixed results. One that I was involved with compared three laboratories and found the same questions to be problematic, but for different reasons across the different laboratories. Is that a good thing? Or a bad thing? It wasn't really clear.

A more recent study was by DeMaio and Landreth at the Census Bureau – who also made a three-team comparison. They were unable to tell which is best, in the horse race sense, or whether there were certain procedures that were better than others, except, from a procedural point of view, they did decide that it's important to listen to recordings of the interviews in order to have reliable results. So, we get some information, in any case, from doing these studies.

Further, for a project that I was involved in, the investigators tested a self-administered questionnaire on perceptions of either breast or prostate cancer risk, depending whether you're male or female, and where the functioning of the questions was unknown in advance. That is, these were newly scripted questions. So, I thought it was a good opportunity to have four different labs do parallel testing. Labs were told: Use whatever your techniques are, test this, and tell me what you find, and then I could compare the reports. This was a large cognitive interviewing study, almost 150 interviews in 4 different cultures or languages. Again, the research question was: Were the written results going to be very similar, or very different? So, we wound up with 148 interviews between the NCI, Westat, NCHS, Public Health Institute, in 4 languages.

The questionnaire said: "Please circle the single number, on a scale from 1 to 5, that best describes how concerned you feel right now about the following things." And then we had "Feelings of concern now," to remind them, and then things like, "breast cancer occurring in me, my family's history of cancer," – 24 things. Finally, the response categories were "Not at all," to, "Extremely." This is a common kind of self-administered form. We had a preconceived view of how the respondent would handle the task, cognitively: The respondent looks at the first question, and thinks about feeling of concern. Then, he or she connects those cognitively, and then picks a response. So, the four labs tested the questions, to find out if that's what, in fact, happens. What came up instead is, a model of what respondents actually do. First, they didn't read the instructions at all. They also don't read the part about feelings of concern now, at all. So, they were not thinking about the critical element of 'concern.' Instead, they just see "Breast cancer occurring in me," and, "Not at all," to "Extremely," and they just circle something, whether it's how likely is this, or how much has this occurred, or something else. But this "something else" is something other than "how concerned" I am about it.

So, for example, for "What having breast cancer would do to my body," people would say "Well, it would do very much to my body." That's not what we're trying to ask. And for "My chances of dying of breast cancer," the person would say, "Somewhat" – but it had nothing to do with how concerned they were about that. So, that was interesting, but what was more interesting was that this finding was ubiquitous. Across all the different labs, and the different languages, the reports I got back basically said, "This approach does not measure perceptions of degree of concern because concern was ignored." The important point is that nobody saw this coming. And it was reassuring to see this same result come up repeatedly across the four labs. That study constituted evidence of reliability, but there have also been opposite results. Kristen Miller at NCHS has done multi-country cognitive laboratory studies and experienced real problems, because you can get big differences in results across countries and don't know why. It's got a lot to do, I think, with different training between staffs of different labs. But the point is, this issue still needs to be addressed. Under what conditions are the cognitive interviewing results reliable? And especially: What do we have to do to make that happen?

The second unresolved issue is about procedural evaluation. This means working down in the trenches, on the nuts and bolts that only a cognitive interviewer could love, like "Which of our particular procedures are useful?" There used to be a fair amount of research in this issue. So, investigators contrasted think-aloud, versus retrospective probing, in which you administer the questionnaire; and leave all the probing till the end (debriefing). My Census colleagues compared these two approaches on a paper questionnaire, and they found that the findings were very similar, except that subjects with low educational level tended to blow the skip patterns on the questionnaire under think-aloud. I think that makes sense. If you're trying to think aloud and also follow the questionnaire, and you're not very good at some of this, you tend to mess up. So, that may not be really illustrative of what happens when you're not forced to think aloud. So, it does matter sometimes which procedure you use. But, they wound up recommending, think-aloud for some interviews, and retrospective probing for others, in order to find something useful.

There are a couple of other studies, where the research has been done on concurrent and retrospective probing versus pure think-aloud. The authors said that in general, the problems were similar in both of these for self-administered questionnaires. For example,

Susan Schechter at NCHS did a similar study and again said, "It's good to mix techniques." But that research has, to my knowledge, mainly stopped and we certainly haven't taken it to the next level, which I think is unfortunate. What has been of more recent significant interest is the contentious issue of appropriate sample size. Cognitive interviewing is a qualitative technique, so we don't do hundreds of interviews, usually. The sample sizes are typically small. We're doing no more than 30 interviews often, and sometimes people do less than 10. What do we get from that? Is that useful? Do we reach what they call, in the qualitative field, "saturation", meaning that doing more interviews doesn't give you much more information?

So, how many cognitive inteviews is enough? Thinking about it further; that's a complex question because it could break down to mean, "How many is enough to identify a problem?" That is, how many interviews do you have to do before that problem first occurs? Or, do we mean something like validating in the sense that we set a criterion and say, "I want to see a problem happen in three of ten, or eight of ten interviews, before I conclude that it's a problem." This has not been well studied, and is something that we have to grapple with. There's been a little bit of work on it – Johnny Blair and Fred Conrad did a nice study, assessing what happens when you increase the number of interviews you do.

The critical graph from their paper, in Public Opinion Quarterly, shows that additional interviews continue to produce observations of new problems, although the rate eventually decreases. So, the argument could be made that we should be doing more interviews. Now, there is also pushback against this. I know other people who say that Blair and Conrad picked the wrong dependent variable. Anybody that's motivated to not do more interviews can find all kinds of counter arguments. But this is something that we need to look at a little more closely.

So, turning to the survey of the future, we get into mixed modes and novel administration methods issues, as well as "Is cognitive testing useful for these purposes?" I think it is. I think it's a natural; we're well positioned. Cognitive interviewing has a strong history of attention to administration mode, in particular because interviewer based versus self-administration are very different cognitively, as to whether they deal with auditory-based information versus reading/visually based information. So, we're already there, in what we're dealing with. We've increasingly focused on Web usability. Jennifer Romano-Bergstrom has done a lot of usability testing, and I think that cognitive testing and usability testing are starting to converge. That seems like a good thing because it enables us to use our current methods as we go to new technologies that increasingly involve computerization.

I also think there's interesting new research that's being done, for example, by Jennifer Edgar, Bureau of Labor Statistics, on more electronic types of cognitive interviews, either Skype or Internet-based, where you have your respondents fill out questionnaires and answer probe questions by writing and sending you back the stuff. This may be limited, because you can't conduct flexible or follow-up probing. But you can do a lot of interviews for the same amount of money as one interviewer-administered cognitive interview. So, if you lose some information, maybe it's worth it. I'm open-minded enough to consider the possibility.

The other issue here that's really important is cross-cultural applications. This is a hugely important issue in survey methods generally.

And you've got to know this to do the cross-cultural work, or else you're just not asking the right questions. So, the issue is, can we use our evaluation pretesting techniques that we already have fairly well established to obtain cross-cultural comparability, or do we have to go back to the drawing board? I would argue that there are strong issues of cultural misapplication, where we are just asking the wrong questions. You can't ask poor Hispanic women about physical activity by using a questionnaire that was validated on Harvard grads. And then there are linguistic errors and challenges.

We have to have methodological research into these cross-cultural applications in particular.

Getting back to cognitive interviewing, the question that we always wonder, doing these cross-cultural, multilingual, cognitive interviewing investigations, is: Do cognitive interviews themselves function similarly across groups? This is something that, again, my colleagues at the Census Bureau are really interested in, because if people respond differently to cognitive probes, you can come to the wrong conclusions. If something looks like it's a real problem in Spanish, but is just an artifact of the probing measurement-based process, then we're really not doing our job right. So, the issue really is: What kinds of modifications do we need to, in particular, analysis procedures. This is another place that's ripe for further work.

So, to be very overt about what I think should be studied and funded, perhaps by NSF, would be, cognitive testing reliability overall. We've got to figure that out, because that's vital to continue to support this endeavor at all, I believe. This procedural parametric evaluation is really important if we ever have a hope of developing best practices that are going to be commonly applied.

As far as applications to the survey of the future, we've got to get into, of course, the new administration methods, so we're staying current with developments in the field. And again, cross-cultural applications are really where it's at as far as maintaining, again, our ecological validity. I'll add, too, that a lot of this research is not that expensive to do. But this methodological research, related to cognitive pretesting, or pretesting results, pretesting procedures in general, is fairly low cost, and I think you get a decent bang for the buck out of it.

Gordon Willis is Cognitive Psychologist at the National Cancer Institute, National Institutes of Health. He has previously worked at Research Triangle Institute and at the National Center for Health Statistics, Centers for Disease Control and Prevention, to develop methods for developing and evaluating survey questions. He attended Oberlin College and Northwestern University. He has co-authored the Questionnaire Appraisal System for designing survey items, and the Cognitive Interview Reporting Format for organizing study results; and has written two books: Cognitive Interviewing: A Tool for Improving Survey Questions; and Analysis of the Cognitive Interview in Questionnaire Design. He also teaches questionnaire design and pretesting for the Joint Program in Survey Methodology, and at the Odum Institute, University of North Carolina. His work involves the development of surveys on health topics such as cancer risk factors, and focuses on cross-cultural issues in questionnaire design and pretesting.

55

Interviewer Deviations from Scripts

Nora Cate Schaeffer

When I was asked to contribute a discussion of "Interviewer Deviations from the Script," the first question that I had was, "Exactly which script?" There are at least two different scripts that could be involved. One is the script of the survey questions themselves, and the other is the script of the rules of standardization. Rather than providing a comprehensive review of the literature or summarizing my own work, I will describe some studies that might be thought provoking as a way of thinking about issues that arise as interviewers implement these scripts. Along the way, I will provide some illustrations and some thoughts about future research.

Given the increasing use of web and other self-administered forms of data collection, why think about interviewers at all? Well, interviewers will probably continue to be important in recruiting the first stage of critical panels. In addition, Computer-Assisted Personal Interview (CAPI) surveys have become increasingly complex so that, in addition to asking questions, interviewers sometimes have to complete complex histories and timelines, collect anthropometrics and tests of different kinds, and bodily fluids, such as blood spots or saliva samples; interviewers may also be asked to persuade respondents to give permission to researchers to obtain and link records, such as social security records. Telephone interviews and interviewers also continue to be important for reinterviews in panel studies. So, even though

N.C. Schaeffer (✉)
University of Wisconsin, Madison, United States
e-mail: schaeffe@ssc.wisc.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_55

self-administered modes have increased in importance over the last decade or so, interviewers will continue to play an important role for the foreseeable future.

How can we contextualize what happens in the interview? A model that we have been working on over the last few years (Schaeffer and Dykema 2011b) is an attempt to think systematically about what influences the behavior of the interviewer and the behavior of the respondent in the survey interview. One thing that we do not often pay a lot of attention to, because we take it for granted, is the role of technology in influencing both the behavior of the interviewer directly as she has to manage the instrument, but also indirectly by limiting and shaping the characteristics of the survey questions. So, our model has technology as an important, sort of distant, cause of the behavior in the interview. Question characteristics are a pretty obvious influence – I think we are used to thinking about question characteristics as an influence on the behavior of the interviewer.

The behavior of the interviewer is also shaped by interviewing practices and the training (including regular monitoring) that they receive in how to behave in standardized ways. Once we listen to recordings, which now is easier to do, we can see that the behavior of the interviewer is also shaped by what we could think of as interactional or conversational practices. A very obvious example is the practice that interviewers have of saying "Okay" when they acknowledge an answer or when they return to the agenda from some digression. If the interviewer asks a question, and the respondent gives an answer, the interview may say "Okay" as a way of both closing out that sequence and returning to the agenda. That is very much an interactional or conversational practice that has been observed in other kinds of contexts.

And then, finally, the behavior of the respondent is probably one of the most important influences on the behavior of the interviewer and ways in which the interviewer deviates from standardization. So, all of these different pieces – technology, question characteristics, the rules of standardization, conversational practices – all of these things affect the behavior of the interviewer, directly or indirectly, but they also influence the behavior of the respondent and through that, the behavior of the interviewer.

How do these factors connect to each other in leading to deviations from the script? For an illustration, the work by Marek Fuchs (2000, 2002) that also involved, in various combinations, Mick Couper and Sue Ellen Hansen (Fuchs et al. 2000) about how rosters or grids were implemented is instructive. Grids or rosters are a kind of format that we sometimes use in paper instruments to display the information we want the interviewer to collect. When instruments moved to CAPI, we turned the rosters and grids into series of repetitive questions, partly because there was no way – and it is still very difficult for some software – to get a grid on the screen in the same nice way that we could on paper. Fuchs and colleagues examined the implication for the interaction of a roster with an item-by-item sequence of questions organized by topic compared to the interaction that results when the roster is organized by person. This was a fairly small study, but they found that, in some cases, respondents volunteered a lot of information all at once. That is an instance of a kind of conversational practice occasioned by the fact that the respondent can infer what questions are coming next. For example, the respondent tells the interviewer, "Everybody who lives in this household is white." When that happens, the interviewer now knows the answers – or what the respondent thinks are the answers – to the next questions. Nevertheless, the interviewer must manage the instrument and the interaction in a way that is consistent with the rules of standardization.

This is a kind of situation in which there can be deviations from the script, because the interviewer does not ask all the questions that they were supposed to ask or because the interviewer does not follow other rules of standardization as they try to manage the information that the respondent has supplied unexpectedly. If the instrument is designed as a roster organized by topic, Fuchs' results indicate that the information is provided much more quickly. Presumably, that is because the respondent does things like saying, "All of us who live in this house are white" and so answers many questions at once.

With that as an illustration of how technology, interviewing practices, respondent behavior, question characteristics, and so forth might all be tied up together, we turn to other kinds of deviations. An obvious deviation is "not reading the question as worded." Another is "inadequate follow-up behaviors." For follow-up behaviors, there is not an actual script, but there are principles of standardization that interviewers could follow or deviate from. These follow-up behaviors include providing definitions and feedback or acknowledgements. There are also intrusions of conversational practices that present challenges for the rules of standardization, such as "reports," that is, informative answers that do not match the response format of the question.

A few years ago, we summarized findings about the relationship between behaviors of survey interviewers and the resulting quality of measurement (Schaeffer and Dykema 2011b). We restricted ourselves to studies that were record-check studies or had, in the case of variable errors, some formal analysis of those variable errors. The important study by Groves and Magilavy (1986) did not find any consistent relationship between question reading and interviewer variance when they examined a set of 25 items. Hess, Singer, and Bushery (1999) found no effect of exact reading of question on the test–retest index of inconsistency in their examination of 34 questions. Dykema et al. (1997) used a record-check study, the Health Field Study, to compare answers to survey questions with health clinic or hospital records. They found that a substantive change in the reading of the question had no effect for nine of ten items and increased accuracy in one case. They have fairly complicated tables, and there are several variants of this result. But looking at 11 items, there was no effect of question reading for nine items, decreased accuracy for one, and increased for another.

This is clearly a fairly limited number of studies. But when we look at them together, one thing we might say is that the way a well trained and supervised standardized interviewer reads and delivers the question does not really have much effect. However, it is very important to keep in mind that we are looking at standardized interviewers, and although the level of standardization probably varies over these studies, most of these are telephone studies where the interviewers are pretty carefully monitored and supervised. Thus, research to date suggests that when standardized interviewers are carefully monitored and supervised, most deviations from question wording that they engage in do not seem to have a big effect on the quality of the data.

The other thing we can say is that there are a few examples in which it seems that deviations from the exact question wording made some difference, either improving accuracy or decreasing accuracy. But we do not know why, because the analyses weren't able to go to that next step of, "What actually happened? Was this just one or two interviewers who were responsible for this effect or was there some specific behavior that improved or worsened a question? And, if so, what were those interviewers doing to make the question better or worse?" Those would be good things to know, but they are expensive questions to answer because you can find the cases, but then you have to go through the tape recordings, produce transcriptions, and code them, which is slow.

Deviations from the script also arise during follow-up behaviors. One of the things we know about follow-up behaviors is that when probing happens, data quality is lower. Probing is associated with increased interviewer effects, and follow-up behaviors by the interviewer are associated with decreased accuracy, whether or not the interviewer is following the rules of standardization when she does the follow up. When there is any probing, and it has significant effects, it seems to be associated with reduced accuracy as well as increased interviewer variability.

When follow-up behaviors happen, there is usually a problem of some kind – either the question has a problem, or there's a lack of fit between the question and the respondent's situation, or the respondent can't remember, or there is some behavior of the respondent that expresses some difficulty with answering the question. It is not that the interviewer causes these problems, it is just that when probing is needed, there is a difficult situation. The contribution of the design of the question to such situations is illustrated in a very nice figure from a paper by Schnell and Kreuter (2005) and, you probably could see something similar with the O'Muircheartaigh and Campanelli (1998) paper, if they had a similar plot. The plot shows an index of features of items that might be associated with increased interviewer variability. They find that the fraction of the total cluster variance that is due to the interviewer gets bigger as – and more consistently big – when the survey question has some of these "harmful" properties, such as being an open question.

Another follow-up behavior that we might be interested in is providing definitions. A famous experiment by Schober and Conrad (1997) compared standardized and flexible interviewers. Both sets of interviewers were trained to read the question as worded, but the respondents in the flexible condition were given extensive additional instructions about how important it was to ask questions, and the flexible interviewers were trained to follow up, as needed, to help the respondent understand. The techniques that were used appeared to be effective at improving the understanding of the respondent when the respondent scenario was not a good match to the question, but the interviews took more time.

It seems as though providing definitions can improve respondents' understanding of complex concepts when the respondent's situation needs it. But, we do not have studies that compare that method of providing definitions with other methods of providing definitions, many of which were discussed by Schober and Conrad (1997). It would be helpful, in thinking about production interviewing, to think about expanding the comparison, not just to providing definitions this way versus no definitions, but thinking about other ways of providing definitions and, with enough money, doing it in a way that both variable errors and bias could be assessed simultaneously, something that does not happen very often.

Feedback is another site for possible deviation from the script, and there are not a lot of studies that examine this. One study, Groves and Magilavy (1986), looked at a study that included an experiment that compared completely scripted feedback to a very limited range of ad lib feedback. But, even within this very narrow range of types of feedback, they saw a tendency, although not significant, for the rho_{int} to be smaller for the group with the scripted feedback.

One of the few experiments that look at feedback, is Dijkstra's 1987 study, which used "personal" feedback to motivate respondents. This issue of respondent motivation is very important, and I think likely to be increasingly important if, when we get to peoples' houses, we want to make them work hard and do all kinds of things for hours on end. Paying attention to the respondent's motivation is important, and Dijkstra experimented with this. He found, even with a small number of interviewers and testing for the interviewer-level effect, some significant effects of person-oriented feedback on behaviors of respondents associated with motivation.

One kind of interview that sometimes does not have a script is event history calendar interviewing because it is very flexible. Event history calendar interviewing illustrates the very close relationship among technology, instrument, and interviewing practices, but it also illustrates the challenges that come up in trying to design studies to evaluate different kinds of interviewing. If you want to compare event history calendar interviews with standardized interviews, just finding the right place to make the comparison is very hard. Different behaviors occur in the two different styles of interviews, not surprisingly, because the interviewers are trained to behave differently. One set of interviewers has a script. The other set of interviewers does not. So, how to make assessments of one group versus the other group is not straightforward, and when you observe differences, it is very hard to know to what you should attribute the difference. It could be that some observed differences are due to behaviors that we are not even tracking, such as motivating feedback or something like that. The other thing that we see in thinking about this kind of study is how important it is to have the right kind of statistical design and analysis.

I wanted to return briefly to the idea of a roster or grid. In the example that we started out with, the grid, respondents spontaneously do things that deviate from the way the standardized interview assumes they are going to behave. And we might think about how we would design a method of data collection to get complex information of the kind obtained by grids.

A study that we developed recently at the UW Survey Center aimed to describe complex family structure using a collaborative approach. The label is taken from the Suchman and Jordan (1990) paper in which they discuss having an instrument designed so that the respondent and interviewer could both see the form being filled out. Because this study required repetitive information, we wanted an aid that would convey the structure of the task to the respondent, make collecting the data more efficient, support motivation as well as recall, display the information to help the respondent see the structure of the task, and also give the respondent the opportunity to correct any information that was recorded incorrectly. We wanted the interviewer to be able to record answers the way the respondent provided the answers, because we expected respondents to use the kind of conversational practices reported by Fuchs et al. (2000).

We know little about how visual aids and such are actually used in interviews. When we think about using technology to give us better visual aids, we have to think about what interviewing practices go with the technology. Every technology, whether it is paper or CAPI or some other kind of technology, needs a set of interviewing practices that somehow helps the interviewer manage conversational practices, reduces interviewer variability, supports the motivation of the respondent, and still keeps reliability and validity in mind. So when we introduce something new, we need to think about, "What are the rules for interviewing that goes with it?"

The visual aid we designed for the WiscMoms study was a kind of dynamic household roster. The respondent holds a tablet, and the screen is filled out dynamically. As the respondent gives the names of the people in the household, the names appear. As they identify the relationship and the birth date, all that information appears to the respondent so that the respondent has the big picture, and can also see what information the interviewer has entered. We had a similar display for a timeline about co-residence – when the children were born and when various fathers and boyfriends co-resided with the mother and the children. This, in the interview, is a dynamic display of all the children we have learned about, all the fathers we have learned about, and the time periods of co-residence from the time the oldest child was born. We also had repetitive questions about who contributed to the household. We were able to list all the people that we were going to ask about, all the different kinds of contributions we were going to ask about and, again, display that to the respondent.

Doing all this required coming up with rules for how to train the interviewers. For example, one of the things we needed to develop was rules for how to do a verification. We wanted rules that were more detailed than any rules we were able to find guidance on. We also realized we needed to train interviewers to notice inconsistencies and distinctions of complicated kinds. For example, Andrew and Sarah are brother and sister, but they could have different fathers, and we wanted the interviewers to be aware of that and be able to ask the respondent the appropriate questions in order to confirm the correct information.

The information we trained interviewers to use in verification was extensive. We had all of the people in the household, all the children, and then, the different meals that they could get at their daycare. We wanted the interviewer to be able to take information volunteered by the respondent in a fairly flexible way. Some of the challenges are: how do respondents provide information, how do they use the display, how can interviewers manage the variable provision of information by the respondents, and what is the set of interviewing practices that support the use of the display

Some of the things that we need to look at in future research on interviewing include complex interviewing tasks like the event history calendar, measuring household structure, and such; non-interviewing tasks, including physical measurements, cognitive assessments, and so forth; and how to design questions that are sensitive to conversational practices and stimulate recall.

What do we need in study designs? We need manipulation checks when we have experiments; identification of key behaviors that can be coded reliably; and we need outcomes built into the design that can be compared across conditions so that we can come to strong conclusions. We need experiments that are large enough to have a sufficient number of interviewers and appropriate assignment of respondents to interviewers when we are measuring and assessing an interviewer-level behavior. We need the right analytic models, as well as designs that let us look at, ideally, validity and reliability together. These studies need to use a realistic research context that can serve as a model for large-scale production and include some kind of development and assessment of interviewer training and monitoring. Finally, there continues to be a role for small-scale lab experiments as well as for large-scale field experiments, which tend to be quite expensive.

Nora Cate Schaeffer is Sewell Bascom Professor of Sociology and the Faculty Director of the University of Wisconsin Survey Center at the University of Wisconsin-Madison. Her current research focuses on interaction when the sample member is recruited and during the interview and on instrument design issues.

56

Coding Open Responses

Arthur Lupia

When Jon Krosnick and I were principal investigators of the American National Election Studies (ANES), we learned a number of very important lessons about coding open-ended questions and in this chapter I hope to convey those lessons. What I'm going to describe is based on the work and experiences of a lot of different people. I'll start by outlining some background and describe some challenges in the domain of open-ended coding. I'll tell you about an example that was our trial by fire in this domain. Then, I'll talk about some general attributes of the approach that we have tried to develop and are taking at the ANES to make open-ended coding something that's more legitimate and credible than it has been.

Regarding the ANES, the brief overview is that it's widely considered a gold standard of election studies. Its origins date back to the University of Michigan in the 1940s. Our objective is to try and provide data that can facilitate lots of different hypotheses in the social sciences pertaining to elections, and the incredible scientific opportunities that arise when you have hundreds of millions of people making a comparable choice on the same day. Jon and I were the principal investigators from 2005 to 2009. Now new principal investigators at Michigan and Stanford capably run it. Some of the experiences that I'll tell you about pertain to the 2008 version of the ANES time series.

A. Lupia (✉)
University of Michigan, Ann Arbor, United States
e-mail: lupia@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_56

The main study that the ANES is known for is a time series that takes a core set of questions, asks them every election period, and then adds new ones as circumstances request. Each version of this survey happens in two waves: once in the two months before the election, and once about six weeks afterwards. We try to get in the field for the post-election wave as soon as possible after Election Day, and really try to complete it before people start going away for the holidays. In 2008, we had about 164 minutes of interview time. Lots of questions. Most of the questions we ask are close-ended, typically with very few response options. But there are a couple of questions we ask where we want the respondent to answer in her own words. Examples of such question solicit views on "What's the most important problem facing the nation?" as well as views on what respondents like and dislike about the candidates and political parties.

Probably one of the more used and famous open-ended questions on the ANES has to do with a respondent's ability to recall certain things [aka "political knowledge" questions.] Here's the beginning of a kind of question that's been asked for close to 30 years. "Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public....What about William Rehnquist? What political job or office does he now hold?" So that's the type of question that we're dealing with. Respondents answer in their own words. Users of ANES data have expectations about what we're going to do with answers to these questions. ANES users expect us to convert open-ended answers to numbers. These numbers are then typically used in correlations, regressions, and things of that nature, to make inferences about all kinds of things about choice and the election. So people take the numbers that we produce and they draw an inference. The critical thing about this process is that the users base their inferences on beliefs about what each of these numbers means. That's the key here. The beliefs that people have about these data and are their beliefs correct?

When we think about the beliefs that people have about these numbers, many people believe that open-ended coding is easy to do, that it generates valid measures, and that it's performed well by survey organizations. In fact, for the 30 years that we've been doing the recall questions, we had little to no record of users asking us questions such as: "How did you make decisions about what answers were correct or incorrect?" The ANES has no record of being asked for, or producing, reliability statistics. Yet, the recall questions are widely used. And this broad use without questioning accuracy represents a belief in the user community about important qualities of this data.

When Jon Krosnick and I were principal investigators of the National Election Studies, we discovered a different reality in terms of how these numbers were produced and what they actually meant. A lot of what we learned surprised us and disappointed us, and so we started a further investigation. What we found was the practices that the National Election Studies were using to produce open-ended codes were not unusual. So here's the fundamental question in front of us: What is the correct inference for a user to draw from an open-ended response – from a coded open-ended response to a survey question? And the answer to that question is going to depend on what question we've asked, what the respondent says, and then some decisions that are made after the interview is conducted about converting those words into numbers. My emphasis now will be on those decisions and what we know about them.

I have a goal in mind for how we might think about what we should do. The goal is to produce measures that are credible and legitimate. By credible, I mean that the numbers have some property that make them believable, that make people trust that when see a number, they actually know what it means. A lot of social scientists are interested in credibility in their own inferences. We're also interested in legitimacy. That is, when people have questions about, "What does this number mean?" you can say, "We produced this number using a set of principles and standards that we can defend and that you might be willing to accept. So that when you run regressions, you can better understand what the number represents." Those are the big picture goals and Jon Krosnick and I tried to achieve those goals at the ANES through increased procedural transparency. As you will see, that was badly needed. A lot of people in the user community didn't understand how the codes were produced. The ANES also needed more rigorous documentation. One of the reasons that users drew mistaken inferences about ANES data, and data from lots of other surveys, is that the ANES and many other surveys do not rigorously document many of their coding practices. They do not record the instructions that one person gave to another person or the sequence that led from words being converted to numbers. By doing those things better, we hope to increase the credibility of the ANES's codes of open-ended data.

I'll take a moment now to tell you about our trial by fire. And that has to do with what political scientists call "political knowledge." Scholars and the public have a lot of beliefs about what the public knows about politics. I have a quote that is representative of a wide set of critiques. This quote says "close to one-third of Americans can be categorized as know-nothings, which is not to say that the other two-thirds are well-informed, right." And so the basic claim is that we have a knowledge base, where we ask people pretty simple questions, and they appear not to be able to give correct answers. But this is not just a pop culture thing – Bob Luskin is an academic who studies this as well. He uses ANES data to support his broad critique of the American public. So what is the knowledge base from which this type of claim is drawn? The National Election Studies data supply a political knowledge measurement for a lot of people, and the actual data for others.

A study done by Jim Gibson, of Washington University, and Greg Caldiera of Ohio State really brought to Jon Krosnick and my attention a big challenge we had with open-ended coding. To see the challenge, here again is the ANES question about William Rehnquist. Let's look at the data, what was reported, and how people used data from this question. In the 2004 National Election Studies, only 12 percent of the people in the sample were coded as having answered this question correctly. So that's interesting, this type of result appears to confirm the type of claims we saw earlier about how ignorant the public is about political topics. Let's think about where that 12 percent number came from. Prior to 2008, the way that this data was produced was that we asked the recall questions, (that's what you just saw here) in an open-ended format, but only numerical codes based on respondents' answers were released to users. Prior to 2008, the transcribed verbatim responses that record the respondents actually said were not released. The ANES had beliefs about what would happen if we released those things. In particular, there were some concerns about privacy. So, the long-standing policy of the ANES was to release only the numerical codes and not the transcribed responses.

But we did have a way for people who are interested in seeing the verbatim response to get the data. They could go through a restricted data access program to basically sign their life and their children away in exchange for being able to look at some of our data that we don't release. Gibson and Caldera did that. It was a good thing that they did, because at the same time that the ANES was in the field, Gibson and Caldiera conducted a study about public knowledge of the Supreme Court – and they had asked a question just like the one that the ANES asked.

And, yet, the answer that they got was very different than ours. They found that upwards of 40–50 percent of the public could answer this question correctly, and they couldn't understand why they got such a different answer than the ANES staff. After they got access to the restricted data, they revealed the reason for the difference. What they determined was that in 2004, an ANES respondent was graded as answering the question about William

Rehnquist correctly only if their answer said both "chief justice" and "Supreme Court." So a respondent who said only one of those phrases was graded as incorrect. And it turns out that another 30 percent of ANES respondents actually identified William Rehnquist as a Supreme Court justice, but were marked as incorrect because they didn't say both "chief justice" and "Supreme Court."

Gibson and Caldera also asked the same type of question in different ways. This work too gives a sense of what the ANES's 12 percent number does and does not mean. They also asked a multiple-choice question. You can see it here on the screen asking who was chief justice at the time of the interview, was it William Rehnquist, Louis Powell, or Byron White? Seventy-one percent of people got that answer right. This finding also raised serious questions about what, if anything, users could learn from the ANES data.

As we started our own investigation, we looked at other ANES studies. For example, we looked at the 2000 ANES, where the same open-ended question and coding scheme was used. And we found that of the nearly 1,600 respondents, close to a quarter of them said that William Rehnquist was a judge, or that he was on the Supreme Court. And, yet, if they didn't say both things, they were coded as having answered incorrectly. This slide shows a list of responses that shows different things that 2000 ANES knew about William Rehnquist. All were coded as "incorrect."

So you can see we have a problem here, but this wasn't the end of our problems because Jon and I then started trying to answer questions such as "How is this happening? What was the process that led to this type of outcome?" And in the process of lifting up rocks, we found another error. One of the other identification recall questions asks about Tony Blair, remember him? There he is [a photo is shown]. The ANES asked the question: "What job or political office does he now hold?" Now what's interesting about this question is that people who code the data were given some instructions. We've looked high and low for these instructions. As a general matter, we couldn't find any. But we were able to find a paragraph about what to do in this case. The paragraph read: "The reference must be specifically to Great Britain. The United Kingdom is not acceptable because Blair is not the head of Ireland."

Okay, so anyone in 2004 who said that Tony Blair was the prime minister of the United Kingdom was marked "incorrect." So now this is a pretty serious error. Once again, we asked "How did this happen? How did this happen not just once, but over a period of multiple years?"

The answer begins with the fact that during an interview, an interviewer asks the question and transcribes the respondent's answer. One thing that we can talk about is the quality of these transcriptions. As we investigated that part of the process, we found the transcriptions to be of varying quality. When we took over the ANES, we arranged to have these parts of the interviews recorded. We think there might be some virtue in recording all of these interviews so that we don't have to deal with the variation in transcription quality in the future. Going back to how the ANES had been operating, after the interviews occur, several weeks after, the staff implemented a coding scheme. Now as we looked into this coding scheme it was reminding me of people looking for weapons of mass destruction in Iraq. We kept expecting to find the documentation about exactly how coding was done and what instructions were followed, how codes were validated and so on. But what we found was really little to no record of instructions to the staff. Basically, the way that the coding was done was staff members would hire undergraduates or other people and give them a list of codes and then basically tell them to make a judgment. There was no documentation, and no reliability analysis. Moreover, it appears that a single coder did much of the coding – with no reliability checks.

So this was problematic, and we had an initial response. The first thing we did was to send out a letter to the user community alerting them to the situation and talking about what we would do next. What we would do next is based on the principle that you see on the screen now. "A basic expectation of scientific research is that you be prepared to document, archive, and share everything so that what you do is available for scrutiny by others." I think this is how you develop legitimacy and credibility. In this respect, we're channeling Richard Feynman. Feynman at CalTech's 74th Commencement Address talked about honesty, and the idea of giving people all the information, to help people judge not just the judgment that you make, but how you got there. And this type of philosophy was really what we started to aim the coding process towards.

So what we did immediately is, to the extent that we could for 2008, we made redacted transcripts available. And the reason we were able to do this for 2008 is because we had written the consent form in a way that would allow us to do this. We would have liked to, and may still like to release earlier transcripts, but there's a legal concern about whether our consent forms from those years allow us to release that information. Since the pre-2008 respondents were not given a chance to explicitly consent to this, we haven't released them.

We then approached NSF, who was incredibly supportive and realized the types of problems that not just we had, but realized that a lot of other data collections had similar problems with documentation. We then ran a conference in Ann Arbor bringing together a number of people that are in the room right now, some machine coding people, linguists and others, to try to work through and develop new best practices for coding.

Next, we started to assemble expert committees for each of the ANES open-ended questions. We asked these groups to help us develop for each question mutually exclusive and collectively exhaustive coding schemes. Mutual exclusivity and collective exhaustion is important – it means that every response fits into one and only one category. When data has this quality it is easier to understand accurately. When we started to look at some of the coding schemes that the ANES was using, we did not find mutual exclusivity and collective exhaustion. Instead, we found big piles of codes where we were actually questioning where we would put certain responses because there were 10 or 15 codes that, to us, looked indistinguishable. At the same time, you'd have wide spaces of uncovered territory – that is, a lack of coverage for large substantive areas. So we assembled expert committees on each of these issues to help us figure out a scheme where we could have a consistent relationship between the words that people say and the numbers that we produce.

We also wanted the codes to be replicable. One of the consequences of what we found is we had to delay the release of the open-ended codes for 2008. And we'd get questions. People saying: "You know, I really need the codes to compare 2000 to 1996 and so forth. When are the codes coming out?" One of the things that we had to say in a constructive and friendly way as possible is: "We don't have the documentation – there is no documentation from previous years. As a result, we could try to pretend to replicate what had been done in the past, but really I don't think we can do it. So we're trying to put forward a set of materials that you and others can use to make valid and replicable inferences across years."

Our first expert committee focused on the recall questions. A big part of that debate with that group had to do with which responses should we count as correct; which should we count as incorrect, and how should we deal with partial knowledge? To see some of the problems, consider "What's the job or political office", when you ask the question about Dick Cheney. In 2004, a lot of people say "vice-president", but some people say "anti-Christ", and some people say "chief puppeteer". And the question is what do you do with that? And so a lot of people wanted to count responses like "puppeteer" as representing a kind of knowledge. So this debate actually went on within this group, and it was the focus of how we developed a new coding scheme. It's actually not that simple to find a way to grade partial knowledge. There are a lot of slippery slopes. I'd say that we've reached a breakthrough on this topic when we actually asked the committee to do something, which you wouldn't think you would have to ask a group of users to do, which is go back and *read the question*. The question doesn't ask you to free associate about these individuals. It asks you, "What is the job or political office that the person now holds?" And so as we started to think about how to code that question, we wanted to code it with respect to this question. So, if a respondent says, "Nancy Pelosi is from California", it suggests that they know something about her, but it is not a correct answer to the question that we asked.

So if you say that Dick Cheney shot his friend in a hunting accident, you have decided not to answer the question that we asked. You may be giving us knowledge about Dick Cheney, but now we're getting, maybe, a biased sample of what you know. Because to get that type of information, we have to (a) count on you not to answer the question that we actually asked, and then (b) work on telling us some other things. We reached a consensus on grading "partial knowledge" by agreeing that "partial" should be defined with respect to the question that we actually asked. The coding framework that we now use basically first focuses on political office. Did the respondent say something about the political office? Did they identify any part of the title of this person's political office correctly? Some people have multiple offices; for example, the vice-president of the United States is also president of the senate. The person who is the speaker of the house is also a congressperson. For people like this, we'll take any mention of a title that they actually hold as correct. We also have a code for people who identify part of the title correctly, but only a fragment of it, and who also didn't say anything incorrect about the title. This allows us to give credit for someone who said part of a title, like "speaker", when they were talking about Nancy Pelosi.

The first thing we focused on is did you say anything correct or partially correct *about the political office*? Since the question asks about a political office or job, a correct answer to the question that was asked would also constitute descriptions of what this person does, of what their job is, maybe making legislation or organizing a political party, and so forth. So for each of the questions we asked, we've gone through textbooks and other sources things of that nature to identify a long list of jobs that are associated with the relevant political offices.

Unlike before where the ANES produced a simple correct or incorrect code for each question, now we have a more informative code: Do you give a completely correct answer or a partially correct answer? Did you give a complete description of a job or an incomplete description? And then finally there's "other". And "other" is anything that the person says that is not pertaining to the job or political office of that person.

Now with respect to the "other" responses, what we've done with them is we haven't made any judgments about them. What we have a code saying "other". The truth value of the various claims that are made can be difficult to assess. Suppose somebody says, "Nancy Pelosi is a liberal devil." What is the truth value of that statement? We don't have a special code for that sort of thing, because it does not provide a correct answer to the question that we asked. In effect, the new ANES code reflects: Did you name the political office, did you name the job, or have you said anything else? Now the verbatims or at least the transcripts are also publicly available. So if someone wants to create their own variable for what they think is partially correct, they can do that and the research community can argue about it. Again, we decided not to give detailed codes to "other" responses because these responses do not answer the question that was asked. It is not a good question to elicit free association. If we want to measure general recall, we need a different question.

The main attribute of the new coding scheme is that it's theoretically defensible. You could think of different ways of coding this, but now we can go back to a set of first principles and say, "If you want legitimacy and transparency, here's how you get it." It turns out that when we wrote the instructions for this question and explained it to coders at Ascribe through an iterative process, we had very high inter-coder reliability. So now we have some confidence that if we showed these instructions and this data to any coder, we would get the same numerical representations of the responses. This coding scheme also has the property of being mutually exclusive and collectively exhaustive. Moreover, scholars can use the transcripts. If they want to create a variation of what we've done, they can do that. So with that example in mind, let me just talk about our more general approach to open-ended coding.

For each of our open-ended questions, we first tried to identify a theoretical framework. The basic idea is that words don't define their own categories. These words, when analysts want to use them, are with respect to some sort of theory about how the world is organized and what concepts are important. So what we did is we really leaned on, and worked with, the expert communities to figure out whether is there a consensual theoretical framework from which we could develop a set of categories that will be analytically useful to a large and general population. We would then take that theoretical framework and develop the code frame, develop the relationship between numbers and sets of words. In every case, it was really an iterative process with the experts. For each code frame, we would take what the experts gave us and then do the math ourselves, and then bring it back to the experts and say, "Is this an implementation of what you spoke to us about?"

Next, we would go through a process called "chunking." The idea of chunking is if a person gives a longer response to a question, what we do is try and break it down to discrete utterances and discrete thoughts to allow for the fact that the person may say different things during an answer. To help us implement such chunking, we hired an external vendor. We would say, that we don't want a code to give a subjective impression of an answer. We want to produce a precise as possible a numeric representation of each thing that a person said in their answer. We developed with the vendor a set of rigorous instructions so that the coding can be implemented by a human being that is really at arm's length from us, which is a way to validate whether our code frame and whether our instructions really lead to a set of numbers that don't depend on having come from us. They depend on it coming from a logic that people can understand. So that's the sequence. To further maximize legitimacy and increase transparency, we sought to document everything. And that was our goal. We documented our correspondence with the expert communities and the vendor. Since a lot of our communication was electronic and we have a written record, we're hoping to make so much of that available so people can see how the theoretical frameworks were developed.

We have written correspondences with the vendor about implementation. We'd like people to be able to see how the code frame was implemented so they can figure out whether there is something about our practices that skew the numbers. Written documentation of all decisions, including every decision we made when we ran into trouble. Written documentation of all conversations we'd have, including conversations with the vendor when we'd deliver a draft of the instructions and then find coders would be confused by them. We'd like people who are using the data, if they want to, or people, who are developing their own coding schemes, to be able see "What did the ANES do. And if I'm getting different results than the ANES, is it because of a difference in practice?" We also wanted multiple independent assessments of our decisions. Actually producing this documentation and evaluation is incredibly time-consuming. A decision that we made was to unshackle ourselves of the time constraint. We did not rush to get the codes out. So our main constraint in doing this was really financial.

To give a sense of just how thorough we're being, if a coder had a question, what we asked the vendor to do is not have that conversation verbally; we asked them to have it in writing so that we could see it and record it. So that in the future, when the ANES is trying to do their coding scheme, they can see what types of challenges we had and perhaps decide to make the same decision from the same logic, or perhaps use our experience as the basis for choosing an improved method. To the extent that conversations like these remain hidden, it makes replication harder, makes comparability harder, and makes it harder to answer questions such as "What does that code actually mean?"

In terms of our current practices, we have much better documentation at every stage of this process. We have evaluations of various kinds at every stage of this process. We have increased procedural transparency as a result. And because we've had great expert communities and great vendors to work with, we've actually accomplished very high inter-coder reliability, statistics that we will be releasing in lots of different ways.

So the second example I want to talk about is the ANES' "most important problem" question. And here the challenge is going to be – I'm not going to go through all the various problems, but the question here is going to be: How do you develop a code frame for a question where there's very little prior guidance? Here's the question: "What do you think is the most political problem facing the United States today?" Variants of this question had been asked on the National Election Studies for many years, and it's a question in which a lot of people are interested. It's a question that is copied on other surveys including the National Election Studies of other countries.

Now when we started looking into the coding scheme for the most important problem question, we identified some challenges. One was that the coding categories seemed to vary a lot from year-to-year. There would be codes dropped, codes added. Again, because of the lack of documentation, we really couldn't figure out what the criteria were for adding or dropping categories. So, for example, there's a code for education that was used in 2000, in other words, some number would be assigned to be the code for any mention of education. And so the idea here would be if someone had mentioned financial assistance or quality of education, or so forth, they would use this code frame in 2000. Now in 2004, a new code was added to include the high cost of college. This change makes year-to-year comparisons difficult.There are lots of examples like this, but the point is these code frames were shifting from year-to-year. Now if this was "one-off" survey, where people just looked at a particular year and then left, these changes wouldn't be a problem. But one of the most common uses of the ANES is to compare from year-to-year. And so this type of categorical and definitional variation, particularly if users don't understand that it's happening, can cause big problems. It can be a cause of the fact that education is chosen by 8 percent of respondents as the most important problem in 2004, but only 4 percent in 2000.

Such differences could also be caused not the people who wrote the question or the people answering them, but by other people varying in how they apply codes to these responses weeks after the interviews are completed. When we looked at the 2004 code framework, there are 154 categories. In other words, there were 154 different buckets into which a coder could enter a representation of what people said. We saw no clear theoretical framework organizing these categories. There was no comprehensive hierarchical logical organization of what was happening. Many of the categories, as I mentioned before, were not mutually exclusive and collectively exhausted. So the notion of any sort of consistent relationship between what was said and the number – it was impossible for us to see how that could have happened in the past. We couldn't find written instructions to coders. We did not find validation statistics. I spent the better part of a year trying to find this information, like Jon Krosnick and I calling like every person in the building who could have had a record like this. We never found them. They don't exist. And what we also found is when we looked at how coders used these categories; actually 154 aren't used. What we found is most users trying to take the initial categories, convert them to a small number of mega-categories, and then using them. In other words, even the coders were confused by the scheme. It also worthwhile saying in 2004; only 14 of the 154 categories had more than five non-zero entries. So in terms of this coding scheme, its effectiveness or efficiency, I think there are some problems there. So here we convened an expert committee. We talked to them but weren't coming to a consensus about what we should do. One thing that did come up is that lots of survey organizations ask a most important problem question. So the committee, Jon Krosnick and I asked Matt Berent to conduct extensive interviews of other polling firms that made use of "most important problem" questions, and they were wonderful. Now a difference between them and us, of course, is most of them are using these questions in "one-off surveys." And so the comparability over time, I think, is not the same issue for them. So what we found when talking to them is that they have very limited code frames and did not have anywhere near 154 categories. I don't remember what the median number of categories was, but it was very small; it may be 15.

When we talked to a lot of the groups, they didn't have a set rule about how codes were added or subtracted. And, in fact, a common thing we heard is that the organization would collect the data first, and then develop the code for frame later, which, again, for that one-off study is defensible; but for us, and the demand for codes that are comparable over time, it was not plausible. What we didn't see in these survey houses was systematic analyses of inter-coder reliability. It may have happened, it just didn't come out in those conversations. So this was the challenge for us. What do we do, given that we have a code frame in the ANES that people aren't using most of and that we can't replicate.

So we wanted to start with the idea that if you want a coding scheme that has meaning for a large group of scholars, it has to be defined with respect to some theory of language and some theory of meaning. As the result of a lot of conversations, we once again developed a mutually exclusive and collectively exhaustive frame that is stable over time, can be used over and over again every four years, and is replicable. The main source for the new scheme is *federal budget categories*. In the new most important problem-coding scheme, the main eight categories reflect the main eight categories in the federal budget. Then there was a secondary scheme, and that was called "the rule of two." What we did there was look at the other main credible houses that ran the most important problem question, and document their coding decisions. So, our coding scheme is hierarchical: first use federal categories, then use the rule of two. With respect to the federal budget categories, what's interesting about them is that these categories have been stable since 1940. That is if you look at the federal budget and the main categories were all things are organized, the federal categories are actually stable; they don't change over the history of the ANES. Some subcategories change because the government evolves, but the main things, like how government labels what it does in the budget, has remained constant. Every federal governmental program and activity is listed within this framework, or at least everyone that needs funding. When you look at this budget, what you find that is that all major federal government functions are categorized.

The reason that is relevant for the National Election Studies is, of course, when you ask voters around an election, what is the most important problem facing the nation, they're usually thinking about things the government should do or shouldn't do, and many of these activities are in the budget. As a result, our coding schemes can be identical to the federal budget categories. So that's the basic framework that we work from. And now when we get down to the subcategories, these are – now we're getting to a level of details where most of our respondents don't use when answering the question. And so that's where we looked at how major polling organizations were using this data to help us fill in subcategories.

So our "rule of two" was we took any category that was used by two or more groups. And if they had used a category repeated times regularly, that would be one of our subcategories. I should just say that not everyone, when they talk about a most important problem, mentions a government activity. We had another nine categories, starting with 900, which is "other", and that would mention those types of things that end up in the "rule of two." Now one thing I can say about the advantages of this code frame; again, we have aimed for mutual exclusivity and collective exhaustion. This is derivable from a transparent logic. One of the nice things ex-post is with this code frame, we have achieved very high inter-coder reliability, and, I think, a set of categories that users are more likely to find useful. There are other ways to develop a coding scheme for this question. What we have is one that is defensible and replicable.

The last thing I want to mention that is the challenge for us is that documentation and validation of what we do, of how we turn words into numbers, it's time-consuming and expensive. I think Jon would agree. I mean some of it was frustrating, but actually a lot of it was fascinating. Personally, I learned a lot in the process from lifting up rocks. But I think the bigger picture, if we go back to the open-ended coding conference that we held and if you consider how many people had the same kinds of challenges that we did, is that greater documentation and transparency benefits us all – in terms of legitimacy and credibility, and society benefits in terms of their ability to trust what we say when we have rigorous public accounts of how we produce those numbers.

Because at the end of the day, the numbers we produce are a function of what we write, a function of our samples, a function of what respondents say, and then a function of a lot of decisions that we make that at least in the ANES's case had not been written down or shared. So that was leading to people making erroneous inferences about ANES data and then staking their own credibility in that data. But we've really tried to turn that around for the ANES and we've had a lot of support in doing that particularly from NSF. So I'd like to acknowledge that in where we ended up.

Arthur Lupia is the Hal R. Varian Professor of Political Science at the University of Michigan and research professor at its Institute for Social Research. He examines how people learn about politics and policy and how to improve science communication. His books include Uninformed: Why Citizens Know So Little About Politics and What We Can Do About It.

He has been a Guggenheim fellow, a Carnegie Fellow, is an American Association for the Advancement of Science fellow, and is an elected member of the American Academy of Arts and Sciences. His awards include the National Academy of Sciences Award for Initiatives in Research and the American Association for Public Opinion's Innovators Award. He is Chair of the National Academy of Sciences Roundtable on the Application of the Social and Behavioral Science and is Chairman of the Board of Directors for the Center for Open Science.

57

What HLT Can Do for You (and Vice Versa)

Mark Liberman

Human Language Technology, or HLT, is a term that originated in the Defense Advanced Projects Research Agency (DARPA) Speech and Language program, starting in the mid-1980s. It's more general than just smart things computers can do with text — it also includes speech input and output, and algorithms for handling text, speech, and communicative content in the context of other applications.

HLT covers many different individual technologies: document retrieval, which is what Googling used to be called; document classification; document understanding; information extraction from text; summarization; question answering; machine translation; what's come to be called sentiment analysis, or sometimes opinion mining, and so on. In the area of spoken language, there is speech recognition, where speech is transformed to text or to some representation of meaning or communicative intent; there is speech synthesis, whether from text or from some more fundamental representation of meaning; there's speech activity detection, and language recognition, and speaker recognition or verification, and determination who spoke when in multi-speaker recordings And there is spoken document retrieval; information extraction from speech; question answering; human–computer interaction via speech, a la Siri; speech to speech translation; and so on again.

M. Liberman (✉)
University of Pennsylvania, Philadelphia, United States
e-mail: myl@cis.upenn.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_57

50 years ago, most people believed that the path to effective human language technology would necessarily involve a process of layered understanding, based on a form of artificial intelligence with the general ability to understand what people say, what they write, what they mean, and why they mean it. This intelligence would include not only broad and deep knowledge, but also the general ability to reason about about its observations based on its background knowledge, and to plan and execute its own communicative acts so as to advance the goals of its developers or its users.

This view of HLTreflected a more general view of artificial intelligence as applied logic.

The field's recent successes have emerged from a different view: artificial intelligence is applied statistics.

And the recent successes of human language technology share this perspective. A typical development process starts with a very large collection of training material, whether speech or text or video, supplied with human annotations: transcription, classification, translation, tagging of regions corresponding to interesting entities such as people, places, times, dollar amounts, governments, genes, protein, disease states, or whatever. And then there's a statistical model of some kind. It might be a very simple model, or a complex one, or a model in the form of "neural net", but in all cases the goal is the same: to learn from the annotated training data how to reproduce similar annotations on new inputs.

In a sense, you can see this simply as a form of regression. There are some independent variables and some dependent variables; there's a statistical model that learns predictive parameters; and in the end you run the model to map available inputs to desired outputs. Inside the machine there's no explicit knowledge and no logical inference process.

Attempted applications of these ideas in survey methodology go back at least several decades. So a natural question to ask is why there hasn't been more progress, and broader application of the techniques — why aren't social scientists using these methods every day? In the first place, the techniques haven't really worked well enough until quite recently. In the second place, the general approach depends on large amounts of consistently annotated and shareable training data, and this is both expensive and potentially problematic in terms of privacy and confidentiality. And finally, there's a significant cultural gap between social scientists interested in survey methodology and engineers skilled in human language technology.

When I first met Jon Krosnick, a leading survey researcher at Stanford, I learned that he didn't know Dan Jurafsky, a computational linguist who is also at Stanford. But it's natural to expect that this cultural gap will narrow.

As HLT become more effective and as implementations become more accessible, such methods diffuse across disciplines. The extent and speed of the diffusion will depend on the obvious cost-benefit calculation: how well the techniques work, how timeconsuming and expensive it is to find out how well they work, and what cost savings or outcome improvements result even from successful applications.

The answer to these questions, unfortunately, is that it depends on the details of each case. Out-of-the-box HLT solutions will sometimes solve a survey research problem easily, but sometimes the results will be disappointing. The difficulty and expense of exploring such methods depends not on the nature of problem but also on things like how hard it is to clean up the data for computer analysis, what knowledge and skills the survey researchers have, what HLT collaborators are interested in participating, and so on. And as long as the adaptation of HLT techniques to survey data remains a novel and case-by-case task for experts, the development costs are likely to be greater than the cost of the semi-skilled labor that can be displaced. On the other hand, there may still be value in the methodological experience and in the possible development of iproved techniques for predicting opinions. And at some point, HLT methods will become sufficiently routinized to be cost-effective even for solving standard problems.

I wanted to say a little bit about our experience over 25 or 30 years of research in human language technology and on human coding, and one point is that natural annotation is extremely inconsistent. If you give annotators a few examples, or a simple definition, and turn them loose, what you get is very poor agreement. Now, I've done work on the identification of entities, which are things like people, places, organizations, or, in a biomedical domain, genes, gene products, proteins, disease states, organisms, and so on. So, if you were to take a bunch of Ph.D. geneticists, give them scientific papers that are about areas in their specialization, ask them if they can determine when genes are mentioned in those papers, and they'll look at you like you're an idiot because of how simple the task would be. But if you take two of them, put them in separate rooms, and ask them to do that task, and then look at how well they agree on the output. If they agree 50 percent of the time, you're very lucky.

It's worse for what we call normalized entities, that is, where you don't just say, "All right, this is the name of a gene," but you want to know which gene is it the name of. This is – not just, "This is the name of a politician," but "which politician is it the name of?" It is worse yet, for relations among entities. This is because human generalization from examples to principles is variable, and human application of principles is equally variable. And the natural language context raises a bunch of additional hard questions. The result is that the “gold standard,” as we call it, is not naturally very golden. The resulting learning metrics are noisy, and F-score, which is the harmonic mean of precision and recall of.3 or.5, it’s just not a very attractive goal. If you tell people that you can agree with their intuitions 30 percent of the time, they’re not very impressed, even if their intuitions and their neighbor’s intuitions only agree 30 percent of the time. So, the traditional solution is an iterative refinement of guidelines of exactly the kind that we heard about before lunch, try some annotation, compare and contrast, adjudicate and generalize, go back to step one and repeat, at least until inter-annotator agreement is adequate.

What we usually do is about ten percent blind dual annotation throughout the task. That is about one instance – one thing to be annotated in ten is actually being annotated by someone else, and you don’t know which ones those are. The process of convergence is slow. We heard that it took them 4 years. We can sometimes get it down to months, but that’s usually because DARPA insists that we deliver the data, not so much because we’re really convinced that we solved the problem. Now, the result can be quite high inter-annotator agreement. For things like entities, we typically get over 90 percent. But this is based on a complex accretion of what is really quite like the common law. It’s like you have a simple statute that says what you’re allowed to do and what you’re not allowed to do, what’s legal and what’s not legal. Then when you try to apply that to cases, it gets complicated. And if you want to do it in a consistent way, the only way that seems to work is to accumulate a long list of semi-generalized particular examples. Our guidelines for typical annotation tasks can sometimes be hundreds of pages. They’re typically at least 50. They’re slow to develop and hard to learn, but this is more consistent than natural annotation. It’s the only way that we know to produce high-quality inputs to the machine learning process that could be used to automate, as we learned earlier.

Okay, now getting back to the potential for interaction, the simplest reason to want to do it is that there’s a not inconsiderable overlap between what surveys do and what human language technology or human language technology researchers do. For example, open-ended response classification is effectively equivalent to spoken document – to document classification, whether written or spoken. And this was the overlap that was featured in the 1998 book that I mentioned earlier. I think that there’s a larger and maybe better set of reasons, such as saving money and getting things done faster and to a higher standard is a very important goal, and so I don’t mean in any way to minimize it. But as an academic researcher, I’m always interested in the things that we haven't done yet. And there's a large area of what surveys could do, and some, I think, do already, that also overlaps with interests and applications in human language technology. Probably the most obvious one, and the most widely explored these days is sentiment analysis or opinion mining and perhaps its generalization to things like trying to figure out how strongly respondents believe the answers that they give you, and therefore, how likely they are to change their minds, that sort of thing.

I thought I would give you an example of a fun, easy case that we did in a class last semester. So, I got 84,000 reviews. One review is from *Wine Enthusiast Magazine* online, which has about 4 million lexical tokens corresponding to around 30,000 distinct words, of which about 5,500 occurred at least 20 times. Each of those reviews comes with a rating, which is a number between 80 and 100, which is supposed to be how good the person who created the wine-tasting notes thought the wine was. So, we did a regression of the ratings on the words, limiting it to the 5,500 words that occurred at least 20 times, and the multiple R number for the regression was.86. So, about 75 percent of the variance in the ratings is accounted for by this simple bag of words model. And you can sort of see why that is, if you look at the 20 words with the biggest regression coefficients, "incredibly," "gorgeous," "superb," "brilliant," "beautiful," "wonderful," "massive," "wonderfully," beautifully," "opulent," "delicious," "impressive," "powerful," "excellent," "luscious," "huge," "power," "intense," "long," and "richness." And similarly, if you look at the 20 words with the smallest coefficients, "heavy," "lean," "sour," "modest," "everyday," "rough," "rustic," "simple," "short," "dimensional," which presumably comes from one dimensional, I would guess, "thin," "vegetal," "watery," "lack," "lacking," "dull," "lacks," "harsh," "sugary," and the one that's interesting to me, the single word with the most negative regressions coefficient was "acceptable." That's like the worst possible thing for a wine to be apparently.

I think there's an even larger and even better set of reasons, which is that there are things that researchers in human language technology would like to do that are somewhat different from but also overlap with possible future survey research, and most importantly would benefit from access to the kind of datasets that survey researchers have.

So, what will the future bring? As I said before, the out of the box solutions, some of them, I think, may work now. As time goes on, more and more of them will work, but in my opinion, the biggest potential benefit is in collaborations where both sides are breaking new ground. And I'd like to say a little bit about how that could work. And the best way that I can explain this to you, I think, is by telling a story, and this is my attempt at intercultural communication from the other side. I want to give you some insight into an interesting and perhaps, from your point of view, slightly peculiar feature of researchers in this area, and the culture of that field.

So, the story begins in the 1960s, with some interventions by John Pierce, who was an executive at Bell Labs, who invented the word "transistor." He supervised the group that invented the transistor, and he also supervised development of the first communication satellite. It was his organization in Bell Labs that I got a job with when I got out of graduate school. So, in 1966, he chaired a committee, assigned by the National Academy of Sciences, to report on research in machine translation in automated translation of text. And in 1969, he wrote a letter to the journal of the acoustical society of America. The Automatic Language Processing Advisory Committee (ALPAC) report, which was the machine translation report, was diplomatic. In 1966, machine translation was not very good, and ALPAC said that the committee couldn't judge what the total expenditure for research and development should be; however, it should be spent hard-headedly toward important, realistic, and relatively short-range goals. And in fact, U.S. government funding for machine translation research went essentially to zero for more than 20 years. The committee felt that science should precede engineering in such cases. We see that the computer has opened up to linguists a host of challenges, partial insights, and potentialities. We believe that these can be aptly compared with the challenges, problems, and insights of particle physics. Certainly, language is second to no phenomenon in importance, and the tools of computational linguistics are considerably less costly than the multibillion volt accelerators of particle physics." So, this was 1966, and it's – we don't yet have the quantum field theory of linguistics, but I guess they don't have a quantum field theory in physics either.

Pierce's views about automatic speech recognition were similar, but his letter was his own personal composition. It wasn't a committee report, and so it was substantially blunter. He said, "A general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of English, a knowledge of language comparable to those of a native speaker of English."

Most recognizers, by which he meant researchers working on recognition, behave not like scientists, but like mad inventors or untrustworthy engineers. "The typical recognizer gets it into his head that he can solve the problem. The basis for this is either individual inspiration, or acceptance of untested rules, the untrustworthy engineer approach. The typical recognizer builds their programs an elaborate system that either does very little, or flops in an obscure way. A lot of money and time are spent. No simple, clear, sure knowledge is gained. The work has been an experience, not an experiment." And then he went on to say, "We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon," which they did after – anyway. "One doesn't attract thoughtlessly-given dollars by means of schemes for cutting the cost of soap by ten percent. To sell suckers, one uses deceit and offers glamour." Now, this quote is pretty – he's not mincing any words here: "It is clear that glamour and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect."

So, the first idea at the government level was, "Okay, there's this new notion that's come along, artificial intelligence, and very smart people at MIT, and Stanford, and CMU, and various other places have come up with all of these brilliant ideas about applying logic and proof theory and stuff like that to solving problems, so, let's try it." So, there was a DARPA speech understanding research project, 1972–1975, which tried out those ideas, and it was canceled early on the grounds that it didn't seem to be getting anywhere. Basically, the Piercians in the government funding agencies won. And the second idea was just to give up. So, between 1975 and 1986, there was no U.S. research funding in these areas at all. There was a fair amount of work in Europe; there was a certain amount of work going on in Japan; some work going on in U.S. companies, though relatively little, but nothing really funded by the U.S. government – hardly anything.

And Pierce was not the only person with this jaundiced view of this area. By the mid-1980s, most informed American research managers were skeptical about the prospects. But there were also many people who thought that some kind of human language technology was needed, and in principle, ought to be feasible, at least to some level of performance. So, in 1985, the question was posed, "Should the Defense Advanced Research Projects Agency restart human language technology?" Charles Wayne, the DARPA program manager, had an idea, which was to design a speech recognition program that would protect against glamour and deceit, because there's a well-defined objective evaluation metric that's applied by a neutral third-party agency – in this case, the National Institute of Standards and Technologies. It would be applied on shared datasets, and this would ensure that simple, clear knowledge is gained, because the participants have to reveal their methods to the sponsor and to one another at the time that the evaluation results are revealed. And in 1986, in the United States, there couldn't have been any other kind of automatic speech research program that could have gotten the funding.

So, in 1985, Dave Pallett at the National Institute of Science and Technology (NIST) wrote a paper called "Performance Assessment of Automatic Speech Recognizers," and I won't read this at length, but the basic idea was that replicability is key. That you have to have published data; you have to have explicit, published, quantitative evaluation criteria. You put the data into the recognizer; you apply the evaluation program; you get out a number.

Have some other recognizer put in the same data, look at the output, compare it through the published evaluation criterion, which is generally published in the form of a program, as well as a description, you get out a number. So, this resulted in what came to be called the "common task structure." There was a detailed evaluation plan, often quite a long document that was developed in consultation with researchers and was published as the first step in the project. There was automatic evaluation software, producing a quantitative evaluation that was written and maintained by NIST. It was also published at the start of the project. And, critically, there was shared data, training, and dev. test or development test data that was published at the start of the project, and then the evaluation test data would be withheld for periodic public evaluations.

Now, not everybody liked this, to say the least. A lot of the Piercians were skeptical. The basic idea was, "It doesn't matter what you measure, you can't turn water into gasoline." And a lot of researchers were disgruntled. One well-known researcher in this area told me at the time, "It's like being in first grade again. You're told exactly what to do, and then you're tested over and over." You know, it's like "No Child Left Behind" for speech technology research. But it worked. Why? The most obvious reason was that it allowed funding to start; it turned the spigot on. It also allowed funding to continue, because the funders could measure progress over time and point to their superiors at the graphs that showed that things were genuinely and believably getting better, even though the technology was still not ready for application.

Less obviously, it allowed project internal hill climbing, because the evaluation metrics were automatic, and the evaluation code was public, and so, the obvious way of working was to develop a new algorithm, or, more likely, a slight tweak on an existing algorithm, run it on the evaluation test set, see if you get better. If you do, you keep the innovation. If it doesn't, you throw it away. So, you're doing, in effect, algorithmic hill climbing, kind of like, you know, improving the internal combustion engine over the course of 50 years or so. And the people who were complaining about being tested every six months actually started testing themselves several times a day in order to improve.

Perhaps the least obvious thing in advance, but the most obvious in retrospect, was that it created a culture because researchers shared methods and results, on shared data, with a common metric. There came to be a large group of researchers, in the U.S. and around the world, who really spoke a common language because they had common problems and common solutions. In fact, participation in this culture became so valuable, that many research groups began to join without funding. And in fact, after a few years, DARPA realized that they could get research done in areas of interest to them without funding it, just by creating one of these common task evaluations and inviting people to come compete for gold stars. And people did. It created a positive feedback loop, when everybody's program has to interpret the same evidence, ambiguity resolution becomes a sort of gambling game, and this rewards the use of statistical methods. And given the nature of speech and language, statistical methods need the largest possible training set, which reinforces the value of shared data, and these iterated train-and-test cycles allowed you to do the kind of algorithmic hill climbing that I talked about.

Now, over the past 25 years, this method has been applied to lots and lots of other problems: machine translation, speaker identification, language identification, parsing, sense disambiguation, information retrieval, information extraction, summarization, question answering, optical character recognition, sentiment analysis, and so on.

The general experience is that error rates – as long as people keep working on the problem, error rates decline by a fixed percentage every year. That's a relative percentage, obviously, not an absolute percentage, or you would soon get error rates below zero. But you get a kind of asymptotic decay to some error rate that's determined basically by the noise and level of noise in the data. Progress usually comes from many small improvements, and a change of one percent in performance can be a reason to break out the champagne. The conferences in this area are either heartening or depressing, depending on your attitude, because you hear talk after talk after talk in which somebody says, "Well, we did this incredibly elaborate experiment, and we improved performance by three-quarters of a percent." And then there's another talk, "And we did this other elaborate thing, and we improved performance by half a percent," and so on.

And an improvement of one percent is a big deal. So, you could view that as being depressing. I mean where is the basic scientific advance? But of course, if you've got 50 improvements by half a percent, that adds up over time, as long as at least they're semi-independent and don't cancel one another out.

The shared data plays a crucial role. It can be reused in unexpected ways, and glamour and deceit have been avoided, and a sort of self-sustaining process was started. If you do a Google Scholar search for some of the datasets that have been published, in association with these evaluations, you get counts like 12,000. This is 335 in the past year for that one, 6,192, and so on. The number of tracks that the number of kinds of evaluation that NIST runs, or has run over the last 25 years, is numbered in the dozens. So, this happens to be one about text analysis. So, there's question answering; recognizing textual entailment; summarization; and what they call "knowledge base population," which is sort of learning facts from bodies of text.

The same kind of pattern now exists in lots of non-governmental entities. So, the CoNLL is the association for computing machineries. It's a special interest group on natural language learning. At their annual meetings, since 1999, they've included a shared task in which training and the organizers provide test data, which allows participating systems to be evaluated and compared in a systematic way. This is a technical society. This is not a government agency, it is entirely volunteer based; people are doing it because they want to and not because someone is making them.

My own organization, the Linguistic Data Consortium, was founded in the early days of this process with seed money from DARPA, to help create and maintain and distribute the data involved. And over our 20 years of existence, we've distributed more than 90,000 copies, of more than 1,300 titles, to more than 3,200 organizations, in 70 odd companies. About half of the titles are common task datasets developed for these technology evaluation programs, and we add about 30 titles to our catalog every year. Now, in 1983, the first conference on applied natural language processing had 34 presentations, none of which used a published dataset, none of which used a formal evaluation metric. That doesn't mean they weren't interesting. Here are a couple of examples. Wendy Lehnert and Steven Schwartz, "Natural Language Processing System for Oil Exploration." They described the problem in system architecture; they give examples of queries and responses. Larry Reeker et al., "Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions." So, this is an attempt to do information extraction from the chemical literature.

In 2010, there were 274 presentations at a comparable meeting, and essentially every single one of them used published data and published evaluation metrics from one or another of these evaluation series. And the few that didn't dealt with a new dataset creation or new evaluation metrics. The point is that this is more or less the state of this field now. This is how people work in the area of human language technology exploration. That is, the research is almost inevitably on published sharable datasets using – by reference to published evaluation metrics, so that replication in reference to previous or subsequent algorithms is available.

There are some exceptions, especially now, because there are some companies such as Google, Facebook, and Twitter, that have incredible bodies of data internally that they can't publish for privacy reasons, or for reasons of confidentiality, or corporate advantage. And so, some of the people working in those areas do publish results on datasets that – where things can't really be replicated. You sort of have to take their word for it. But that's a minority; that remains a minority of a situation. So, the culture of social science and the survey area is somewhat different, but maybe not all that different. I mean when I look over the history of the ANES survey, for example, which I've done recently, I'm certainly struck by the fact that many of the same values are kind of implicit in its design and its evolution.

Sharing data and problems lowers cost and barriers to entry. It creates intellectual communities, in which that body of data certainly has done. It speeds up replication and extension, and it guards against most forms of glamour and deceit, as well as simple confusion. The thing that hasn't quite happened, and someone this morning said that it would be helpful to some kinds of research, for people involved in survey work to loosen up a little bit with respect to sharing data with other researchers. And I would like to strongly underline and support that view. I know that there are serious problems about privacy and confidentiality and Institutional Review Board (IRB) protocols and so on, but those can generally be negotiated; those can be dealt with where there's a will. The example that I'd like to give, as a point of reference, that I invite you to look into, is something called the Alzheimer's Disease Neuroimaging Initiative. This is something that was started by the National Institutes of Mental Health, and it has enrolled a very large number of patients funneled through at least 30 clinics. For each of these patients, there's detailed demographic information and medical histories. There are Functional Magnetic Resonance Imaging scans, and I think structural MRI scans as well, in some cases positron emission tomography scans. There are blood tests and cerebrospinal fluid, assays. There are cognitive tests, and there are physicians' notes from the examination. Obviously, the patients are anonymized, that is, you don't know their name and address. They're given some kind of key. But all of the rest of that information is, in effect, available to anyone. Any researcher could get it. Now, it's not up on the Internet for anybody to download. You have to send them a note and tell them what you want to do with it and sort of establish that you're a bona fide researcher. And you have to sign something that imposes some protections on what you're going to do with the data. But once you've done that, and it's not hard to do, you can get all of this stuff.

Why did they do that? Because the current state of the art, as I understand it, what I'm told, is that if I were to go to see a neurologist today, exhibiting symptoms of "mild cognitive impairment," as they call it, that is, my short-term memory is beginning to decay; my ability to remember names is even worse than it ever was, and some other things are leading me to worry a little bit, so she does and examination. She does a brain scan; he takes blood; she takes cerebrospinal fluid; he puts me through some cognitive tests, and he can tell me where I am. She can put me in a diagnostic space. She can tell me nothing about what to expect. She can tell me, "In five years, you could be a vegetable. In five years, you could be more or less just like you are, maybe a little bit worse." There's nothing in that body of data that enables people to do a reliable projection any significant distance into the future.

So, what they would like to do is to encourage all sorts of people: computer scientists, statisticians, other biomedical researchers, anybody with any reasonable standing to take a flying leap at this task, to come in and see if they can do better. Because they've got longitudinal data from these patients, and more accumulating all the time. Now, if they can do that with complete medical records for thousands of people, I think it ought to be possible to do something similar with opinion survey results, or lifestyle survey results for similarly sized collections. Obviously, it requires appropriate IRB protocols to be filed. It requires appropriate informed consent on the part of the participants. It requires some kind of procedure for sort of taking care of how the data is handed out and what's done with it after it's handed out. But it seems to me that it should both legally, and from the point of view of government policy, and also ethically be quite feasible.

Mark Liberman is an American linguist. He has a dual appointment at the University of Pennsylvania, as Christopher H. Browne Distinguished Professor of Linguistics, and as a professor in the Department of Computer and Information Sciences. He is the founder and director of the Linguistic Data Consortium. Liberman's main research interests lie in phonetics, prosody, and other aspects of speech communication. His early research established the linguistic subfield of metrical phonology. Much of his current research is conducted through computational analyses of linguistic corpora.

58

Confidentiality, Privacy, and Anonymity

Roger Tourangeau

This chapter examines the issues of privacy, confidentiality, and anonymity in surveys. It also concerns the issues involved in asking sensitive questions and the evidence regarding the self-administration of surveys and its utility for collecting information about sensitive topics. Finally, it addresses methods for improving reporting sensitive information.

According to Singer (e.g., Singer et al. 1993), privacy refers to unwillingness to reveal information at all. Respondents who say (or feel) that "It's none of your business," are exhibiting privacy concerns. Respondents who are willing to reveal the information to a researcher, but worry that it will fall into the wrong hands are concerned about *confidentiality*. In the survey setting, respondents may worry that someone else in the household may overhear what they report in an interview or that another federal agency or a corporation would learn learn what they said.

The work of Singer and her colleagues showed that a substantial portion of the American population believed that federal agencies shared information and the information reported in the decennial could be used by other agencies. For example, one of the confidentiality items asked "Do the police and FBI use the Census to keep track of troublemakers?" Another asked "Is the Census is an invasion of privacy?" the latter is a classic privacy item.

R. Tourangeau (✉)
Westat, Rockville, MD, USA
e-mail: RogerTourangeau@westat.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_58

It turns out that a lot of interviews aren't done in private so there is a genuine risk that other household members will overhear information revealed in a survey interview. Mneimneh and her colleagues (Mneimneh et al. 2015) examined data from the World Mental Health interviews, which are conducted in many countries around the world. And, for example, in Japan, about 12.6 percent of the interviews are done with somebody else present; the country she and her colleagues studied with the least interview privacy was India, where 70 percent of the interviews were done with another person present. These surveys are about mental health symptoms; for example, one question asked "Have you ever tried to commit suicide?" So, this is kind of sensitive stuff, and you wouldn't necessarily want to have somebody else listening in. Mneimneh and her colleagues identified several factors that help determine whether an interview is done in private. An obvious factor is household size. People living alone are much more likely to be interviewed alone.

Still, even controlling for differences in household size, there were large differences across countries in the proportion of interviews done in privacy. These cross-country differences partly reflect differences in cultural norms regarding privacy. In some countries, respecting people's privacy is a value; in other countries, sharing with other people – for example, collectivist cultures – privacy is not such a value. The third factor that Mneimneh and her colleagues point to is the interviewer; they find large variance components in the proportion of interviews done privately. Some interviewers seem to understand the importance of privacy, whereas others do not.

What difference does it make whether an interview is conducted in private? A model by Aquilino, Wright, and Supple (2000) identifies two main factors that determine whether the presence of third party during an interview reduces respondent truthfulness. If the bystander already knows the information, it may increase the chances the respondent will tell the truth, since it would be embarrassing to be overheard telling a lie.

But if the other person does not already know the information, then the respondent may not want third party to find out. In analysis of a national survey on drug use, Brittingham and her colleagues (Brittingham et al. 1998) found that parental presence during the interview lowered the percentage of teens reporting that they smoke. In a meta-analysis, Ting Yan and I (Tourangeau and Yan 2007) analyzed studies of the impact of third-party presence, and found a large effect of parental presence when young people are interviewed. More generally, Mneimneh and her colleagues found that there are signs of increased social desirability bias across cultures when third parties are present during the interview.

A step that's sometimes taken in order to convey a greater sense of privacy and confidentiality is not identifying the respondents – that is, collecting data anonymously. It may be very hard for face-to-face surveys to do this convincingly. If an interviewer comes to a house, then clearly he or she has the respondent's address and it may be difficult to convince the respondent that the data are anonymous. Still, a few studies try to do carry out anonymous data collection. Monitoring the Future is a study of high school seniors about drug use, and the design is to send questionnaires to schools; the questionnaires are distributed and completed in classrooms, collecting no identifying information.

There is some evidence that anonymity has some disadvantages. The idea is that if the respondents are not held accountable for the quality of the data, then they will take cognitive shortcuts or exhibit other forms of satisficing (Lelkes et al. 2012).

The three topics – privacy, confidentiality, anonymity – are often talked about in conjunction with another variable – collecting sensitive information. It may be more important that the data be seen as confidential, the data collection be seen as anonymous, and the data collection be carried out in private when there's something about the questions that may bother the respondents.

There are three somewhat distinct meanings that question sensitivity is taken to have. Sometimes questions are seen as inherently offensive. There is a distinction between a sensitive question and a sensitive answer. If a federal survey asked respondents about their religion, say, many respondents would doubtless object to that. They actually contemplated putting a religion item on the decennial Census in the 1950s, but decided against it.

Thus, one type of sensitivity involves inherently offensive questions. There are, in addition, questions that raise concerns about disclosure to third parties, such as other family members or other government agencies. Survey researchers all worry about disclosure risks – that some analyst will correctly infer who a specific respondent was and what they reported in the survey. Although to the best of my knowledge, that's never happened, many datasets are altered and geographical information is removed to ensure that identification of individual respondents does not happen in the future.

Apart from questions that are inherently offensive and those that raise concerns about disclosure to third parties, there is a third type of sensitive questions – those that raise social desirability concerns. With these questions, there is a socially approved answer and a socially disapproved answer. There are two conceptions of social desirability bias. The older, psychological conception, dating back to Crowne and Marlowe (1964), is that there's a trait involved. Some people are very worried about how they're perceived, and they consistently present themselves in a very positive light. Most survey researchers take a more social psychological vantage point on this, believing that an item is more sensitive for those in the socially undesirable category.

The Crowne–Marlowe items are designed to measure social desirability. They consist of statements that are true of hardly anybody, (for example, "Before voting, I thoroughly investigate the qualification of all the candidates"). However, the problem with these items is acquiescence – the socially desirable answers are "true" to most of the items.

When the topic of the survey is sensitive, researcher worry about unit non-response (sample members not doing the survey at all), item non-response (respondents skipping offensive or embarrassing questions), or reporting errors. Since the 1970s, several methods have been applied to the issue of misreporting on sensitive questions. The early work of Bradburn and his colleagues (1979) found that self-administration and open items reduced misreporting.

Later studies have suggested that hte randomized response technique is useful. Turner, Lessler, and Devore (1992) conducted a study, where more than twice as many people admitted in a paper self-administered questionnaire that they'd used cocaine in the past month than in a paper-and-pencil interview.

More recent studies also examine these issues. Several studies compare self-administration and interviewer administration when records data are available to determine which is more accurate. Kreuter, Presser, and Tourangeau (2008) compared three modes of data collection. We conducted a survey of University of Maryland alumni and the registrar agreed to give us access to their academic records. We compared things the false negative rate in reporting a bad GPA, for example, in the three modes. When an interviewer collected the data, they were the least accurate. When respondents had, in fact, gotten D's or F's only 20 percent in the self-administered Web condition denied it versus about 33 percent in the interviewer administered condition.

In a study by Tourangeau, Groves, and Redline (2010), we examined reports about voting, a socially desirable behavior. Respondents were telephoned or mailed a questionnaire about whether they had voted in the last three elections. We found an effect for mail administration relative to telephone administration. We examined the percentage of non-voters according to their records; non-voters over-reported significantly more in the telephone survey than in the mail survey. Still, non-voters did not report perfectly on the mail survey either.

Apart from self-administration, some researchers advocate the use of the randomized response technique for sensitive questions. Despite its popularity among statisticians, no production survey uses it and it is not clear whether any survey will ever adopt it.

A related approach is the item count technique. In the item count technique, respondents get a list of four or five things and are asked how many are true of them. Holbrook and Krosnick did four experiments comparing these techniques; the conclusion seems to be that the item count does not add to self-administration in improving reporting of sensitive information.

Tourangeau and Yan's meta-analysis of studies on the item count technique found mixed results – sometimes it seems to work and sometimes it does not. And sometimes it just gives totally implausible results, percentages that are negative or greater than one hundred. We need to invent new methods that are actually workable, that go beyond self-administration.

Another method is the bogus pipeline. This is a device that the respondent thinks can detect a lie.

For example, a A study by Bauman and Dent asked students whether they smoked; the researchers also took breath samples, from which they could determine whether the respondent had smoked recently. Some respondents were told that the breath sample would reveal whether they had smoked recently; the others were not. Apart from showing the impact of warning respondent that misreporting would be detected, the results of the Bauman and Dent (1982) study were impressive in that all the errors were in the expected direction. Students who do smoke deny it, but those who don't smoke report accurately.

There is some work on the relationship between having voted and survey participation. Tourangeau, Groves, and Redline (2010) presented some results on this issue. Some 47.6 percent of their sample had voted in 2004, according to Aristotle. Of the *respondents*, 11 percent more had voted; there were 11 percent more voters among respondents than in the original sample. Among the respondents, 80 percent said they voted, a measurement bias of about 21 percent. The study by Tourangeau, Groves, and Redline indicates that measurement error contributes more than non-response error to the total error in the estimate, and both contribute more than sampling error,

One conclusion from this work is that self-administration reduces reporting error, but hardly eliminates it. Even when the questions were self-administered, half the non-voters said they voted (Tourangeau, Groves, and Redline, 2010). There are lots of clever methodsin the literature, like the randomized response and item count techniques, but it's not clear whether they actually add much to self-administration. They are rarely used in practice because they do not provide an individual-level measure of the variable of interest.

Thus, we need to devise new methods for collecting accurate information on sensitive topics. It seems to me they fall under three headings: causes, consequences, and fixes.

Under the causes heading I do not think we understand this process very well. I think the process is, at best, semiconscious People may be so adept at dodging embarrassing questions in everyday life, they do it unthinkingly. For this reason, lying seems too strong a word for what happens in surveys. In the social psychology literature on lying, most lying is of the "white lie" type.

We are adept at avoiding embarrassing questions. Still, it would be useful to understand more thoroughly how people fend off such questions in surveys.

On the causes side, it is not very clear when people are inclined to misreport in surveys. Kreuter, Presser, and Tourangeau (2008) found that the level of sensitivity of the question related to the level of misreporting. It would be useful to understand what questions people find to be sensitive.

In terms of consequences, I would advocate more studies like Tourangeau, Groves, and Redline (2010) that provide some estimate of the relative magnitude of the errors from different sources, such as non-response and measurement. Without a large body of such studies, it is not clear which error sources survey methodologists should be worrying about.

References and Further Reading

Aquilino, W. S., Wright, D. L., & Supple, A. J. 2000. "Response effects due to bystander presence in CASI and paper-and-pencil surveys of drug use and alcohol use," *Substance Use and Misuse*, 35, 845–867.

Bauman, K., & Dent, C. (1982). Influence of an objective measure on self-reports of behavior. *Journal of Applied Psychology*, 67, 623–628.

Bradburn, N. M., Sudman, S., & Associates. 1979. *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.

Brittingham, A., Tourangeau, R., & Kay, W. (1998). "Reports of smoking in a national survey: Self and proxy reports in self- and interviewer-administered questionnaires," *Annals of Epidemiology*, 8, 393–401.

Crowne, D., & Marlowe, D. 1964. The approval motive. New York: John Wiley.

Kreuter, F., Presser, S., & Tourangeau, R. 2008. "Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity," *Public Opinion Quarterly*, 72, 847–865.

Mneimneh, Z. M., Tourangeau, R., Pennell, B.-E., Heeringa, S. E., & Elliott, M. R. 2015. "Cultural variations in the effect of interview privacy and the need for social conformity on reporting sensitive information." *Journal of Official Statistics*, 31, 673–697.

Singer, E., Mathiowetz, N., & Couper, M. 1993. "The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. census," *Public Opinion Quarterly*, 57, 465–482.

Tourangeau, R., Groves, R. M., & Redline, C. D. 2010. "Sensitive topics and reluctant respondents: Demonstrating a link between nonresponse bias and measurement error," *Public Opinion Quarterly*, 74, 413–432.

Tourangeau, R., & Yan, T. 2007. "Sensitive questions in surveys," *Psychological Bulletin*, 133, 859–883.

Turner, C. F., Lessler, J. T.& Devore, J. 1992. "Effects of mode of administration and wording on reporting of drug use." In C. Turner, J. Lessler, & J. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 177–220). Rockville, Md.: National Institute on Drug Abuse.

Roger Tourangeau is a Vice President in the Statistics Group and co-director of Westat's Survey Methods Group. Tourangeau is known for his research on survey methods, especially on different modes of data collection and on the cognitive processes underlying survey responses. He is the lead author of The Psychology of Survey Response, which received the 2006 AAPOR Book Award, and he was given the 2002 Helen Dinerman Award, the highest honor of the World Association for Public Opinion Research for his work on cognitive aspects of survey methodology. He is also the lead author of The Science of Web Surveys. Before coming to Westat, he worked at NORC, the Gallup Organization, and the University of Michigan; while at the University of Michigan, he directed the Joint Program in Survey Methodology for nine years. He has a Ph.D. from Yale University and is Fellow of the American Statistical Association.

59

Panel Attrition

Randall Olsen

I think it's fair to say, for the last 25 years, longitudinal surveys have been seen as an important part of the future of social science research. The National Science Foundation (NSF) supports one clearly longitudinal study, the Panel Study of Income Dynamics (PSID), and the other two main studies that it has funded over a great many years are the American National Election Studies (ANES) and the General Social Survey (GSS). The latter two surveys have very significant panel attributes. In addition, a lot of our quintessential cross-sectional surveys aren't entirely cross-sectional either; the Decennial Census data, courtesy of Steven Ruggles and the Integrated Public Use Microdata Series (IPUMS) project has been turned into a panel survey of sorts. The American Community Survey (ACS) has, essentially, been used as a screener for other studies, such as the NSF Scientists and Engineers Statistical Data System effort. And so, the ACS might morph into a panel. As others have noted, the Current Population Survey with its rotating structure allows you to actually make this a high-frequency panel, although I think it's fair to say the Census Bureau does its best to prevent that from happening.

So, the point here is that people and organizations are always in motion and it's unusual when a cross-sectional survey doesn't ask backward-looking questions. So, when we look at the intertemporal aspect of social science data, we really have two choices. We either collect this intertemporal data

R. Olsen (✉)
Ohio State University, Columbus, United States
e-mail: olsen.6@osu.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_59

prospectively or retrospectively. If you are collecting data that covers many years, you do one or the other, you pick your poison. You either have problems with retrospective accuracy or you have attrition. In one of my papers, I show the panel aspects of the ANES – which are fairly considerable – and there's the GSS that has had, historically, many re-interviews. And then, of late, the GSS has started to go to a panel design as well.

So, let's consider a convenient sample of longitudinal surveys around the world. I have to acknowledge the assistance I've received from people at RTI and the University of Michigan Institute for Social Research (ISR), the University of Essex and, of course, colleagues at the National Opinion Research Center (NORC) and U.S. Census Bureau, who for years, have collected all these data. The Educational Longitudinal Survey (ELS), at its intake, was able to complete interviews on 88 percent of the selected students, which is very high, but, of course, it's a school frame. I don't think this number accounts for non-cooperating schools. I'm not sure to what extent schools opt out of the ELS, but insofar as that happens, this 88 percent is somewhat overstated. At wave 2, the ELS completed interviews on 91 percent of the wave 1 respondents; and at wave 3, 87 percent of the wave 1 respondents. So, the attrition rate from one to two and then, from two to three, fell from nine to four percent. In addition, in wave three, the ELS returned to wave two non-respondents and for selected domains, recovered data missed in that wave two interview that was skipped. That's an important point and I'll develop it more.

In terms of the attrition problem, I believe there's a case to be made that the attrition problem is greatly overrated, and in the ELS, by the third wave, had effectively recovered 96 percent of the data that you would have expected them to have collected in the wave 2 interview, which is quite good. Compare that to the cross-sectional attrition going into wave 1 at 12 percent. So, the data loss in these selected domains for wave two was really only four percent. So, let's keep things in perspective.

I want to discuss two other examples. I'm going to compare the PSID and The British Household Panel Survey (BHPS). These are really nicely matched studies. They're both based on tracking households. They use proxy reports from a cooperative informant, which makes a big difference. PSID started in 1968 from a Survey of Economic Opportunity (SEO) frame they got from the Census Bureau, and somebody at the ISR had some real pull to get the PSID out of the Title 13. The PSID was augmented with an ISR augmentation frame. The BHPS started in 1991, based on an address frame. One, perhaps, little appreciated difference, and I think we have to keep this in mind when we compare panel studies internationally, is that, in this country, we fund survey organizations and interviewers based on cost reimbursement. Interviewers get paid for their time and their expenses of going out there and pursing respondents and trying to bring these people into the fold; whereas, in Europe and Asia, a lot of studies compensate interviewers, essentially, on a per head basis. And so, the incentives for an interviewer to really push hard to bring in a respondent aren't there because the risk is all on the interviewer and the capitation model of reimbursing interviewers. So, with that in mind, and also, bearing in mind that The British Household Panel Survey started, 23 years after PSID, the data about the retention rate of the two follow the same pattern – a steep drop and then somewhat leveling off with The British Household Panel being somewhat below the PSID.

Now, let's think about the hazard rate for attrition for the PSID versus the BHPS, the data indicate a similar sort of pattern that was present with the ELS, a fairly high attrition rate after the first wave and it falls, and then it starts to level off, although the ELS didn't run long enough to really see this pattern. But, both these household panels show the same pattern. So, what do we make of this? And, incidentally, the same pattern exists in the ANES. In the 2008 panel, there was heavy attrition in those high-frequency panels early on and then, it leveled off. This is perfectly consistent with the notion that attrition is being driven by heterogeneity in innate respondent cooperativeness, which means that these panel surveys slowly prune out the less cooperative respondents and, in time, we're left with the more cooperative people.

So, with that in mind, focusing on attrition, first of all, the attrition for these surveys is not at all that bad and we'll see later on that the news is even, potentially, a lot better. But, looking at attrition in panels, is equivalent to single entry bookkeeping because we focus on one of the problems of panel surveys, but we miss the substantial benefits; namely, it doesn't cost you anything to enroll the wave 2 respondents or the wave 3 respondents. You've already got them. So, an equivalent repeated cross-section effort is going to be substantially more expensive because you've got to go out there and you've got to sample and recruit people every time.

Now, over time, the panel studies have a tendency to de-cluster, that is, in the National Longitudinal Surveys (NLS), after about 25 years, if you look at a scatter plot of where the respondents are, it looks exactly like a simple random sample. So, these longitudinal surveys do de-cluster, so you do lose the economies of clustering that drive its use in the first place. But, nonetheless, you also have to remember that in repeated cross-sections, you have to keep collecting the same background information. And so, a lot more of your interview time is absorbed by the bookkeeping and keeping up to date and, "Who is this person that I'm talking to?" that the panel surveys just paid for once. And, in addition, the time varying data, I would argue, is more accurately collected prospectively than retrospectively. So, one action item for NSF to consider supporting is an effort, a comparative effort to ask, "How do we compare the problems of attrition with the problems of recall"? And unless you consider the problems of recall, which we know are serious, especially for non-salient events, the focus on attrition is going to be misleading.

The National Longitudinal Surveys are surveys of individuals rather than households with six cohorts started between 1966 and 1997, and what's interesting is that the four original cohorts were collected by The Census Bureau and NORC collects the data for the latter two. So, these are very similar surveys, but if you look at the young women, a survey which was begun in 1968, versus the NLSY79 and the NLSY97, we've got information that spans three different generations with starting points two decades apart with two different survey organizations. So, we've got a little bit of generalization that might be going on here.

These surveys returned to non-respondents, covering questions in many domains and sought to collect the data missing from the missed waves. And so, what I'd like to do is to develop the theme of respondent attrition versus data attrition. The completion rate for the young women's survey from the beginning to 35 years after the beginning has been quite variable; it had its good times and its bad times. If you plot a line for the normally computed attrition rate, you would say, "Oh, my goodness. Look at this attrition! At 30 years out, we're down to 60 percent." The issue is that this would actually be very deceptive, because if you go back and you collect the data missed, when you re-interview a respondent who had not responded to a previous round and use that to fill in the history, then the extent to which your data history is complete is actually much higher. So, over time, this trend line slowly lifts up as you pick more and more respondents who've left the survey earlier. So, instead of looking at a given point and saying, "Oh, gee whiz, at 31 years, we've got a completion rate of 60 percent", so, you might say, "Oh, we've got a respondent attrition rate of 40 percent, but our data attrition rate is really only 30 percent." And this isn't even as good as it gets. In the NLSY79, which has been remarkable for a high completion rate over many, many years, you see the same story. The normal completion rate, again, shows the sharp drop after the first wave and we actually have the completion rates going up for some of the subsequent rounds. And so, you can say, "Oh, after 30 years, we're at an 80 percent completion rate." Well, that's true, but, at 25 years, we had collected essentially 90 percent of the data that was available to be had from this group of people. Now, that's not too shabby.

So, if you look at these longitudinal surveys and if NSF provides the sort of funding for surveys such as the PSID to go back to people who've left and recover the data, some of these concerns about attrition start to go away. In the NLSY97, we see the exact same pattern – a sharp drop initially, showing the pruning out of the uncooperative – by innate respondent heterogeneity. And, you know, we have some rounds in here where the completion rate actually goes up. So, that's something that's rarely appreciated when people talk about attrition and longitudinal surveys. Sometimes, the field organization does a really great job, such as NORC did here and the completion rate goes up. But, look at how much of the data is complete. Out here at 12, 13 years, we're recovering 90 percent of the data and that's a lot better picture than what appears to be when the normally completion rate is in the mid-'80's. Looking at the hazard rates of attrition, there are some blank spots here where the hazard rate, as computed, actually goes negative because you actually got more respondents than you did the round before. So, the hazard is not well defined. But, again, you see this sharp, sharp drop in the hazard for these various cohorts and then a leveling off.

Now, going back to the young women's survey done by the Census Bureau, you can see that this solid line isn't as far above the dotted line as we saw for the two cohorts fielded by NORC and the explanation here was in the Census Bureau's following rule. The Census Bureau said, "Oh, as soon as these people are not responders for two waves, we're not going to follow them anymore," so, you know, we'd complain and grouse. And so, finally, I guess, I made a sufficient nuisance out of myself and they decided to go back to these respondents. Some of these bumps upward in the completion rate were generated by the Census Bureau recanting this decision not to return to past non-responders and, instead, go back to them and recover more of the data. Had they done this all along, the data attrition rate would have been substantially more favorable.

So, my argument here is that standard respondent attrition rates are a flawed metric; that we really ought not to be talking about them quite so much; that what we really should care about, I believe, is the completeness of the data. And so, we structure these longitudinal surveys around the notion that we're going to recover data when respondents drop out temporarily – and a lot of respondents drop out for idiosyncratic reasons. You know, their marriage goes into the dumper. They don't want to talk, you know? They lost custody of their kids. The last thing they want to do is to talk about their life. But 5 years later, 10 years later, we found that if you go back to these people, all of a sudden, "Oh, so happy to see you." And away they go. You pick up a lot of the data. Now, sometimes, this generates a long retrospective and the recall problem pops up yet again, but as the comments earlier revealed, someone said "You're better off with recall data than no data whatsoever, right? No data is the worst data you can possibly imagine." So, now, operationally, in the paper, there's a table three that estimates our ability to predict who's going to be a respondent and it turns out that easily observable factors, based on the previous survey or the respondent's history with the survey, will reveal and explain a lot of the variation in who is at risk to attrition. And so, once you know who is at risk of attrition, you can start to do something about it. So, the data attrition rate is really what you want to look at, not the respondent attrition rate. So, just in case you find that completely unconvincing, which would not be too surprising, how about turning to how we would reduce respondent attrition. So, we'll leave the data attrition problem to the side for now. I'll just assert that data attrition is the metric you want to use. But, if you're interested in reducing respondent attrition, which is also another way of reducing data attrition, what can we do? And it comes down to are we willing to walk the walk?

This table is also in – in case you're at the back and you can't read it, it's in the paper. This is an experiment done in 2000 for the NLSY79 and, the truth be told, this experiment just didn't materialize out of the blue. This really reflected a request from the Office of Management and Budget (OMB). We had been doing differential incentive fees for years and years and years because it was part of the folk wisdom of the NLS. And, finally, OMB got on our case and said, "Can you prove that these work?" And we said, "Yeah, we'll do the experiment." But, we'd been doing something like this for 10 years under the radar and we knew it would work.

So, what we did was consider two levels of respondent fees, a $40.00 fee and a $80.00 fee. And we made the distinction between what happened in round 19 for respondents who were not responders in round 18 versus those who were responders in round 18, this captured the notion that their past behavior has labeled them in terms of their innate cooperativeness. The bottom line is that the fee costs per incremental complete among the round 18 non-respondents was $133.00. For round 18 respondents, the incremental fee cost was $272.00. And, of course, what this reflects is that some of the people you give a fee boost to were going to complete the interview anyway. Now, this experiment was done at the end of the field period, so we'd already sorted out the most cooperative respondents and it was a matter of, "Okay, we're at the end. What can we do to get the completion rate up?" And so, "Well, we'll try this – bolt this experiment onto it."

So, after somebody's been in the panel for 18 waves, if you can get them to do the survey for $133.00, that's not a bad deal. So, money can convert some respondents. But, the point is, these fees are most cost effective when they're targeted to the people that you're pretty sure are most at risk of attrition. What happened in this experiment was no surprise at all, we knew exactly what was going to happen. We just had to tell OMB that's what happened. But, what did surprise us was that when we did this experiment, we saw something we had never seen before and that was that, all of a sudden, the field costs fell. And, in fact, they fell so much that among this stratum, the round 18 non-respondents, the fee experiment very nearly paid for itself. That is, field costs fell by an amount per case that was pretty darn close to the additional fee that we were paying.

Alright. Let's move on. This is an interesting example. We were finally able to persuade the Census Bureau, the "Last Bastian of Don't Pay Anybody," to do an experiment much like that in 2003. There was a group of women who had never gotten an incentive payment. So, in 2003, there was a zero fee, a $20.00 fee and a $40.00 fee and these were all based on non-respondents in 2001. That is, we had learned our lesson from the formal experiment. We'd made the case. "Yeah, if you're going to do incentives, target the difficult cases." So the Census Bureau agreed to do this and what we got was that the $20.00 incentive versus no incentive had an incremental fee cost of $52.00 per case and the $40.00 incentive was the incremental cost per complete was $82.00 a case. So, but, again, after 35 years of following a panel, if you're offered a completed case for $82.00, take that deal, it's a no-brainer.

The NLSY97 round 10 experiment used much the same strategy of stratifying the experiment by former non-responders versus responders and what we got was that among the non-responders, the incremental fee cost per complete was about $50.00 a case, but the incremental fee cost per incremental complete among the responders was $230.00. So, again, we don't have a lot of money to pay incentives. The money we spend has to be spent wisely. This tells you, pretty much, if you want to spend your money wisely, put your money where the payoff is highest.

So, that leaves us where we are. In collecting time-varying information, we have to either deal with recall problems and cross-sectional surveys or attrition in longitudinal surveys. And so, the open question is, "Which is the bigger threat?" It would be interesting to test the heterogeneity theory by administering the big five personality inventory in the first wave of a panel survey to see whether attrition is related to agreeableness, openness, conscientiousness and maybe even neuroticism. Again, we need to recast, I believe, the discussion of attrition in terms of data attrition rather than respondent attrition, although the things we do in terms of differential respondent fees that can stem respondent attrition help with data attrition. Of course, a lot of people find differential respondent fees problematic. Interviewers hate them because they think it's unfair because they feel like you are treating the uncooperative people better. But how many of us pay the same amount for our airfare even if we are on the same flight? You know, everybody on the airplane paid a different amount. So these objectors just have to get over it.

So, I think the bad rap that panel studies get for attrition is misplaced. Several waves of a panel study will almost surely cost less than an equal number of similarly sized, equally lengthy cross-section surveys and you wouldn't be able to get as much data in those cross-sectional surveys because you have to keep going back and collecting the background information over, over and over. And so, the basic conclusion I come to, and I'll admit that I have a dog in this fight, is that really, panel studies are an excellent value for money and we really need to think about that when we consider panel studies and keeping them going and supporting them.

Having, in addition, the initial data from a panel study also helps us do things with attrition corrections because we know more about the people who've left after that initial wave of a panel study. I'm not sure whether the PSID has done model-based weights or not where you take into account Round 1 characteristics of people and use that to up weight people with similar characteristics, aside from the usual sampling strata. I would argue that, perhaps, an equally important problem to attrition is the lack of accretion; that is, inflows of migrants into the country. My guess is migrants are probably less well represented by existing panel members because migrants are just very, very different. And so, in times of heavy in-migration, which probably described the situation from about 1980 through 2008, the seriousness of lack of accretion bias may actually rival attrition bias and that's something on which future research is needed.

Randall Olsen is Professor Emeritus of Economics at Ohio State University and Senior Research Scientist.After graduating from the University of Chicago, he was a post-doc at the University of Minnesota and then joined the Department of Economics at Yale. From there he moved to Ohio State University. He served as Director of the Center for Human Resource Research where he guided the National

Longitudinal Surveys of Labor Market Experience for nearly 30 years. He also helped found the initiative in Population Research at Ohio State. He responsible for many innovations in longitudinal surveys relating to survey content, methods of data collection and data dissemination. He also designed a reformulation of how respondents are engaged and encouraged to cooperate in long running longitudinal surveys. He has served of several panels for the National Academies related to survey work. He also chaired the Federal Advisory Committee for the National Children's Study.

60

Computation of Survey Weights

Matthew DeBell

The basic theoretical foundation for weighting survey data is not controversial. Weights account for each respondent's selection or inclusion probability, and as such, weights say how many people each person who responds to the survey represents in the population. We know that we need weighting because surveys have uneven probability of household selection, either by area sampling for clusters in face-to-face surveys or other clustered designs, also for oversampling of target populations where we want to increase the number of cases so that we get enough statistical power for subgroups.

We also know that weights are required because of unequal probability of respondent selection when you select a respondent within a household. If you pick somebody out of a household with multiple people, then their chance of inclusion is one out of all those people, where if it's somebody that lives alone, they're included with certainty, so this is also very straightforward. It is also common to weight based on unequal response rates within areas or groups where we can see that the response rates are variable. These ideas go back to the beginnings of sampling theory as it's been applied in social research for decades. This is the easy part.

Weights are also warranted where there are known errors, when we see that there are differences between the sample estimates and the known population benchmarks. For instance, when we compare to the Current

M. DeBell (✉)
Stanford, United States
e-mail: debell@stanford.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_60

Population Survey (CPS) and we see that we have 55 percent female instead of 52 percent female we see there's an error there. Now, in theory, to take care of all these things, we would apply weights in steps: we would have an adjustment factor for household probability of selection, we would have an adjustment factor for person probability of selection, and we would have an adjustment factor for non-response in observable categories. And generally, for the response rate within a geographic area, because, assuming you're doing a face-to-face or an address-based sample, what you know about every element of the sample is where it is, and you may or may not know anything else. So your non-response rate assessments might be limited to geography, although in some cases you might know more, depending on what sample information you have. And then the step after those simple steps is the harder, more interesting one, which is adjusting the weights so the sample better matches known characteristics of the population.

I'm going to focus on raking as post-stratification in this chapter, because that is what we worked on. It's a commonly used technique. But there are a few limitations, bear in mind. Weights do not fix non-coverage error, or at least people try to do these things, sometimes, but whether you're going to succeed is iffy. They don't fix non-coverage, because weights say how much the sample members count to represent the population, and if your count of a subgroup is zero due to non-coverage, you can't weight up zero to anything. Weights can't fix extreme non-response bias, generally. Weights also can't fix non-response bias for factors that aren't correlated with the factors that you used for weighting. You can only fix non-response bias if your weighting factors are correlated with the causes of the bias. And we also can't know if we're fixing errors on factors where we can't look at the population benchmarks. We only know we're fixing errors if we can look at the benchmark. Now, maybe we have a good theory for things that we can't measure, but in terms of what we can actually verify, we can't know that we're fixing errors on factors where we don't know the population benchmarks. We also have to bear in mind that weights are costly in the sense that they increase the variance, thereby increasing sampling errors or standard errors, and reducing the precision of estimates.

Next I'll introduce how weighting tends to be done. The points I want to hit on are that the literature doesn't have a lot of implementation specifics, which I'll elaborate on, and that inconsistent methods are used. I will give you some examples from the American National Election Studies (ANES) and mention some other studies – and then the last element of practice is how data analysts, not data producers, how data analysts handle weights.

The general literature: Groves' *Survey Methodology* textbook is obviously important. But it has just two paragraphs on how to do post-stratification weights. That's typical of many texts that are intended to provide a comprehensive view of survey research, and I'll just mention a few more citations there that are taking the same approach that weighting is just something they don't get into. That's something they assume will be left to the sampling statisticians. The sampling statisticians know how to weight. The problem is that the statistical literature addresses weighting in painstaking detail. That's wonderful for mathematical statisticians. But there isn't a more general set of recommendations in the literature that would tell people how to implement weighting in practice.

That's something that's left to individual judgment. This leads to inconsistent methods in practice. So Voss, Gelman, and King looked at polling, political polling data from over 20 years ago, from mostly 1988 and 1992, and they found that across the political polling organizations, everybody was weighting differently.

An analogy to consider might be the notion that common law lays out principles and then judges have to apply those principles in particular cases. That's kind of how weighting is done, except that we don't have law in statistics. So there's no binding precedent. Just because one statistician did something doesn't mean that others feel obligated to do the same. So we had quite varied practices 20 years ago, and we still do, that still describes the situation. Now, this is reflected in the ANES. The ANES has a few salient characteristics, the sample size is 1,000–2,000, it's a cluster sample, and historically it's face-to-face interviews but a web component is now starting to be included.

The ANES has dealt with weights in a variety of ways over the years. In the earliest studies, the outlook was they would draw self-representing samples, and then you could analyze the data; no weights were required. They provided none. And that was perfectly realistic I think. Most people who were going to look at survey data in the 1950s didn't have a means to do weighted statistical procedures on a computer anyway. So the harm there may have been slight. Now, some of the ANES studies include the number of eligible people in the household. So you could weight by that, but that's the only factor of the study that you could use for weighting. Other studies have cell-based post-stratification, so you make a cross tab with age categories. Maybe you'll have five age categories and four education categories. You look at the percentage of survey respondents who fall into each cell, you compare those percentages to the population percentages in each cell, and then you apply an adjustment factor, cell by cell, to make the dataset look like the population.

That's cell-based weighting, and that is a helpful thing to do, but it has its limitations. And then starting with 2008, we have post-stratified weights with raking. So over time, ANES has not been consistent in its weighting methods by any means, and the progression in its weighting methods reflects, I think, the development in attitudes within the discipline of "do weights matter?" Which started with "yeah, they matter a little bit, sometimes." Then "they matter more" and ultimately: "we need to take them more seriously." That's the trajectory.

Other studies are also inconsistent. And many more lack documentation. I think this situation is improving, that practitioners are increasingly heeding the call to be transparent about their methods, but the gospel has not fully percolated into the community, so we need to keep reaching out. And then the last issue I want to address about practice is user awareness, or user unawareness. So in political science, it is generally not the case that people use weights. That's slowly changing, but we did a review of about 100 studies in three of the top political science journals that used ANES data over a period of several years. Of those 100 studies, about three reported that they weighted the data and they calculated statistical significance using design-consistent methods. That's about 3 out of 100. One of those authors was actually a statistician. So users are not generally conscious of the importance of weights or the need to weight. That is the state of practice.

Some of the problems are probably evident in what I've just been saying, but I want to highlight them for you explicitly. One issue is that the average survey researcher doesn't know how to weight, doesn't know how to weight well, and there are different classes of survey researchers. Most survey researchers are not survey methodologists. They are more area topical specialists: political scientists, demographers, economists who have interests in particular questions that surveys can answer. They're not in the business of creating surveys, they're more consumers of survey data. So analysts need to have it explained to them that weighting should be done, and it's actually not hard for them to do that properly. And then survey researchers who are involved in survey production but who are not methodologists need to learn the lesson that weighting is worth their attention, not merely an afterthought. Survey statisticians know how to weight, but they're doing it ad hoc. They're applying lots of different methods: it's inconsistent. The consequences of these problems are that the results of weighting are not always transparent, they're not often replicable, they're not always comparable to one another, and that may mean that the weights are not optimal. So we think there's room for improvement.

In that vein, we should want to promote four things, I would argue. The first is more accessible guidance about weighting, which means training people. And that's obviously happening. We also want methods that are transparent, we want them to be replicable, we want them to be comparable to one another, and these things are all achievable through a standard of practice. This is not the same thing as dictating a particular single method of how weights should be done, but rather a standard of practice that calls for transparency and disclosure. It doesn't mean closing off alternatives. It just means having a common frame of reference for progress. And I want to say there's definitely more than one way to compute weights. It's not a matter that there's one correct set of weights for a dataset. Rather, there are weights that optimize one factor or another factor. For instance, if you want to maximize the accuracy of a certain set of statistics because they are the ones that led you to do the study in the first place, those weights might reduce bias in some areas, but they might increase bias on other variables you don't care about – that's just something you live with to minimize bias on the variables you do care about.

In terms of improvements that we have been working on: you'll recall I told you that ANES had been kind of inconsistent about weights. Leading up to our 2008 study we wanted to provide the best weights that we could, so we put together an all-star blue ribbon panel of experts onto a committee to give us advice, and they provided wonderful service. And the question that we asked them when we pulled them together and recruited them was, "we want to put together a set of recommendations for how weighting should be performed in general." We developed a procedure and that procedure has led to a tool that hopefully will be broadly useful.

One of the first things that the chair of the committee did was to ask me a bunch of very specific questions about the 2008 study design. What's the sample design; what are the areas, how big are they, a lot of minutia. And we said to him, and this is illustrative of how weights are done ad hoc, and so we said to him, "Well, thank you for your focus on detail, but what we really want to do is think about this generally. We want general advice that can be applied to more than one study, not just this one study. So please indulge me and I won't give you any of that information you want. Let's talk in generalities and put together a procedure that's a little more broadly applicable."

So that's what the committee advised us on, and the aim was to develop a general approach. We wanted to take some of the guesswork and what you might call the art out of weighting and make it a more rigorous scientific process. That was our goal. So the first set of recommendations is straightforward. They follow from this well-founded long-standing statistical theory of applying the steps of adjusting for probabilities. Some of the points are very simple. Others such as one about adjusting for unequal response rates is the tougher one because sometimes you have to make judgment calls about which areas or groups or subgroups you want to pay attention to when evaluating response rates.

The last part of the process, the part that's more involved, is how to do the raking. That proceeds in a series of steps. The procedure that we put together after getting the recommendations from this committee, and I should say the committee members don't all unanimously endorse all the recommendations that I'm articulating to you today, it's more that they all provided wonderful advice to us and we tried to synthesize that into something that they might quibble with pieces of, so just to acknowledge that, but the steps would begin by first conducting a benchmark comparison.

Once we've got a dataset, and we've handled the straightforward standard steps, we want to proceed to post-stratification, making the survey estimates look like known population estimates. How do we do that? The first thing we should do is gather the statistics for all the variables in the study for which we have good, comparable measures where we think there's low measurement error and we can find population benchmarks, maybe from the CPS, maybe from other kinds of government records or other surveys – the American Community Survey (ACS) is another good possibility. But the key thing is to get comparably measured variables where there are reliable population statistics, and some of those are hopefully for most studies going to be age, sex, and other demographics.

Then you want to interpret your benchmark comparisons. So look at the survey's estimate, look at the benchmark, find the error. It is the simplest arithmetic you could do. The error that you identify highlights the areas of potential problems that your post-stratification should fix. Next, you want to apply a selection criterion. So we're going to say we want errors to be no greater than an error level that you think is substantively significant. Maybe you'll say it's three percentage points, maybe you'll say it's five percentage points, but choose a threshold.

We recommend five points as a common threshold, but it really depends on the particular needs of a study. Five points is arbitrary, frankly. Rake to correct the errors exceeding the threshold. Then the key element is to assess the errors on all the factors that you have data on, because raking to fix errors on certain variables can introduce new errors on variables that you didn't change. So you have to look back at the benchmark comparison and do that again. Also, you want to look at the design effects, make sure that you're not inflating the design effect too much with your weights, and then repeat the raking as necessary.

What I've just described is the general outline of the process. There are a lot of details specified in our report that is on the ElectionStudies.org website, and I don't need to go into all those details now, because they're all spelled out in explicit form in writing there. The advantages of this procedure are that by explicitly spelling out each step in detail, we would make it easier for more people to handle weighting in a consistent way, and that in so doing, it would be easier for other people to evaluate each other's work and to make comparisons and to understand the sources of difference between two different sets of weights.

To make this even easier, Josh Pasek, who is now an assistant professor of Communication Studies at the University of Michigan, wrote an R statistical software package that implements the raking, and it's pretty easy to use. I'm not an 'R' user, but I've used his "anesrake" package to weight. But if you don't want to touch 'R' at all, there's an even easier way to do it. Gaurav Sood wrote a tool that is on this website: https://web.stanford.edu/group/iriss/cgi-bin/anesrake/raking.php. So you can go to this website, upload your data, specify the variables for which you have benchmarks, compare all the benchmarks to the dataset, establish a threshold for how big an error you find tolerable, and then the software will rake and produce the weights in a few moments.

Those are the improvements that we have worked on, and I will conclude this chapter by suggesting some additional improvements that we would hope to see. These focus on transparency and replication, disclosure risk as it relates to weights and transparency and replication, and the comparison of methods.

Prioritizing transparent and replicable methods is important because, just to repeat myself and themes that have already come up, these are clearly the hallmarks of scientific research and we really should be trying to weight by scientific means and not by having ad hoc attempts by disparate researchers with disparate methods that are not necessarily fully reported. If we prioritize transparency and replicability, this should increase the pace of scientific progress in the area.

One way to promote that is by making weighting a factor for funding. And I'm not suggesting that we're anywhere near a point where a funding agency could establish a standard where they say, "You will weight this way." We're not there. But what could be done is that a funding agency could say, in a request for proposals, "We expect to hear from proposers about their plans for weighting, and incorporate that in the standards for grant making, and that transparency and replicability of such plans would be important considerations."

This also raises some concerns about disclosure risk, so I want to run through that very quickly. Disclosure risk is the possibility that confidential information about respondents could be somehow made public as a result of the survey experience. The way that might happen is that somebody could download a survey dataset, look at the cases, and figure out, "Ah, that respondent is Matt DeBell." We don't want that to happen; that would be very bad. For one thing, we've promised the respondents confidentiality, so we can't let it happen. But we also have just a general professional ethical responsibility, our Institutional Review Boards don't want this to happen, the funding agencies don't want to happen, and the universities we work for don't want it to happen. So we have an obligation to take disclosure risk seriously. The reason this matters for weighting is that weighting may reveal characteristics in respondents like the exact size of the census tract where they live that would increase the chances that you could find the respondent. So in area probability sample or in an address-based sample, a selection unit is often a relatively small geographic area, and on the publically released data files, we don't say exactly where anybody lives.

On the ANES, the smallest area we talk about is a congressional district, which is half a million people or more. If we were to disclose that somebody lives in a census tract with a population of some precise number, there's probably only going to be one tract that's going to match that description within a given identified congressional district, so we would then be effectively saying, "This person lives in this very small area." That increases the chance that they could be somehow identified. And we can't have that. And the mechanism by which this might happen in disclosing information in detail about weights is that if we disclose all the factors that are used in weighting, one of the factors is going to be an adjustment for the size of the area where the person lives. So that's the problem, that by disclosing all the pieces of the weight, somebody could reverse engineer this information about exactly where somebody lives.

This is a problem if we want to have transparency and yet live up to our obligations regarding disclosure risk. How do we deal with this? One option is that we could not publicly release those detailed weight components. That's the way ANES has worked in the past. We've never released those details before. The problem is, this limits transparency. This is the very thing

I'm saying we need to do, that we need to be open about. So that would be unfortunate.

Another approach would be rounding off the numbers or jittering the data, entering some random error or deviation in the factors before weighting, but this might detrimentally affect accuracy, so that's something to be concerned about and it's something we have never done before, but it's on the agenda to think about. This is an area for further research. We need to think about how we can be as transparent as possible without posing any significant disclosure risk.

Lastly, we need comparison of methods of weighting. We put together this ANES procedure and the ANES raking tool and we're pretty confident that they'll produce reasonable weights. We hope that they'll produce better weights than other techniques, but this needs to be put to an empirical test. There's already a small literature comparing different weighting methods, but this is literature that it would be great to see grow and to see it grow with weights that are calculated using methods that anybody can very easily get into and make tweaks. The ANES raking package makes that easy, because anybody can get into it and make tweaks to it. It's all completely open, so that's something that we would hope to see in the future.

So, recapping what this chapter has covered, the theory for weighting is well founded but in practice, the implementation is uneven. We want to get the word out to data users that they need to use the weights, and we want to make it easier to calculate weights using methods that are comparable and transparent. Researchers need more guidance on this, and if funding agencies will establish a standard that they should be attentive to these issues, we can have progress in that area.

61

Paradata

Frauke Kreuter

We have seen over the last few years some examples of successful use of paradata to gain efficiency in the survey data collection, and some alerts to errors that can happen. However we still face serious challenges in using paradata on the fly, in production. We also still see challenges, with the tailored collection of paradata, and in particular the transfer across modes and survey organizations. I think there is a need for future research here, and there is a great opportunity here for funding agencies like the National Science Foundation (NSF) to help. Also I think we have failed to really use the wide variety of paradata that are out there, or could be out there, and have mostly focused on non-response and measurement error in using them. I think a lot could be learned outside of that. The U.S. Census Bureau has invested in combining cost information and paradata and increase the modeling of paradata, and a lot more can be done in that area as well.

These are the main points that I hope to convey, and I will now back up and define paradata before I come back to these points. Using the Total Survey Error (TSE) framework, we have the representation arm of building a survey statistic, moving from the target population to the sampling frame, picking an actual sample, and hopefully gaining some response. Along each of these steps in the survey data production, we can identify paradata that are

F. Kreuter (✉)
University of Mannheim, Mannheim, Germany
Institute for Employment Research, Nuremberg, Germany
e-mail: fkreuter@umd.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_61

produced in the process. For those that are not familiar with what paradata are, the thing to remember is that paradata really are data that are produced as we create survey data. They sometimes are automatically produced; sometimes they are collected, but they wouldn't exist had we not done the survey.

I think that distinguishes them from other auxiliary data sources that are discussed in this volume, such as data from commercial vendors or government records: paradata are data created throughout the process of data collection. Let me give you an example of what I mean. Call record data are the typical example for paradata. They are collected by some survey firms in the process of trying to reach sample cases. If a contact attempt is made Monday mornings, an interviewer would write down the time and the day this contact was made and what was the result, you could presumably think there is useful information in these data. Because if I try to reach you between 9:00 and 5:00, all five workdays of the week, and there is never anyone at home, but I do reach you on Saturday, I probably am right to assume that you are working and not available during working hours, but reachable outside those hours. So there is a piece of information here about the respondent, kind of a footprint in the call record data. That's sort of the idea, and the same is true for other footprints that we leave.

Switching over to the measurement arm of the TSE framework, we look at an example from keystroke files, clicks and mouse movements. If you think about the process of answering say a survey, one thing that you can always capture in addition to the actual number or a sentence that someone would enter into Web survey is what else happened as input. Did the respondent go back a page? Did she look at a different webpage? Did she switch screens? How long does it take to get the next screen? You might be somewhat more limited in what you can capture on the respondent side if it's a Web survey, where you need to ask for permisson to collect such data.

More options are available in Computer Assisted Personal Interviews and in particular if an interviewer is working with software that you or your company designed. A lot of these data can be captured there. One interesting example is looking at the vocal characteristics of interviewers and respondents. They can be used to inform both the non-response process and the measurement error process. There is an NSF funded study at the University of Michigan, where a lot of introductions of calls to cases were recorded. Bob Groves at one point gave a talk called "A Thousand Hellos," and was playing all the hellos throughout his talk. So that's basically what this body of data is, vocal recordings where you can measure properties of these hellos and anything else that is said in those introductory sentences. What this research has shown, in part, is that it doesn't really matter if there is a rising intonation at the end of a "hello" that the interviewer says, but it matters differentially for cooperation how welcoming and rising the intonation in the "hello" is that the respondent said when picking up the phone. If you hit a rather mute respondent that says, "Hello?" and you're not an enthusiastic interviewer, then that's sort of the worst scenario we can think of in terms of achieving cooperation.

We could come up with other pieces of data that one could collect in this process. But researchers mostly focus so far on these two, on keystrokes in various forms and response times that can be collected out of them, as well as information from call record data and interview observations. I'm not going to talk about interview observations because Brady West has chapters in this volume that discuss them in more detail. But I will use response times and contact data to give you a flavor of what has been done with paradata and give a little insight on what may be still missing. To do so I will distinguish between *post hoc* use of paradata and concurrent use of paradata.

Using response time as an example, we can distinguish these two uses, *post hoc* and concurrent, and there is, of course, a substantive use too.

Psychologists have previously used response time to evaluate attitude stability. Methodologists have used response times to evaluate post-hoc characteristics of instruments and settings. They have found poor wording, poor layout, and complex questions to increase response times in answering. Jon Krosnick's chapters in this volume talk about satisficing and that tends to go along with faster response. There you already have one indication for the difficulty in using these kinds of data. It is difficutl to distinguish what is actually going on, and you need multiple measures to actually tease that apart.

Another example where response times have been been used successfully is to evaluate interview administration and errors that can happen there. In the extreme case, you can see falsification of interviews, if the total interview time is faster than you possibly could read out a question, let alone hear and answer. In this case, most likely this interview or questionnaire was filled out at home without any interaction with a respondent. This can vary by mode. In a phone setting there is much more supervision, and this would be different. But even in phone settings, as we have seen in work done by Kristen Olsen quite a while ago, the time that interviewers take for surveys varyies over the course of the data collection period. In the field, the difference between interview one and interview 50, we see that interviewers do get faster over time. However, there is also an effect on the respondent with faster rate of speech. If I get faster and faster as an interviewer and then interview elderly respondents, that can be quite a challenge for them with respect to remembering all the answer options or even hearing what was said.

But all of these examples that I have mentioned are examples of this *post hoc* use of the paradata. There are very few examples of concurrent usage. Roger Tourangeau, Fred Conrad, and Mick Couper have been involved in experiments trying this. In their study, in the process of answering a Web survey respondents get feedback relative to the length they take to answer survey questions.

When it was seen that a respondent would be faster than a typical response to a particular question, respondents were asked to take their time in answering. The notion is that if we prompt respondents to slow down, the quality of the data could be enhanced and error is reduced. Similarly, we can think of these different uses for call record data or interview observations as post-hoc and concurrent. What is interesting though, here we see a lot of effort to increase efficiencies. How can I use these call record data to get faster, add more respondents? In the very first attempt for increasing efficiency, the focus was in phone centers: optimal call schedules based on prediction of when are good times to reach respondents. Data from face-to-face data collection from the National Survey of Family Growth have been used to predict response itself, and best times to call.

Increasingly more of these data are used in the field. Instead of just afterwards looking at how many contact attempts did it take to reach a certain person, if you have the data available, you can say, "Okay, when was that person contacted last time and what is the probability for that person to be at home, or not at home," the next day or the next time you try to reach that particular respondent. Not only the call history, but also auxiliary information about the area the respondent lives in can be used to predict when it might be likely that someone is at home. So if you have information about people living in Adams Morgan in Washington D.C., you probably do know that there are few kids around, and most of the adults are working. Thus you might not even try during the day. Those kind of concurrent uses to increase efficiency are much more prominent, sometimes guiding interventions, but sometimes just for monitoring and supervision.

Now, this kind of use of paradata is not new at all; we just haven't always called it paradata. The regional managers of offices at the Census Bureau would probably say, "Oh yeah, I've been looking at these data." But one thing that is clear is that it had been done on the fly, not necessarily model-based, and not in a valid, documented way. Also often not in a way that one could, afterwards, reproduce decisions and learn from them. So that's why, right now, I see a lot of survey organization trying to figure out how should we do this. There's a lot of reinventing of the wheel, or finding wheels to begin with, but that's what I see going on here.

There's also use of these data to focus on reducing error, and assessing non-response bias. There have been a number of studies where early respondents are compared to late respondents and one couldn't do that kind of analysis without paradata; otherwise, you wouldn't know who is early and who is late. That is one very simple example. Interview observations, which are covered by Brady West's chapters in this volume in much more depth, are often used for non-response bias assessment or adjustment. Mostly, though, even this approach to reduce error has been a *post hoc* approach, so the contact data or the interview observation used after the fact for adjustment and not necessarily for intervention in-between.

One striking thing is that while the whole discussion of paradata started with the keystroke data (Mick Couper in his 1998 presentation at JSM had coined the term "paradata" for those kinds of keystroke data), since then, very little research has been done to really do systematic work with those keystroke data. The reason for that, I think, has in part to do with how these keystroke data look like, depending on the system, they are very messy. They are not clean data. A lot of people who work in field data production are not statisticians or data scientists and know how to dissect strings of text and easily form datasets that then can be analyzed. Other reasons why we see these struggles with analyzing contact protocol data – or other paradata for that matter, are that they have missing data; they are erroneous; or they have a very complex hierarchical structure. They come from different datasets and they often don't easily align.

One thing that we found is that interviewers tend to fill out their contact protocols diligently, hopefully, for all the contact attempts they made. But then that last contact attempt, when the interview took place does not show up on the contact protocols because now they get access to the respondent and there is not even time to fill out that contact protocol. They did the interview and they did their job. Why would I record that in that other dataset? Often these systems are separated out. When you then merge interview data and contact protocol data, for example, from the European Social Survey, you may find that the same case was contacted 11 times. But according to the contact protocol datasets, you would only see ten contact attempts. According to the contact data you would have concluded that this was a case with no result at the end, or non-contact. But if you merge in the interview data, you see actually there is an interview for that particular case. So this makes it messy, and a lot of people shy away from using these data.

Now, not just the analysts, also the interviewers shy away from even producing these data, and I think that Nancy Bates at the Census Bureau can attest to how difficult it can be to convince the interviewers. Why is that? Well, in a lot of surveys, these entry systems look like tax forms. They are not designed for good entry. We have put as much effort into making this task easy for the interviewers as we do for respondents. It's a survey, right? In this case, our respondents are the interviewers that provide us these data, but we're not using our skills to help them.

We see statisticians or analysts that plan on making an adjustment, not necessary getting all wild and crazy about it because they do realize there are errors in these observations or recordings. Not just the missing data and the complicated structure, but the data might just be wrong to begin with.

I think it would be very beneficial to develop some form of open-access code repository to use these strings of keystrokes and transform them into something useable. There are some places like the University of Michigan Survey Research Operation Center, that have developed some SAS code that if you ask them nicely, they will share with you, but it's not easy to get and not all organizations have that. What I see is that organizations that I talk to, in particular those outside of the U.S., where this is not quite as hot of a topic, shy away from that initial investment. I know that Westat is building such tools right now, and I am sure that other organizations do that too. It would be nice to see some NSF-funded research to help with access. We also need more development for proper statistical methods to deal with these kinds of data. They have an odd structure, there is a time component there, there are discrete time points. They are very sparse; these sequences of different lengths of standard sequence analyses techniques don't work.

We might have a shot with the functional data analysis that we see in statistics. I don't know. There is just not enough data development right now in that area. Likewise, when you think about the entry system for interviewer observations, I think it would help to have consolidated efforts for entry systems, presumably there is an app for that soon, but we don't know that yet. Having an open-source application that is used by many would also lead to data that are more in sync and more comparable across different surveys. Finally, we do need some research and more tailored measures. I know that there was one question – or is one question on the table of how standardized should these paradata be, and depending on the needs and the users of them, they need to be tailored and not standardized.

The systems we do develop need to be flexible enough to allow tailored survey-specific implementations as well, and not just standardizations across them. Looking beyond the current problems, we have not learned enough about real time use of these paradata, in part because it's only more recently that they are available in real time during the field period. But we also haven't tackled on how to integrate their use across the different modes, and that's tagging onto the discussions about using different modes in the first place, that are covered in other chapters in this volume. One thing would be particularly interesting to discuss with respect to NSF funding is whether or not these new data management plans that are written for the NSF proposals should include a mentioning of paradata and their availability. My feeling is that they should, in part because of the striking success of most national election polls with only a nine percent response rate, as Scott Keeter discusses in his chapters. We know that the quality indicators that we have don't work or are insufficient to tell us something.

These paradata or this process information can document a good process and a good procedure. Just like in the good old Deming days, if you can ensure that the process is correct, if every step on the way was just like it should be, then you probably have a higher chance of a good quality product in the end, even if your response rate is only nine percent. If you can look at the contact data and can see calls were only made between 5:00 and 7:00 p.m. on Mondays, then that's probably a less mixed respondent pool than if call attempts were spread out over the week and over the different time windows, just to give a simple example here. One thing, though, that I often hear in terms of making paradata available is: "Oh, we can't do this. It's all confidential." This problem is probably solvable since addresses and geocodes are not needed to make the paradata information usable.

Frauke Kreuter is Director of the Joint Program in Survey Methodology at the University of Maryland; Professor of Statistics and Methodology at the University of Mannheim, Germany; and head of the Statistical Methods Research Department at the Institute for Employment Research (IAB) Nuremberg, Germany. Her prior appointments were with the Ludwig-Maximilians University of Munich and the University of California, Los Angeles. Frauke Kreuter is a Fellow of the American Statistical Association and recipient of the Gertrude Cox Award. Her recent textbooks include Data Analysis Using Stata (Stata Press); Big Data and Social Science (CRC Press); Practical Tools for Designing and Weighting Survey Samples (Springer). Her Massive Open Online Course on Questionnaire Design has taught roughly 100,000 students. Currently, she is building an International Program on Survey and Data Science with funding from the German government's Open University program.

62

Interviewer Observations

Brady T. West

There are a lot of exciting opportunities in the field of survey research, particularly with new kinds of data that can be collected and analyzed. One of the areas that I've done some research in has to do with interviewer observations, and I just want to review some of the research that has been conducted in this area, and then overview some future research and funding priorities in this area.

Interviewer observations are observations that might be recorded by survey interviewers for all sampled units, so everybody who was sampled to potentially participate in a given survey. And these observations could describe selected features of those units, including how many attempts have been made to recruit, features of neighborhoods, things of that sort. Interviewer observations could also be classified as those assessments that are recorded by survey interviewers for respondents only. This is an important distinction: We could have observations collected on every single sampled unit, but it's also common to request interviewers to actually collect observations and opinions about their thoughts on the actual interview, and how everything went. Did the respondent understand what the survey questions were trying to get at? Did the respondent seem to take enough time? Did they run into cognitive challenges with the actual survey? And those observations are also

B.T. West (✉)
University of Michigan, Ann Arbor, United States
e-mail: bwest@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_62

often collected in different surveys. In both cases, we have to determine whether what we have is process paradata, or observational paradata.

This chapter will focus mostly on the observational class of paradata, and these are items that are actually recorded by human interviewers, not paradata that are collected automatically. These are variables where interviewers have to make judgments and then record those judgments as a part of the data collection process. So I'll anchor this chapter with some other examples from studies that I've been associated with. There could be observations collected on the sampled area, for example, a small area as part of a multistage probability sample. In the Los Angeles Family and Neighborhood Survey (L. A.FANS), for example, interviewers were asked to record evidence of crime or social disorder, looking for things like trash or graffiti or various other elements of potential social breakdowns in the particular areas where they were working. So that is one type of subjective judgment, with the idea being that if we can ask the interviewer to record what's going on in a particular area, that information might be useful to correlate with some of the survey features that we're trying to collect in that particular area. Interviewers might be asked to record observations on the sampled household. For example, in the National Survey of Family Growth (NSFG), on which I work, interviewers are asked, among other things, to record the presence of young children.

The NSFG interviewers make this observation the very first time that they visit a housing unit. Before they first attempt to contact the household, the interviewers are asked to guess whether they think that young children under the age of 15 are present in that particular household. And there are many variables related to the presence of children that are collected in the NSFG that could be predicted by that interviewer's judgment.

Interviewers might also be asked to collect observations on the survey respondent. And I say the potential survey respondent because often times these kinds of observations are collected during screening processes. Considering another example from the NSFG, interviewers are asked to record their judgment of whether the respondent who has been selected from a screening interview is currently in a sexually active relationship with a member of the opposite sex. And a lot of people, when I say that, they kind of scratch their head and they say, "What in the world are you asking these interviewers to do? How are you going to guess whether somebody is in a sexually active relationship?" And it's been fascinating to see some of the things that these interviewers look for when they're trying to guess whether or not this person meets that kind of classification.

The stories that we've heard from NSFG are just all over the map. But interviewers use different strategies to make this kind of observation, and that's one of the things that I'm going to be talking about. And I mentioned the quality of the data recorded post interview. For example, in the Panel Study of Income Dynamics (PSID): Did the particular interviewer observe that the person was referring to records, or did they seem to have a good understanding of the various financial data that was being collected? There are a variety of observations being collected after the interview by PSID interviews to basically give the PSID study staff an idea of whether or not that interview went well in terms of the particular information being collected there.

So there are different stages of the survey data collection process where the interviewers might collect these kinds of observations. So why? Why even do this? Why do we care? Why do we ask these interviewers to make observations? Well, primarily it's an inexpensive source of potentially useful auxiliary information on sample units. We're going to hear a lot about linking commercial data and linking administrative records to sampling frames, but it's often a difficult task to try to collect good auxiliary variables. And by good, I mean auxiliary variables that are correlated with many of our different measures of interest that we're collecting in the survey, but also potentially correlated with an indicator of whether or not a person or a unit is going to respond to the survey. So if these observations are relevant for the survey measures that we're trying to collect, they may very well be worthwhile. And also, of course, we're asking interviewers to make observations, so we're talking about face-to-face surveys. How much longer are face-to-face surveys going to be conducted? Can we make observations effectively in Computer-Assisted Telephone Interview (CATI) surveys or Random Digit Dial (RDD) surveys? Guessing respondent gender is something that has been done in CATI surveys, but it is unclear if this is an effective practice given the potential errors in these observations.

In addition, existing literature has shown that interviewer observations can in fact be correlated with both response propensity and key survey variables. If we're thinking about non-response adjustments, we really need to have those two criteria met for an auxiliary variable to be useful when constructing weighting class adjustments, or whatever it is that we might be trying to do. And some initial literature has shown that this is the case for observations that are tailored to the content of the survey. In other words, it's not useful to have the interviewers just collect observations at random, but rather have them collect relevant observations that could potentially be correlated with key survey measures. I think there is a lot of work that could be done in the area of responsive survey design and how interviewer observations could play a role in adaptive survey design decisions while data collection is ongoing.

That is something that we started looking at in the NSFG, but I'm going to come back to that as an exciting area for potential future research. So why might we not want to do interviewer observations? Interviewer observations do offer some advantages, but there are problems certainly with this process of asking interviewers to record this information. Not all interviewers will view these observations as easy to collect or worthy of their time and effort. This can lead to missing data; this can lead to data with poor quality, which Gary was also mentioning. We have to think about this as another layer of data quality. We're essentially asking the interviewers to fill out surveys for non-respondents in the hope that this information can tell us something about the non-respondents. But if the interviewers don't care and they are exclusively focused on just completing surveys, which is what we're asking them to do, they might not take the time to really record quality information. So it's really on the survey organization to emphasize why these observations are important to record in interviewer training. We typically dedicate a whole hour in an NSFG interviewer training for the most recent year of data collection to emphasize these observations, and showing why it's important to collect these observations for the larger goals of the survey research.

Existing literature has also shown that these observations can be error-prone and frequently missing when interviewers are asked to collect them. So certainly they are not free from error, and depending on what we want to use these observations for, that could have critical implications for non-response adjustment, for responsive survey design, a whole variety of things – imputation of missing values, if we use these observations to do that. And in the volume that Frauke Kreuter has edited on paradata, Jennifer Sinibaldi and I conducted a review of all studies that have looked at the quality of interviewer observations and other paradata to date, and it can be quite shocking how low quality these observations are. And that's one of the first things that I said when I started my doctoral studies and I heard that NSFG was asking people to guess whether people were sexually active, I just kind of shook my head and I said, "That's ridiculous."

So one of the first things I wanted to do was say: Out of those observations that are actually correlated with what people say in this survey, are they actually useful for nonresponse adjustment? Well, it depends. It depends on the observation being made. And another aspect of why not to do interviewer observation is that some types of observations take large amounts of time for interviewers to record. And as we all know with survey research, time is money. And if we're collecting 15–20 minutes of interviewer observations for every completed respondent, and then throwing those data in a closet and never looking at them, what is the point? And if we're talking about PSID or one of these other large national studies, where you have 15–20 minutes per completed interview, and you have 10,000 completed interviews, that's a lot of money. And if we don't use those observations, we're just throwing it away and wasting the public's money. So we have to see, do the benefits of collecting these observations really outweigh the costs associated with collecting them?

In terms of best practices, at least based on what we've learned with NSFG and some other surveys, every observation that interviewers are asked to record should have a purpose. So, again, not just collecting observations at random. Every observation should have a purpose. It's all too easy to say, "Well, we should try to collect this," and, "We should try to collect that." There should really be a well-defined purpose for each observation being collected. Are they going to be used ultimately for non-response adjustment? Are they going to be used to predict response propensity and adaptive design? Are they going to be used to profile active cases for targeting? Are they going to be used to assess data quality? Why are we collecting this data? And there should be a plan in place for exactly what we're going to do with every single observation being collected. Following this idea, we cut the interviewer observation form by about 50 percent in the most recent cycle of the NSFG because we realized that nobody was using about half of the observations being collected. And certainly the interviewers found this to be more worthwhile.

So I just want to emphasize that collecting observations with no apparent purpose is going to be a waste of time and money, and that's not really useful to do that. Observations collected on all the sample units, so now I'm talking about trying to collect information on respondents and non-respondents, should again be correlated with key variables and/or response propensity. And to establish that this practice is worthwhile, we certainly need more studies of whether or not these relationships actually exist across a variety of surveys. I recently published a study like this just about NSFG, but that's a sample size of one in terms of how many surveys we're looking at. We need to understand whether these relationships actually hold in a variety of different surveys where observations are being collected. So more studies need to assess these different relationships because observations can in fact be useful for informing response propensity models. But if we're trying to conduct non-response adjustment, if the observations are only predictive of response propensity, but they have nothing at all to do with survey variables, we're not going to fix any of the bias due to non-response. So we need more empirical studies showing that it's worthwhile that we collect these observations if they have these properties. So the best observation should really be designed to serve as a proxy of key measures that are actually going to be collected in the survey.

So, for example, the National Health Interview Survey (NHIS) is currently considering whether or not to ask interviewers to look for things like cigarette butts, indications of aid for the handicap in a given housing unit, something of that sort because these are all potential proxies of different variables that are going to be collected in the NHIS. So we want to make sure that these observations, again, are relevant. Observations collected on respondents should, again, actually be analyzed. At least speaking for the SRC, it's been shocking to me how rarely this is done on the large surveys that are performed by the Survey Research Center. Frauke Kreuter and my colleague, Ting Yan, at the University of Michigan, and myself are trying to put together a grant proposal that's going to be looking at analyzing these data for many large nationally representative surveys. And seeing whether or not post-survey observations can be combined in some statistical – seeing whether or not these post-survey observations can be cluster-analyzed, or combined in some way to produce a data quality indicator that secondary analysts could actually use in their analyses. So can we put these observations together, and provide something in a public-use dataset that says these are certain cases that may not be worthwhile to analyze, because the post-survey observations really suggest that these respondents didn't care at all about the survey and may have just been providing garbage data. But we want to see if these post-survey observations have that kind of utility across a variety of different large nationally representative surveys. They take a great deal of time and effort on the time of the interviewer, so we should use these data to improve operations. And that's something that I've been talking to PSID staff about further in terms of looking about how this can be done.

These post-survey operations could really point to problems with the questionnaire, in general, or indicate potential data quality issues. We want to explore the idea that we could put something in the public use dataset that alerts the analyst to these kinds of issues, without, of course, raising confidentiality concerns. When possible, we should try to assess the quality of these observations using any available evaluation data. So, again, if we're trying to have the interviewers collect data on relevant variables for the survey that we're conducting, we can try to link those observations with what's actually said in the survey. It sounds pretty simple, but there are very few studies that have actually tried to do this. Again, we're going to be hearing about linking administrative records; do interviewer observations, of selected characteristics, match with what we know based on administrative records, assuming that the admin records are of decent quality?

We should also consider the reliability of observations by independent interviewers: if we ask multiple interviewers to rate the same thing, do those ratings tend to agree with each or not? Reduced quality in these kinds of observations based on validation data has been shown to impair non-response adjustments. And in the same volume that Frauke Kreuter has published, I have a chapter that looked at a variety of very complicated simulation studies, to look at the impact of higher levels of error in interviewer observations on subsequent non-response adjustments. And it doesn't take long for error levels to basically overwhelm the effectiveness of non-response adjustments, and basically make those non-response adjustments worthless. So this is something else that we have to consider. Interviewers vary substantially in terms of the accuracy of their observations. So when we study the quality of these observations, some interviewers are great; some interviewers are terrible. Interviewers don't have the same accuracy across the board; they're all over the map. So interviewer variance in the quality and the time spent on these kinds of observations is another area that we could certainly study, with the idea being that if certain interviewers who are collecting high-quality observations are using more effective strategies, then we may want to adopt those strategies for interviewer training purposes. But this is an area that has received no research attention.

Methods are also needed for standardizing the ways in which these observations are collected. For example, the European Social Survey, as a part of their training, has taken to using visual examples of how to make effective observations. So they will present a scenario of what an interviewer might see in practice. And then they have a general discussion of, okay, what kind of observation would you make when faced with this particular situation? And then the survey managers say, "Here would be the best approach to making this kind of observation." Again, something that is empirically driven based on past experiences that could be used to improve training. In the NSFG, we've taken to providing interviewers with known and observable correlates of what we're asking them to observe. So if they're absolutely at a loss, they can refer to other auxiliary information that we know about that particular sampled unit. We've programmed that into the Computer Assisted Personal Interview (CAPI) system. So they can say, well, if we're looking at a locked building and we're trying to judge whether children are present in an apartment on the 59th floor, we can take at least something that we know about this particular area, and make a somewhat informed decision rather than just completely guessing. And then asking interviewers to provide open-ended justifications for why they might record a particular value, so, in other words, not just having them record an observation, but also having them record a reason for the observation. This is something that we've tried doing in the NSFG for a couple of observations to see whether or not we can qualitatively analyze those justifications, and understand why there might be interviewer variance in how they're approaching this task of recording the observations, and whether or not those justifications could indicate that different strategies are being used in the field. Because many times these interviewers are just left on their own, and we say, "Do your best at making this particular judgment." We don't offer them any empirical evidence for how to make these judgments.

So that's just a summary of the work that has been getting done in this area; this is a brand new area of research. In reviewing the literature for my dissertation in this area, there's really been nothing. And a lot of the work that has been done is in gray literature. So conference proceedings, internal technical reports, the field really has not benefited from seeing the results of these kinds of validation studies. I think that's an area where more publications are needed.

There are a number of open research questions; I just want to kind of raise these for consideration. What is the accuracy and reliability of interviewer observations across a variety of different face-to-face surveys? So not just looking at one survey, but if we look nationally and internationally, are these observations of high quality? Again, this requires some kind of validation data, which can be difficult. There has been research conducted in this area so far, but, again, these findings don't get published in academic journals. So the larger field doesn't benefit from understanding what may or may not be effective in different surveys across the world. What are the drivers of observation accuracy in different surveys? So Frauke Kreuter and I have a paper in Public Opinion Quarterly (POQ), where we use multilevel modeling to try to predict the accuracy of interviewer observations in the NSFG, to see if accuracy is driven by respondent features, interviewer features, area features. What's making the observational task more difficult, and how can we understand that process a little bit more? For example, you could use these kinds of predictive models as a survey manager to see if it's basically pointless to ask for observations in a certain area because the accuracy of those observations would be predicted to be so low, and it's basically fruitless to have the interviewers try to do that. So we might replace the interviewer observations with predictions based on other auxiliary data, or something like that, rather than just having them collect garbage information. Another good question is what do interviewer observations add to what we already have on the sampling frame? If we've purchased commercial data or if we've attempted to link administrative records to a sampling frame, is it even worthwhile to ask the interviewers to collect these observations? Are they really giving us something unique on top on what we've already linked to our sampling frame?

In telephone surveys, obviously, there is much less that we can do in terms of record linkage. But we really want to see whether these observations explain additional variance in survey outcomes on top of what we already have on the sampling frame. If they don't, again, there is really no reason to have the interviewers take this kind of time. So we have to think about what we already know about the sample, when thinking about whether to ask the interviewers to do this. Another question is: "What are the impacts, both statistical and operational, of reduced observation quality on adaptive survey design strategies?" This is something that we are closely following every day in the NSFG, I think we're up to like 80 or 90 different daily dashboard charts of what's happening in the NSFG. But how can the observations inform daily decisions about how to proceed with a survey design? And do decisions that are based on the observations backfire because the observations have reduced quality? So we're trying to tell interviewers maybe target this case, or target that case, but are those targeting suggestions misguided because of the observations being of such poor quality? Or do the observations, on the other hand, really improve efficiency? Are we finding benefits from using these observations to make daily decisions in that kind of framework?

Again, I've mentioned these simulations that I've looked at to date. Another area that could use a lot more work is how these observations that could be prone to error affect the various types of statistical adjustments that are used for survey non-response. So far I've only really studied weighting class adjustments, Paul Biemer has also looked at callback models for non-response to see how error in interviewer observations could affect those types of adjustments. But there are a whole variety of other adjustments that are used in practice, like calibration adjustments, things of that sort. We need to study the impact of errors in these different auxiliary variables on more than just popular non-response adjustments, and think about all the different statistical adjustment techniques that might be used in practice, like generalized regression estimators, or anything else like that. These are areas that need more research.

Other questions include, what are effective design strategies for improving observation quality? And can we provide the interviewers with known auxiliary information, like I mentioned in the NSFG? Can we do something about training in terms of helping to improve that process of collecting these observations, or should we just use predictions based on auxiliary variables? These kinds of comparisons are needed. I mentioned the post-survey observations. How can these observations be used to improve survey estimates? Phil Kott mentioned this notion of using calibration estimators based on respondents only, and using auxiliary information that's only collected on the respondent set to adjust our estimates. And post-survey observations could be useful auxiliary variables for these kinds of estimators that are based on respondents only.

On indicators of data quality: I wrote earlier this idea that we're trying to propose of analyzing the post survey observations, and seeing whether there might be latent classes of respondents based on all of those observations, and then including some kind of indicator based on those latent classes that analysts could use when doing their analyses. And seeing how sensitive estimates are to including or excluding certain potential problem cases based on those post-survey observations. What are the sources of interviewer variance in the observation quality? I mentioned that the interviewers are really all over the map in terms of how well they do at this. Why is that? If all the interviewers are just guessing, why isn't there a flat line in terms of their observation quality? Why are some interviewers at 90 percent, and other interviewers at 40 percent? Are some just dedicating more time to the task?

We really have no idea why that is happening right now. This is certainly something that needs to be addressed. And here is a bright idea, I think it was Norman Bradburn that wrote about the disconnect between the survey researchers and the field staff. We could actually try talking to the interviewers and seeing what they think when they are collecting these kinds of observations. This is something we've tried to do in NSFG, and we've tried to do this in the PASS study in Germany, we tried to identify the ten best interviewers and the ten worst interviewers in terms of their accuracy. And then we held half-hour phone conversations with the interviewers to see what the ten best and what the ten worst were doing in the field, and we saw wildly different strategies based on those conversations. And that work was presented in a conference paper at the American Association for Public Opinion Research annual meeting.

So, again, this is something where actually communicating with the interviewers, not only to emphasize the importance of this practice, but to understand what they're actually doing in the field, is certainly going to be beneficial. And then what are the empirical tradeoffs between costs of collecting these observations and ultimately improvement in the survey estimates that come from collecting the observations. The big question that we really need to answer is whether it is worthwhile for the interviewers to take the time to record these observations, take in their surroundings, understand what they are seeing, and then provide us with that information. Is it really cost-efficient to do that? Are we getting something useful out of these different observations – because if we're not, again, it's probably just a waste of money.

So I want to conclude with some important considerations about this practice. Funding for this type of research will remain important, as long as the U.S. Government and other agencies are conducting large face-to-face data collections. Again, for interviewers to make these observations, they need to be actually seeing and hearing things out in the field. And if face-to-face data collections are being phased out due to expense, or whatever the case may be, it's questionable how long this kind of research is going to last as a potential track.

People don't think that the face-to-face mode is ever really going to die out, but a lot of these issues may no longer become relevant. But at that point, if face-to-face does go away, the quality of other auxiliary variables – for example, if we were to use Google Earth observations to make observations about particular areas, or satellite imagery or whatever the case may be, the quality of judgments based on those data may then become relevant. So the general theme of making sure that we're not using auxiliary information that's too prone to error remains important, really regardless of what type of mode we're talking about. And as Frauke mentioned, again, some studies have looked at the quality of telephone observations. But when you think about it, how much else can we really ask the interviewers to guess in a telephone environment? That's something else to consider.

Brady T. West is a Research Associate Professor in the Survey Methodology Program, located within the Survey Research Center at the Institute for Social Research on the University of Michigan-Ann Arbor (U-M) campus. He also serves as a Statistical Consultant on the U-M Consulting for Statistics, Computing, and Analytics Research (CSCAR) team. He earned his Ph.D. from the Michigan Program in Survey Methodology in 2011. Before that, he received an MA in Applied Statistics from the U-M Statistics Department in 2002, being recognized as an outstanding first-year Applied Masters student, and a BS in Statistics with Highest Honors and Highest Distinction from the U-M Statistics Department in 2001. His current research interests include the implications of measurement error in auxiliary variables and survey paradata for survey estimation, survey nonresponse, interviewer variance, and multilevel regression models for clustered and longitudinal data. He is the lead author of a book comparing different statistical software packages in terms of their mixed-effects modeling procedures (*Linear Mixed Models: A Practical Guide using Statistical Software, Second Edition*, Chapman Hall/CRC Press, 2014), and he is a co-author of a second book entitled *Applied Survey Data Analysis* (with Steven Heeringa and Pat Berglund), which was published by Chapman Hall in April 2010 and has a second edition in press that will be available in mid-2017. He lives in Dexter, MI, with his wife Laura, his son Carter, his daughter Everleigh, and his American Cocker Spaniel Bailey.

63

Leave-Behind Measurement Supplements

Michael Link

This chapter goes in a direction where we don't know a whole lot, which I think is very exciting. In fact, when the editors were asking me to contribute and gave me the topic of leave-behind surveys, I said, "Hmm, well, first, that's not address-based sampling; that's not emerging technologies." But I welcomed this as a learning opportunity as I delved into this topic. "Leave-behind" is something that sounds intuitive – but I think if I were to ask the readers of this book for definitions, what we're going to find out is we each have a different definition of, what exactly we mean by leave-behinds.

Let me start by previewing the punch line of this chapter. What do we know about leave-behinds? A couple of different things. One is that they're fairly prevalent in practice. But in reality, when you start looking up research on this niche of leave-behind methodology, you don't find a whole lot to really understand how this is done and how well it works. Much is known about the components of leave-behind, however. But again, we don't know much about the methodology itself. So, we know a lot about self-reports; we know a lot about different modes; we know a lot about non-response. But when you look at it within this context, you don't find a whole lot in the published literature. Typically, it's also the province of very large, complex surveys. You don't have this type of methodology being used in smaller, mid-sized shops, whether they're private, or they're academic. They typically tend

M. Link (✉)
Abt Associates, Cambridge, MA, United States
e-mail: Linkmi01@gmail.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_63

to be part of very large, complex data operations. And they're used for many different purposes, which I'll go through in just a few minutes.

Last but not least, they come in many different sizes and forms, and where I will take this is they don't even have to be traditional surveys when we're talking about the leave-behind themselves. In fact, the growth of some of the new technologies that we have, mobile and online, combined with the fact that we're getting less and less funding, might make this type of methodology much more attractive for future studies.

So, as Paul Harvey would say, "Let's turn to the rest of the story." So, what exactly is a leave-behind, and what exactly is it not? Well, first of all, when I wanted to look this up, I said, "Let me get at least a basic definition." So, I turned to where we always go, Paul Lavrakas' *Encyclopedia of Survey Research*. It's not in there. There's nothing in there about leave-behinds, leave-behind surveys, surveys that you're going to give to somebody afterwards. So, the question is, if it's not in there, is it really a concept for us? I guess it is; we'll continue to move forward. But then I started to look at trying to use all of our various search engines that we have, looking at all of the different – the key journals that we look at, looking at Google, simply didn't find this term as a focus of research at all. In fact, there was only one, and you had to go back to 1969, where the term leave-behind survey, or any variant thereof, was actually utilized as the focus of specific research itself. And then I started picking up the phone, asking a few of my professional colleagues, and they also kind of gave me this, "I don't know what this means; I have no idea what's going on here."

So, I did what we all do, turned to my social networks. Put a call out to all of my friends at various different places and talked to individuals at a variety of different organizations representing everything from government to large, private, not-for-profits to a number of university colleagues. So, a lot of what I have here today, that I'm going to present, is more input that they gave me on their experiences and the research. And what we're going to find is that the research exists, but usually not in published form. It's lying in the archives of everybody's studies, or buried in the appendix somewhere.

So, first of all, what are the characteristics of leave-behinds? Well, a couple of things kind of came to the fore when I was talking to folks. First and foremost, it almost always involves self-administration. Right? You've gone through a survey concept, and then you're going to give this individual a survey; you're going to give them a task, and it winds up you're sending them off to do this on their own. So, self-administration is a key characteristic here.

Second, data collection mode is often different than the initial mode, which means we're in the world here of mixed-mode designs. Oftentimes these are associated first with an interviewer-administered survey, and then you're handing it off again to self-administered. Maybe it's a paper survey in both instances. These days, oftentimes, it's a Computer-Assisted Personal Interview (CAPI) survey, followed by paper, or it's a CAPI survey followed by something that's completely different than a normal survey.

Typically, what you're trying to do is provide additional information to that main survey that you just collected. But it doesn't always have to be that way. In fact, I'll show you a major example that we have from the government, where the leave-behind really provides the key data, and the up-front survey is more providing some of the analytic data that's going to go along with that. The key thing here, and this is where we are definitionally going to distinguish a leave-behind from some other types of things is it's a task that's done immediately after the initial data collection period, and you are requested to do this at that initial period, which distinguishes this from a panel survey. It distinguishes it from a follow-on, where you have a set of individuals, after a study, and later somebody says, "Hey, let's use that group to do something different."

This is a planned part of the data collection process, but you're just simply not going to do it when you're doing that initial data collection. And nearly always, again, we see that this is part of large complex studies, smaller shops simply don't do this. Smaller shops seem to just collect whatever data they want at the initial time. They're not usually using these types of things. And then again, last but not least, these may involve surveys, but more and more what we're seeing is that this is involving very different types of things, whether it's keeping a diary throughout a week or two weeks, whether it's using electronic monitoring, whether it's collecting physical specimens. There are a number of variations and dimensions that this whole leave-behind is taking on these days.

So, moving forward, this is how I define this concept of leave-behind: essentially a form of data collection that's self-administered by the respondent and completed sometime after the completion of an initial survey. Pretty broad, but that helps us, again, to figure out what we're going to include here, and what we're going to exclude. So, for example, what would we include in this definition? Straightforward, one is a survey that perhaps you start with face to face. Once the face to face is done, you hand them off a paper survey, say, "Please take your time, fill this out, and mail it back into us later." That's a classic example of, I think, how this term has been used most often. Another example, though, might be that you take a phone interview, and you're interviewing an individual, and then you say, "Hey, I'm going to transfer you now to an interactive voice response system." Charles Turner's has used a lot of these types of studies examining HIV-related attitudes and behaviors. Public health uses this quite a bit. And again, I would argue this is a form of leave-behind, because you're leaving the main interview, which was with an interviewer, and you're passing them on, and it's self-administered in another format. Another mode then is diaries. We see this quite a bit, where again, you might have a modest-to-large, up-front survey that's done with an interviewer, and then some form of diary is left behind. It could be a television diary by Nielsen. It could be a physical activity diary. It could be a transportation diary. It could be the expenditure diary, but I think diaries typically fall into this category as well.

Well, what methodologies are excluded then, using this definition? Well, a couple. One, traditional mail surveys or self-starting online surveys, again where that is the main data collection; you're just mailing out to people. So, this distinguishes the leave-behind from all of the rest of the self-administered formats that we have. Another example, audio computer-assisted interviewing. A lot of times, again, in a large study, you might have a CAPI interview that's done. You have a sensitive portion. You give the person the ACASI to go in the other room. They fill out that part. You come back, and you finish the interview. Again, that's not what we mean here, because that's all part of that same main data collection. So, the question is if you don't have the right permission, you have to – it would have been part of the main study, but you leave it behind because you need the permissions later. I would argue that that probably doesn't – that's a variation here. But what we're really talking about here is part of the design. You were going to have an initial data collection, and then you've got this other piece that you know you're going to ask them for at the end, and they have to go off and complete that. That's more the way I would use it. But again, we're just kind of creating a definition for discussion here.

So, why use leave-behinds? Again, when we start looking at individuals, when I was talking to them about why they – not only did they have examples but then why did they utilize this methodology, first and foremost, the easiest thing is, "Well, we wanted more stuff. You know, we have these people, and why not? We need more stuff, and we can't seem to constrain ourselves to 500 questions. We needed all 800 questions." And so, expanding the data collection effort seemed to be one of the key things. Others see it as reducing respondent burden, allowing them to complete a portion of the data collection on their schedule. So, they had some portion of this that they felt they needed to have the interviewer there, and they wanted to make sure they captured that. But then the rest of it, they didn't have to keep the individuals there the entire time. They could allow them to finish it off on their own time and reduce burden.

A third reason that comes up oftentimes, and I've used this example already, is the privacy issue, where you're administering sensitive questions, and you're trying to get less socially desirable, higher quality data, and so you allow the individual to utilize a leave-behind format. Fourth, is data quality. Oftentimes this is seen sometimes as more effective than trying to do just a straight-up recall survey, especially if you're trying to capture multiple frequent events. Again, this is the argument sometimes for using a diary. Why would you use a diary rather than ask individuals to recount and catalog over the phone what they did last month? Because it seems to be that if you give them the diary, and they can keep it, and they keep it diligently, that's going to give you a better measure than the recall.

And then last but not least, there might be unique information that needs to be collected with the leave-behind. This might be the only way that you can capture certain types of information. Again, if you're looking at physiological measures, those types of things, you might have to depend more on a leave-behind than a survey.

So, what does the empirical research tell us again? We know not a lot about those components, as I mentioned before, and the concepts that are associated with leave-behinds, but we really don't know much about leave-behinds as a methodology. It has not, again, been really studied, or at least published, in the context of that leave-behind context. And again, I looked through a whole lot of various different journals, the usual suspects for public opinion and methodologists, as well as some of the things that my colleagues in public health use. I even went to Google Scholar, looking at the first several hundred entries, to see if we could find anything on leave-behind surveys. And again, simply didn't find a whole lot. There were four articles that somewhat mentioned this approach. The one I mentioned before was the only one that was kind of really focused on it. So, why the lack of research? Well, a couple of things. My guess is that studies have been conducted, but they're not described as leave-behinds. And again, I tried to use various different combinations of language, but it simply didn't come up as that. Public health is probably the place that we would find more of this, because I think this type of thing is done more in the public health realm than anything else.

But if you're really going to get at this then, we're going to have to have some advance knowledge of the fact that the study had this design, or you're going to have to really dig into each and every article to look through the methodology section to see, "Okay, does the methodology fit that?" So again, it's not an easy thing to research in the way that we do some concepts that are much more easily defined. Second of all, again, we discussed as a component of a broader study, but it's not typically the focus. And then last, it can often be viewed as an adjunct to the data, but not the main focus, and hence, the researchers themselves don't see that as necessarily important to publish. They're focused on the main pieces of the study that they're publishing. And again, you're going to find these write-ups in their methodology reports, those types of things, but it never makes it out to us and not widely shared.

So, with the lack of being able to look at empirical studies, what I've wanted to do is just give you a couple of example studies that folks gave me that I thought kind of highlight the typical – or some of the approaches that are being used these days. One example is the Health and Retirement Study and David Weir's chapters in this volume provide more detail about the various design of this being longitudinal, being a very large study. It's conducted as a panel study. What he didn't mention, though, is that it does have a very nice, classic leave-behind section to it. Ater the initial interview, respondents are left a booklet, which really looks at cognitive status and psychosocial topics, as well as some things on work and retirement. So again, it's the classic view of you have the interview. Some of the bioindicators are collected, and then as a leave-behind, there's a little packet that's left behind that says, "Please complete this when you can and mail it back to us."

Why do they do this? When I talked to some of the folks who were involved with the methodology behind this, they said, "Well, there's two things." One was that they wanted more data. They wanted to expand what they had in conjunction with the face to face. That face to face was already taking 140 minutes, and so this became a way of saying, "Okay, well let's reduce burden a little bit by giving these folks this." And then also, again, the fact that they were sensitive questions made it a good candidate for the leave-behind. Cooperation rates in this respect, being calculated as a percent that returned the completed questionnaire divided by those that were given it – and they got pretty good results, up in the high 80s over the last couple of waves that were done. While it looks like – while there was a decline, the researchers indicated that this wasn't due to a declining response rate, but rather to changing the design over those two periods. Interesting thing is, though, they did have lower rates of cooperation among the usual suspects that we see with all studies, which is Hispanics, Blacks. And then, in this instance, those most recently added to the panel tended to have lower rates of return for the leave-behind than others.

The second example, comes from the Consumer Expenditure Survey (CES) and this is the one where I talked about the fact that the leave-behind action becomes more the primary data collection than it does the – than it does the upfront interview. The CES is a panel study that collects information on buying habits of American consumers, looks at things like expenditures, income, consumer characteristics. One of the key things is that it's part of the revision for the Consumer Price Index, so it's a very visible, very important government study sponsored by the Bureau of Labor Statistics (BLS) and conducted by Census. The population focuses on the U.S. households, non-institutionalized. And the panel has two separate sample components. One is a quarterly interview – that's one set of people; and the other is a purchase diary, and that is a separate set of people. The leave-behind component is focused on the diary components. And the way that that works, first there is a face-to-face recruitment, and then once the household's recruited, there's the administration of a household characteristics questionnaire. This has demographics, household composition, work earnings, expenditures, and those types of things. There is a up-front data collection piece. Then the interviewer leaves behind a diary. It is a one-week diary. It's meant for entire household, and they are to capture essentially all of their expenses that they have throughout the week. This is food, clothing, all other goods and services purchased. And it's laid out in the classic grid, day-by-day calendar formats. Step 3, interviewer comes back at the end of the week, takes the first-week diary, gives them the second-week diary. So, the households actually keep two of these. They again go through that process of keeping the diary for the week. Finally, the interviewer comes back yet again, picks that up, and then has a closeout interview. So, this is an example where you have essentially two diary leave-behinds, kind of bookended by interviews with an interviewer.

Again, I talked to the researchers. Why is it that they utilize this methodology? Well again, they said that they needed to get the detailed activity, and they tried to do this via recall and decided that the recall just wasn't giving them the granularity of detail that they needed for the diary, that they could get from the diary. Cooperation rate for this? The most recent one that's published is from 2009. Just over three-quarters of the individuals returned the diaries. And again, that's the average across the two weeks. But again, what we see is a story here that we see with all other stories, and we saw with the other one lower rates of participation among Hispanic, black, and younger adult households. Now, what's interesting about the CES is that the results that they're getting so far as BLS is concerned aren't good enough. So, they have actually been working. for the last couple of years, on a major redesign, something they call Project Gemini, where they're looking at both the diary and the quarterly interview components. But it's really the diary that they are most concerned about, this leave-behind. And the recent panel recommended to them to move from paper to electronic, using a tablet rather than the paper diary to see if that would improve data quality, completeness, those types of things. But they're also examining other techniques, whether they should use scanning types of devices, have people keep receipts, download credit card information, go to outside sources.

So, this is an agency then that's utilizing a leave-behind methodology. It seems to get fairly good results. As far as I was concerned, if I get 76 percent response rate, I'd be really happy. But this is a gold standard study, and so they're looking for alternative ways to improve upon that.

Third and last example then, mentioned in David Weir's chapters as well, is the National Health and Nutrition Survey (NHANES). This is designed to capture health and nutrition studies for adults and children across the U.S., conducted by the National Center for Health Statistics, my colleagues at the Centers for Disease Control (CDC). This study combines multiple interviews, physical examinations, and some leave-behind components. About 5,000 individuals are done each year in 15 different counties that are selected.

There are in-person recruitments, with an initial questionnaire. Then individuals are taken to these mobile vans – and if you've never seen the NHANES mobile van, it's a very, very impressive thing – where the exams are done: physical exams, blood work, a whole battery of things. But then, there also is a leave-behind component, and this is very different than what we've seen before. There are two different things that are asked of. One is urine. They actually have a home urine kit that is provided. Now, they take urine at the truck itself, but then they want a secondary urine sample as well. So, they provide the kit. The tech explains how to do this. Folks go home; they provide the sample; and they mail this back in.

Now, the second component is a physical activity monitor, and it has taken various forms over the year. These days it's a wristband that's put on. And again, when they leave, the tech puts on the wristband. They keep this for a week, and this is measuring, then, their physical activity, their steps, where they've gone, their motion, and those types of things. At the end of the week, they take that off, and they send that back. So, again, very different type of leave-behind than what we've seen before.

The researchers ask, "Why do you guys use these?" Kind of three things came up. Again, collecting unique information that they couldn't have gotten otherwise. They didn't feel they could get this information through just self-reports. Reduced burden in the sense that by doing it as a leave-behind, particularly the urine sample, the folks don't have to come back to the trucks yet again. They can do that on their own and send it back in. And then last but not least, they feel that they can get more accurate data than they can via a recall by using these methods.

Cooperation rates. The most published results, about 94 percent for the home urine, and about 88 percent for the physical activity. So, what we're seeing with all three of these is – the large, complex studies – again, various different ways that they're using leave-behinds, and actually pretty decent, I think, cooperation rates in terms of what they're getting. Some of the other selected examples, very similar again. This just kind of proves the point that I think a lot of, particularly, the larger shops who are out there doing these types of things. People doing larger studies are using these, but we don't see them published very often.

GfK does a media consumer study. They administer a big questionnaire again. Then they leave behind a consumer questionnaire that gets filled out later. Kaiser Family Foundation has done something with children, where they interview the children, and then the children are given a two-day media diary. CDC does something very similar with a State and Local Area Integrated Telephone Survey-related project that they have, where there's a telephone interview, and then the respondents are sent something to mail back. There are two that are a little bit different. One is by the Pew Research Center. And what this study does is they are interviewing business executives. That's the main study. But then there's a major leave-behind that is part – that captures information about the business. What's interesting about this leave-behind is it's not necessarily the business owner or the business executive that does this, but they might hand this off to the finance department, and then to the human resources department, and then to the IT department because it's capturing various pieces of information about the business.

Last but not least is an example from Nielsen on their metered rating service. You could think of this in the same vein, where they recruit a home. They conduct an interview with the household. And if they agree to be part of our Nielsen panel, they actually then get a meter put into their home: one on the TV and another something that's called the people meter. And the thing about the people meter is essentially it's like a remote. And every time they walk into a room, they push a button that says, "I'm here." And every time they leave, they push a button that says, "I'm not here." So, there is something that they have to do. They do that for two years. But it's, again, another form of leave-behind.

So, that segues us into kind of the new technology areas. And how might new technologies really change or augment what it is that we're doing with leave-behinds? Online mobile platforms and mobile platforms being both phones and tablets, and then all of the various different Bluetooth-enabled types of devices that we have out there really are changing the way – the types of leave-behinds that we can utilize. If you look at online, online provides really easy, relatively inexpensive access at a consistent interface across multiple devices and platforms. And by that, I mean traditional PCs, laptops, tablets, mobile phones, for capturing and transmitting data. So, you could utilize online very much for leave-behinds, and in a couple of different ways. One, you could have a Web-based survey to augment the initial survey. So, in the old days, we would leave behind a paper questionnaire. These days we can simply leave behind a URL and say, "When you can, please go online and complete the rest of the survey online."

Electronic versions of leave-behind tools. I've mentioned before about BLS looking at taking their diary from paper to electronic. Nielsen, we're doing the same thing on TV, seeing if we can transfer it from paper to a tablet type of format. And then we can track and more easily communicate with individuals. When you have online access, you can start sending – you can send messages; you can send reminders; you can see if data's being entered – things that you could never do before with a paper questionnaire. So, I think it gives a whole new dimension to this idea of leave-behind, whereas before, with paper, people were very much on their own unless you called them back up by telephone, or showed up at the door as the Census folks do. Here now, again, if you're using electronic, you know when data is being entered, or when it's not being entered, and you can facilitate, I think, more easy communication.

In terms of mobile, mobile platform really offers us a whole range of new things that we can use, not just in terms of surveys, but image collection, audio collection, GPS, app-based entries, Bluetooth, Twitter, the whole gambit of things that are on that mobile phone can be utilized for various different types of data collection efforts. Examples – again, a leave-behind survey that's accessible via the Web, but you do it through the mobile phone. Web these days doesn't mean simply the PC or the laptop anymore. In fact, probably for more folks it means just accessing that through their mobile devices. Detailed transportation of mobility studies, obviously, is looking at using this, individuals with GPS. We could find out what routes people take to work, what routes they take to the store, what things they go by. Audio journals are something that are being used in public health. I ready a study of individuals that did detailed interviews with cancer patients, and then they had them use the audio portion of a mobile phone to be able to record on a daily basis some of their journal activities. And then that information was qualitatively assessed later.

The folks at RTI and other places are working with Twitter diaries, being able to use Twitter as a vehicle for sending information back and forth. And again, pictures, stories, videos, Bluetooth devices, a whole range of things that are mobile-enabled that we can utilize. And last but not least then, with the Bluetooth – and again, this I think has been much more – our colleagues in the public health side of things have really taken this to the next level in terms of the monitoring types of things that can be done.

Measuring environment hazards via Bluetooth-enabled types of devices, capturing the uptake of medicines. I've seen this where individuals have a special type of dispenser for the meds for each week, and you can tell if the meds have been taken or not taken. Recording blood glucose, blood oxygen, pulse, the whole range of things that we could put down there. But the key thing here is that there are devices out there now that are – that can very readily measure a number of the things we used to depend on recall for that can now be quickly and easily enabled, that can be used as leave-behinds in the studies that we have.

So, what are some of the areas that are ripe for investigation in this particular field? A couple of them. What are the lessons that can be drawn by focusing on leave-behinds as a distinct methodology? Or are leave-behind approaches so unique that you really can't generalize them? That, to me, was one of the questions. Is that one of the reasons why we don't have anything out there? Nobody really views this as a – everybody's looked at the components and not so much this is a methodology in and of itself. Is it worth looking at that?

Unit non-response errors, cooperation, compliance, how does this differ from the primary data collection? I think that's probably the key thing here, if you were to use a self-administered questionnaire as the primary device, are there anything different that we would learn about using self-administration as a leave-behind rather than the primary, or are the lessons the same? If the lessons are the same, then we know what's going on, and we don't have to go any further.

Which techniques and approaches, obviously, are best utilized as a leave-behind? Do leave-behinds actually reduce burden, or do they make us, as researchers, feel better about ourselves? You know? I mean that, to me, was, when I saw the 140-minute questionnaire and, "Oh, by the way, we're going to leave you this additional thing," does that really reduce the burden on the household, or does it just make us feel better that we didn't keep them for, you know, 200 minutes.

And then influence of having/developing a relationship during the primary data collection, how does that affect the leave-behind compliance and quality? So, that to me was kind of an interesting thing, because that is what you have going on here. During that initial interaction, you've developed a relationship with a respondent. How does that influence – some of the things that we know about that primary data collection when now it's in the form of the follow-up itself? Does that interaction carry over or not?

A couple of other areas, in terms of measurement error – are mode effects different in the context of leave-behinds than they are when they're used as a primary vehicle? Again, it gets back to that if you have a relationship, does that minimize any of the effects – the negative effects that we see when using these things kind of straight up? Survey data versus data collected by other techniques. I think that, to me, is just a huge area that as we – David kind of brought this up and was doing some of that exact examination here. Is it better to utilize some of these other techniques, or are we underestimating the power of recall and our ability to use better quality questions up front? Does item non-response differ in leave-behind situations? Is there greater – this is one for Jon – is there greater satisficing in the leave-behind context than in other survey contexts?

And then basic data quality issues are worth considering and researching. How do leave-behind approaches differ, again, from self-administration? Do timing and context change responses that one might obtain otherwise? And finally, compatibility effects, the difficulty of comparing surveys done at different times, by different groups, using different methods – how does that all play into things?

So, again, to kind of wrap things up, what did we learn out of all this? Again, leave-behinds are fairly prevalent in practice, not very prevalent as a research topic, or certainly a published research topic. Much is known about the components, but not much about that as a specific methodology. Province of larger, more complex surveys used it for many different purposes, getting additional information, privacy data quality, unique data collection approaches. Comes in many sizes and forms and may or may not include a traditional survey. And again, the growth of new technologies and the decline in funding I think will make these much more attractive areas in the future.

I will say this, just in terms of my own personal view of if we were to put dollars into this area, where would you put them? I don't personally feel like the traditional paper and pencil follow-up is terribly worthy of much money. Maybe, you know, a dissertation grant or two, but my guess is that the lesson we learned years ago about self-reports pretty much probably follow on, maybe with some nuances as a lead behind. Where the real meat and potatoes, I think, is, is in some of these newer types of leave-behinds, these new tasks. Again, they're not necessarily surveys. Sometimes they're actually tasks of various kinds that we ask individuals to do. That, I think, puts a whole new spin, a whole new dimension on what it is that we do, and that's the area that I think that's probably most right for funding in the future.

Michael W. Link, Ph.D. is Division Vice President of the Data Science, Surveys & Enabling Technologies (DSET) Division at Abt Associates, a leading global providers of policy-based research and evaluation for government, academic, and commercial clients. He is also a past President of the American Association for Public Opinion Research, 2014–2015. His research efforts focus on developing methodologies for confronting the most pressing issues facing measurement and data science, including use of new technologies such as mobile platforms, social media, and other forms of Big Data for understanding public attitudes and behaviors. Along with several colleagues, he received the American Association for Public Opinion Research 2011 Mitofsky Innovator's Award for his research on address-based sampling. His numerous research articles have appeared in leading scientific journals, such as *Public Opinion Quarterly, International Journal of Public Opinion Research,* and *Journal of Official Statistics.*

64

Ecological Momentary Assessment and Experience Sampling

Arthur Stone

This chapter is going to focus on ecological momentary assessment (EMA) and the experience sampling method (ESM) and I'll refer to both as EMA. And I'm going to talk about what they are, what the alternatives are to collect the same kind of information about a day, and then I'll talk a little bit about how some folks are trying to incorporate this into the survey research.

EMA brings the researcher in people's days. It's trying to understand experiences, behaviors, environment, and even physiological indices from the perspective of within a day. We first have to consider why we even want that kind of information, and that will be a question that folks who are doing surveys will have to seriously consider. Yes, it's a cool technique and you get this very granular information, but do you really need it? There should be a strong rationale for going this way. I'll get into a little bit of the conceptual rationale for why EMA is there in the first place, because that provides the rationale for why you might want to use these techniques. I'll get into what EMA and mention several pros and consciousness from a survey perspective and then get into these alternatives, which have to do with end-of-day diaries, yesterday diaries, and reconstructing yesterday.

So why go into a day in detail? There are a few reasons, some of which may be of interest to you, some of which may not. One of them is an accuracy issue and I'll talk about this in a minute in the conceptual section, but

A. Stone (✉)
University of Southern California, Los Angeles, United States
e-mail: arthur.stone@usc.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_64

basically there are validity concerns if you're interested in measuring small events or highly fluctuating states like emotions or symptoms by asking people to summarize them over long time periods,. Another goal of EMA is to generate high levels of ecological validity and I'll come back to that.

Yet another reason that you might want to get into the day is because you might want to know how people are spending their time, that is, time usage. A lot of people are interested in that and a lot of people are interested in linking time usage with other kinds of subjective questions like emotions or symptoms, for example. You might want to study diurnal patterns of experience and behavior. We have physiological diurnal patters of almost everything and it's kind of neat to start getting into questions about how do some of our experiential states go along with those physiological patterns and to get those diurnal patterns accurately you probably need to get into the day in detail.

There is much information that you can derive about contemporaneous relationships between variables, say how does craving go along with having a cigarette or having a drink or using drugs, and then there are lagged relationships you might be interested in. If something happens at time X, what's the probability of something else happening at time X plus 2 hours, for example? And then there's the possibility of linking to physiological data, which has really become a big deal lately because there are more and more ambulatory sensors that are available, everything from monitoring where we are to how much we're moving to glucose levels to all kinds of cardiovascular and even EKG and EEG.

But let's go to why you might want to use EMA in the first place. There's been a huge amount of work about the concepts of experienced information, and Danny Kahneman makes the distinction between the terms "experienced utility" versus "remembered utility." Some kinds of memories are fleeting and they code what we're experiencing at the moment: this is experiential memory. Then there's semantic memory, which stores a different kind of information, more in the realm of beliefs. And when we have a long recall period – that expands from how are you doing right now through how are you doing over a day or over a week or over a month – what happens is there's a shift in the kind of memory that is accessed in order to create the answer; as the recall period increases, people tend to go more toward semantic memory and beliefs.

But for many research questions we don't really want to know about a person's beliefs about their experiences and behavior, but we want to know about what actually happened. This comes up a lot in the world of pharmacological trials, for example. When you're trying to evaluate the efficacy of a drug or a new treatment, we don't want to know what a person believes about treatment efficacy, we want to know about actual efficacy. There's also a philosophical thing here about experience versus remembrance and it turns out that both are important. Because what you remember, whatever you come away from with an experience has implications for about how you will make decisions.

When we talk about concerns about long recall periods we need to be aware that long is a relative term – long may be a week for rapidly shifting experiences such as pain. There are a whole variety of cognitive heuristics, that is, rules of thumbs that the brain uses in order to summarize information. One of the more famous ones is the peak end rule, which is that you tend to remember peaks of experience, things that are salient and things that are relatively proximal to when you're completing the questionnaire. There are a variety of cognitive heuristics that I won't cover here, but basically they all lead you to the conclusion that asking people about certain kinds of experiences, symptoms, or behaviors over relatively long periods may be fraught with bias.

Finally there is an issue of ecological validity, which Brunswick came up with in the 1940s. It really should be called representative design, but basically it's saying you want to be collecting people – not collecting people, you want to be gathering information from people in the environments that they typically inhabit in order to get the best view of the world, to get the best take on how the environment is actually impacting them.

So what can we do to reduce these kinds of biases and get the most accurate information? We can limit the recall period: make it very brief. And there are procedures that we can implement in order to try to help a person remember accurately. We can also try to ask questions in a better way. We can help a person get to where you want them to be. And it may be that some information that we just can't get over long recall periods. Memory is limited in terms of what's encoded and what can be decoded. And some stuff you just can't retrieve from memory. So there are implications for survey research, because survey researchers are often interested in experiences and behaviors that are captured in these kinds of real-time techniques. But there are many issues for survey researchers and some of these are being seen in the literature on subjective well-being. You may or may not be familiar with that area, but there's lot of emphasis on measuring hedonic well-being, which is how people feel throughout the day, in survey research.,

A lot of what I'll be talking about today has to do with capturing a single day, that is, getting into depth in about the experiences of a single day. As mentioned earlier, you have to start out by asking is that reasonable for the goals your survey? Is one day a reasonable sampling of experienced behavior or environment for your purposes? Even if you can get it very accurately, we know that things vary from day to day. So that's a decision you'll have to make and it has implications. How much sampling do you need in order to get to a reasonable signal over the noise? It turns out that this can work out fine if you're doing extremely large-scale surveys, but there are probably limits when you get down to relatively small-scale surveys and getting information from a day may not be very enlightening.

So what is EMA? The main idea is that you sample people asking them about their immediate experiences during a typical day using a schedule that makes sense for the purpose of the study. You might collect this data using a handheld computer, using a cell phone, using a little questionnaire that you carry around with you and you have a watch that beeps. In essence, that is EMA, but there are all different kinds of scheduling routines that you can do and it can get quite complicated.

Another version of EMA could be called the coverage model and instead of asking the person "How are you doing right now, what are you doing? How are you feeling? What are your symptoms? Who are you with?" and so on, you ask "Since the last time you were beeped, what has happened?" So you're trying to cover much more of the day. In yet another type of EMA, there is an end-of-day assessment, and what you're doing there is asking about what happened during the entire day.

Two other models attempt to collect EMA-like data and are based on asking about yesterday. There are two ways of collecting data this way. One way of doing it is to simply say to the person, "[T]ell me about how happy you were yesterday, and tell me about how sad you were yesterday." The second broad way to do EMA is with a reconstructive process, which is much more time-consuming, and I'll describe it briefly. But the idea behind it is that it's a way to reinstantiate the events of yesterday by reconstructing it to get a better view of what the day is.

So the pros of EMA are real-time data capture with little or no recall bias. Mood can be associated with the environmental qualities, get diurnal rhythms but their point estimates, so you're not capturing the whole day. You may miss things that are important to you because even beeping a person 12 times a day which is annoying enough for most people, doesn't get you everything. It can be burdensome and expensive so these are typically selective studies, very small scale and non-representative.

So now moving to alternatives, there are sort of poor man alternatives or reasonable man alternatives for collecting similar kinds of data that you might get from EMA. So one of them is end-of-day diaries. These have actually been around for an awfully long time, since the 1940s or before, and typically what you do is you have a way of having a person report toward the end of the day before they go to bed, about the entire day. And you can do it in lots of different ways with paper and pencil methods, electronic diaries; interactive voice recording has also been used. You can use the internet, you can use cell phones. I would advise people stay away from paper and pencil because there are all kinds of compliance problems there.

So pros, it's a very rapid assessment. It's not burdensome. You can have people do this for 100 consecutive days and get very, very high compliance. Low participant burden can be relatively inexpensive about the consciousness; there is recoil bias even over the course of a day. We have two studies right now that show that some of the cognitive heuristics are actually a play; even over the course of a 12-hour recall, you get peak and end effects. They're not that big but they're there. You have no resolution of the day. You can't get into the day. You can't associate activities with other experiences.

One example, the National Study of Daily Experience, has data basically showing various kinds of positive affect that go along with a number of chronic conditions that you have. But this is a typical kind of study, actually a large one, 1,000 folks for 8 consecutive – and they do this with brief telephone interviews. You don't have to just do this electronically.

Now onto yesterday diaries; we're getting to that model where you're recalling yesterday and the advantage is it can – this is now getting into a realm I think that survey researchers are starting – well have been using because it starts to become practical now for the first time. End-of-day diaries are tough because it means that you have to have your telephone survey, if that's what it is, done in the end of the day and most survey researchers don't want to do that. So, this one can be done any time during the day. Now whether that when you do and it makes a difference, no one knows. But anyway people do it throughout the day. You cover the entire day yesterday just as you do with an end-of-day recall and it's been used very, very successfully.

The Gallup Organization has a daily that has collected telephone interviews via random digit dial telephone interviews since January 2, 2008: every day 1,000 new people have come into the study. There are now well over a million people in that study. The survey includes several questions about how respondents were feeling yesterday, what they were doing. I'm going to describe some data from the survey that was analyzed several years ago. The study examined a third of a million people and it had to do with well-being. There was a pattern of age going from age 20 to 80 and a global well-being scale that asked how happy are you with your life these days? And the results indicate a typical U-shaped pattern. But when you look at the questions about yesterday which is the point, this is what you get when you look at different kinds of mood states (slides indicate a U-shaped pattern for life satisfaction and decline from age 50 onward for negative emotions). For example, they also ask about stress and worry and the data produce a very, very different pattern than you get with the global well-being scale. And the differences are very significant. The data indicate that the proportions of people that are reporting that they had a lot of stress in their lives yesterday are going from 45 percent down to 15 percent. These results were published in Proceedings of the National Academy of Sciences (PNAS) journal a couple of years ago.

A lot of people are using this "yesterday" approach. It uses a very brief assessment, has low participant burden, and has relatively low cost.

Now the fancy, comprehensive method of doing yesterday evaluation of the day is with the day reconstruction method (DRM). This is a technique that that group including me came out with in 2004 and basically what we recommended is to reconstruct the day into episodes of the day, what happened during the episode, who you were with, how long it was, and you have a person reconstruct the episode of the day usually comes out to about 14 episodes that they identify. They label them, then they go back and they answer questions about what was going on in those episodes one at a time sequentially. In some of our work we've asked them about affective states, how happy were you during episode one, how sad were you, etc. And what you're trying to do is recreate, get to the same or similar resolution that you might get from EMA. So to do this, you start out, get some demographics, then you ask people to recreate the day and then you go back and have them answer questions about that recreation. So to give you a feel for each of those episodes that they identify, we might ask them what they were doing. We might ask them about their affective states just to give you a sense of what this is like. This takes anywhere between 30 and 45 minutes. You can do this with paper and pencil. There are now several web-based versions of this that are available. Arie Kapteyn at USC has done this with the American Life Survey and Understanding America Survey.

The patterns in the data from surveys can be then compared with data from EMA studies of much more restrictive samples and, indeed, you get very, very similar patterns and very, very similar effects providing some sense that maybe you are doing a good job. These data allow us to answer some new questions. For instance, one can select episodes when a person was at work and you can say all right, for people who said that they were feeling a lot of pressure at work, what was their affect like? You would see that for that person, their affect is much worse, they're much less happy than the person who did not answer affirmatively to that question. So you can start putting into play all kinds of information about where you are, what you're doing, who you're with, as well as with other kinds of background characteristics of people.

The third version of yesterday is very much related to the DRM and this is the American Time Use Survey, which is run by the Bureau of Labor Statistics, and there's a similar one that Allen Krueger and I worked on, called the Princeton Affective Time Use Survey. It's really the predecessor for what happened with the American Time Use Survey and I'll explain that. The American Time Use Survey involves reconstructing the day, as 12,300 people did in 2010. They go through and they basically do what the DRM does in this time-usage deal. But we received funding from the National Institute of Aging (NIA) to support an experimental module, at the end of the American Time Use Survey. We added in a module that randomly selected three episodes from the episodes they identified and had them do a variety of questions just like the DRM does. So this is the well-being module as it's called in that dataset. All of these data are available online; by the way, you can take a look and download the data if you're interested.

You might be asking how well does the DRM type methodology work, which I include the eight is under, due at reproducing EMA. Allen Krueger and I worked on a study where for 3 days we had people beeping 6 times a day. At the end of each day they had a DRM, end of each next day about the day they had the EMA and this just gives you some idea about levels of happiness derived from the two different methods. All right, not exactly right but not too bad.

I now want to switch gears to what are people doing in order to take some of these different ways of capturing information about a day and making them more palatable for survey researchers. Some of this work is being supported by the NIA and specifically with a target toward the Health and Retirement Study. There are efforts going on to create "survey friendly" versions of single-day assessments. Jackie Smith at the University of Michigan Institute for Social Research is funded to do this right now, again from NIA and has some preliminary results. And the work that Jackie Smith has been doing is now being piloted in the Health and Retirement Survey and the English Longitudinal Study of Aging and in a more local study called ROBUST.

I'd like to give you a sense of how this is being accomplished. The challenge is to do an assessment of affect and time usage in 5 to 7 minutes, which is from a survey researcher's world, is a huge amount of time from our world that's really no time at all in order to do some of this stuff. So they go through a series of questions, then they say, "Yesterday did you feel all of this stuff?" Which is very similar to those general yesterday questions but then they get tricky and what they do is they combine time usage with experience and this is just one of several questions. Yesterday did you work or volunteer? If yes, how much time did you spend working, trying to get a sense of how much time, the time usage. How did you feel and then trying to get a version of what experiences they had. These can be replaced with whatever you like, symptoms, you know whatever you happen to be interested in. So this is her way of trying to get the information that you would get from a DRM, but importantly there's been no reinstantiation of the day. So you might say that's bad. But frankly no one knows because it hasn't been tested. There has been no study of whether reinstantiated views of yesterday versus nonreinstantiated views of yesterday make any difference whatsoever.

We've been doing similar kinds of things. We have time-use items that are very similar to Jackie Smith's but we've been trying to figure out how well a short form to collect time usage does versus a DRM. So we don't have EMA here so what we do is we have people come into the lab, ask them about yesterday in our short form, and then we have them do a complete reinstantiated day using the DRM. So that's not perfect but what we find is that the short form does fairly well at times but then in some things the short form well underestimates leisure time and time with family and friends.

So this is an open question right now. If you look at the correlations that you get in trying to characterize how much happiness or engagement or satisfaction there was during a given day, from the DRM versus the short form, you see that the correlations are up in the 70s, but then they drop fairly low, which raises the question about whether dichotomies versus 0 to 10 made a difference. This is a glass half full versus glass half empty situation. The point is that a lot more work needs to be done on characterizing a day in a very brief way.

So to conclude, I think that there is a lot of interest for characterizing the day. I think survey researchers may have particular hypotheses and the reasons for doing this but you have to really think it through. Several of the studies have been successful with the alternatives to EMA. I think the Gallup work especially; however, I think there are questions remaining about how well these short versions of the day really do.

Arthur A. Stone is Professor of Psychology and Director of the Dornsife Center for Self-Report Science at the University of Southern California. Stone's early work was concerned with improving the measurement of life events and coping with the goal of understanding how events and coping impact our susceptibility to somatic illnesses. These studies led to an interest in psychobiology with a particular emphasis on how environmental events affect biological processes. Concurrently, he was researching how people self-report information about their psychological and symptom states. This led to the development of diaries measuring within-day phenomena, ultimately yielding a set of techniques known as Ecological Momentary Assessment. Stone has been involved with alternative methods for capturing the ebb and flow of daily experience for large-scale surveys, including the development of the Day Reconstruction Method. He has also been involved with the development of questionnaires for use in clinical trials (the PROMIS project).

65

Biomarkers

David Weir

In this chapter, I'm going to try to give you a quick overview of recent developments in biomarkers, but particularly with reference to what we've been doing on the Health and Retirement Study (HRS). I anoint the beginning of this movement with the publication of this volume *Cells and Surveys* from a National Academy panel that endorsed the use of biomarkers in population surveys aimed at social science analysis. And one way to think about it is, it's kind of a migration from studies which were mostly clinic based, even convenience samples, but some of which were working toward getting locally quasi-representative population samples, and then seeing what of the work being done there could be translated to surveys that had much more direct purpose at population representation. There's now been a companion volume in 2007 that reports on the uptake of these biomarker ideas in a number of studies, including a paper by me on how we did it in HRS.

Here are a few, and there aren't too many other, broad nationally representative studies that collect biomarkers, such as the National Health and Nutrition Examination Survey (NHANES) is kind of the granddaddy of them all. It's a study that exists to do biomarkers and always has, does attempt to do nationally representative populations and gives us, in fact, a very nice benchmark for a lot of the work we do, so it's really good for us that

D. Weir (✉)
University of Michigan, Ann Arbor, Michigan, United States
e-mail: dweir@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_65

it's there. The National Longitudinal Study of Adolescent to Adult Health (Add Health) was a longitudinal study of adolescents in the early 1990s, had been filed multiple times, and they've done a number of biomarkers. I'll talk a little more about that later. The National Social Life, Health, and Aging Project (NSHAP) at the National Opinion Research Center is a study that's focused on older people, social life and sexuality, and they've done a number of biomarkers, and then finally the HRS. There's a number of more locally focused studies that are more or less representative of the populations from which they're drawn. The Los Angeles Family and Neighborhood Survey collected biomarkers in the Los Angeles area, for example. The Wisconsin Longitudinal Study's been around for a long time, but they've begun collecting biomarkers on people recruited from Wisconsin high schools in 1957. The Women's Health and Aging study is in the Baltimore area, drawing from a Medicare list sample, and they do a lot of biomarker collection. The Oregon health study is an interesting one because that's an experimental design. Oregon conducted a lottery for Medicaid expansion. So, people who were interested in getting Medicaid, who weren't currently eligible, but met an expanded set of criteria could apply, and by lottery, some got it and some didn't. And so, they had a study design which involved following up, doing a before and an after on both branches of that experiment to see what happened to utilization, which is it went up, and health, which they think improved, but they will be using biomarkers to try and demonstrate that.

What do we mean by biomarkers? Anything, really, that's a direct measure of a biological state, including a disease diagnosis, or measures of physical function, or levels of different biological substances in the body. They can be classified in lots of different ways. Probably, theoretically, the most satisfying would be, "What is it that they're trying to measure?" But, in fact, for surveys, the more relevant classification is, "What do we have to collect, and how do we go about doing it?" The types of collection I'll be focusing on in this chapter I characterize as minimally invasive, and that includes physical measurements, which I'll give you some examples of and biochemical assays taken from blood, although both Add Health and the NSHAP study have used other bodily fluids besides blood to do biomarkers. Imaging is not commonly done because it requires heavy-duty equipment, but that's another type of biomarker. And then DNA I will talk about at the end; it's kind of its own thing.

The motives for doing biomarker collection are to get objective measures of health. I don't think I have a case for how it's going to help with the voter turnout problem. So, you know, resolving whether people are lying to

American National Election Studies is not going to be resolved by biomarkers. But if you're interested in health, I think there's good reason to consider biomarkers, and that is that they're more accurate and less subject to bias than self-report, at least in most cases. And then the second is that some of them, at least, measure things that people don't know themselves, but which may be relevant to things that we care about. The motives for collecting them, in terms of what kinds of analyses these could play into, can be illustrated with three examples.

One is just descriptive statistics of population health, which is the reason NHANES exists, but it's not the only study that might have that interest, and that is just to say how many people have high blood pressure, how many people have glucose problems in the population. And in case you don't know, it turns out in the U.S. a lot more than elsewhere. And so actually, for the U.S., understanding population health at this biological level I think is going to be very important for understanding why our health overall is not as good as other places.

In terms of the kinds of analyses that have driven interest in biomarkers, probably the dominant one is from social epidemiology, where there's endless repeats of observations that low SES or disadvantaged groups have worse health. And the prevailing hypothesis is that this works through some mechanisms involving stress. And so, you want to have biomarkers either as improved measures of the health outcomes, but more often because you think they're going to mark the pathway through which the social experiences affect health. And so, the goal is to get those sorts of measures. Currently, the kinds of biomarkers that are readily available aren't all that well suited to testing that hypothesis, but it's because those stress mechanisms are difficult to mark, not because people haven't wanted to do it.

Thirdly, economists, and I should mention that's what I am, like to put health on the right-hand side of the equation and argue that various economic outcomes are affected by productivity affected by your health. And there again, you might think of having biomarkers of health to mark the pathways through which health differences are manifesting themselves. But you also have, in particular, if you think about, for example, applications for disability, somebody says, "I'm sick, and I can't work," but we don't exactly how they compare to somebody else who says, "I'm sick, but I like to work, and I'm working." And so, understanding people at a biological level gets away from what John Bound calls justification bias, that somebody who's not working and seeking disability might present themselves as more ill or more disabled, in worse health than somebody who's not trying to do that.

There are some concerns for surveys, and I'll address a couple of these in some detail. Cost is obviously important, but also response rates, both to the biomarker request and possible implications for the survey itself, which may have many other goals besides measuring biomarkers. And there is potential risk to participants and ethical concerns about notification. The HRS, which I'll use to illustrate most of this, is a cooperative agreement between NIA and the University of Michigan, with additional support from Social Security. It's a nationally representative study, drawn from area probability samples, classically designed. The study population is 50 and older, currently about 22,000 people. The problem with people 50 and older is they die, and so the actual number at any point in time depends on when we last recruited a new cohort and how many of the old ones have gone on. The content of HRS has always been very multidisciplinary and has become even more so through the addition of these new measures. We're a longitudinal study, with a biennial interview schedule, every 2 years on the even-numbered years. It was primarily a telephone follow-up study through 2004. When we thought about taking biomarkers into the HRS, no one has figured out how to do them over the telephone. So, it does require either bringing people someplace, to an NHANES van, or to a clinic, or someplace else to be examined and have material drawn, or doing something in the home. And for HRS, we made the decision early on that we would focus on those things that could be done by interviewers, traditional interviewers, in the home. And so, for us, the biggest part of the cost, actually, was converting to a face-to-face study from a telephone study. And we did that to spread the cost a little bit with what I call the alternating half-sample design. We took our sample, and we randomly split it in half. In the first year, we did biomarkers half the group got them, and the other half got the phone, and then we switched them, and now they just alternate back and forth. This creates a 4-year interval between biomarker observations for people, which is not unreasonable. Much longer than that, you start to worry you're going to miss things; shorter than that, it's just more expensive. But it also has the advantage of operationally smoothing the work. So, we're doing the same amount of phone and face to face every wave, which means, from a management perspective, we're not doubling the number of interviewers one wave, and then cutting them back the next. It's a smoother operation. And it also means, since we're a relatively large sample, that each wave we're getting a population representative look at biomarkers, so we can actually mark trends from each wave.

A couple of other studies that have been developed in other countries, patterned on the HRS, use different models. The English Longitudinal Study of Aging (ELSA) in England drew its sample from the health survey of

England, and the Health and Safety Executive uses a nurse visit as its primary mechanism. It's kind of halfway between the National Health Interview Survey (NHIS) and NHANES in the American system. ELSA just uses that nurse visit model every 4 years as part of their design. And that's feasible in a country that's relatively small, because you can have a centrally located staff to do that. But it's still an additional expense of lining up that appointment, having a nurse go do it, and so forth. The Irish Longitudinal Study of Aging, which is newer, built two clinics, one in Dublin and one in Cork, and brings willing participants into the clinic for a pretty extensive and interesting set of assessments. So, there are ways to get slightly richer information, but they're not cheap or easy to do in a large country like the U.S. I'll begin the discussion of specific biomarkers with Group 1, physical measurements, which I define as things you can do to someone, with all or most of their clothes on, and without taking anything but the data away with you. So, a standard battery of such measures that's common across a lot of these studies is measured height and weight, measured waist circumference.

It used to be conventional, in a lot of biomedical studies, to measure hip circumference and get waist-to-hip ratio. It turns out, analytically, waist is really what matters, and the dividing by hip doesn't really add a lot. In terms of the interaction between an interviewer and the respondent, measuring around the waist is generally not too problematic. Measuring around the hips gets kind of dicey. And so, it was just, in our view, safe to leave the hip measure aside. We do blood pressure with an automated cuff the kind of thing you get for home use for yourself. Grip strength is a measure of how hard you can grip a calibrated device. That one is actually kind of expensive; that device is fairly expensive. Many studies do some kind of lung function test, which for us is a simple puff test that measures how hard you can blow at a single point in time. Others do more sophisticated spirometry that get you more complete measures of lung capacity and lung function.

We measure walking speed only at age 65 and above, and alternatives used in other studies include chair stands, getting up from a chair, or getting up from a chair and walking a little bit. There are different varieties of these things, but all get at your lower body mobility. And then finally, we also added a measure of balance, which for the elderly, for whom falls are a big risk for a lot of things, understanding balance seemed important. I want to make an important cautionary point. Participation in biomarker collection may be related to the health state you're trying to measure. I'll discuss this using some data from 2004.

We compare response rate to the biomarker request to the self-reported amount of disability. On the horizontal axis is how many self-reported difficulties someone has. So, zero means somebody said they can jog a mile without any difficulty. (1) Typically means they're not so good with jogging a mile, but everything else is fine. (2) Probably they have difficulty walking several blocks, or maybe crouching, kneeling, and stooping is the other common one. And then by the time you get out to six and seven, these are people who have trouble climbing a flight of stairs maybe, or lifting a heavy bag of groceries.

At the far end are people who are really into disability territory, and that's the number of common tasks – like dressing yourself, eating, bathing – with which you need help. What you see is that for the puff test and the grip strength, they both follow a similar pattern of slightly increasing response as you get to the middle here, but then as you get more and more disabled, less likely to respond, and when you get to the very disabled, really kind of unlikely to participate in these measures. And so what that means is that if you look at the distribution of measured phenomena, you're going to miss the most disabled, and you'll look at a slightly optimistic picture in terms of representing the population.

Why are people with zero or one self-reported limitations less likely to participate than people with three or four? That's harder to say, but it's probably just an issue of cooperativeness among people who are busy. The healthiest people tend to be the busiest, and sometimes less cooperative on things. There aren't a lot of people at that extreme in this population. Having seen this pattern in our pilot data from 2004, we made sure in the new design for 2006 that we paired all the physical measures with at least one self-report item or several that pertain to that ability. In that way, we could, at least minimally, through imputation or reweighting, make up for the fact that response rates were different in these different self-report categories and modify the observed distribution back to the full population.

A second reason for collecting biomarkers is to measure things, we think, better than we get in self-report. I'll give you an example of that with body mass index (BMI). A common assumption is that people lie about their weight. I don't know how many people think other people lie about voter turnout, but a lot of people think other people lie about their weight. In fact, it turns out people don't lie that much about their weight. The average difference was about three pounds in this sample and the correlation was .98. So, my personal belief is that measuring weight is not super necessary. It just turns out a lot of journals prefer it. So, we haven't decided to abandon it yet, but the self-report data look really quite good.

Height, in this population, is more interesting. The correlation is a little weaker at .94. The difference is a little less than an inch in the height, with people reporting they are taller than they actually are. But if you look at it by age, what you see is the overstatement of height gets worse as people get older. We think this is related to the fact that people actually shrink as they get older. And so, when you go to a person, and you ask them, “How tall are you,” that’s like, “What’s your Social Security number?” You got it when you were 18 years old, and it’s always going to be the same. And so, people tell you, “I’m 5′10″,” even though they’ve lost an inch-and-a-half since they were that height. And so, that’s actually a big part of the differential.

Whether this makes a big difference in terms of health prognostication from BMI isn’t really established, but that’s what the pattern is. We also see some evidence that people report a little more accurately if they’re better educated, or have a better memory, but it’s not really a strong effect. But here’s another cautionary point. So, when we put these together and measure BMI, which is weight divided by height squared, we find that the self-report underestimates BMI, on average, by about four percent because weight is understated a little and height is overstated a little and that makes BMI smaller. And again that comes more from height misreporting than from weight misreporting. But if we look at the percent who are classified as obese, by the standard definition of a cut point of 30 on BMI, which turns out to be kind of near the population mean right now in the United States, for the older population, a relatively small error in the mean moves a big part of the distribution across that threshold. The self-report BMI is a little under 30 percent obese, but the measured BMI is 38 percent. So, a 4 percent change in mean BMI translates into a 29 percent increase in the fraction of the population that you’d classify as obese. So, my cautionary point number two is beware of cut points. As a continuous measure, the self-report and the measured are going to give you almost identical answers. If you try and make a population description of the fraction of the population that is obese, you’ll get a very different number.

All right, so next I’m going to turn to blood-based biomarkers, which can either be done from whole blood or from dried blood spots (DBS). And what I’ve found is that it’s the blood-based biomarkers that most people have in mind when you mention biomarkers. And what they have in mind is usually not a pretty picture. People are very concerned about what this would do to their interview process and their participants. Whole blood is not an option for an in-home interview using a regular interviewer. Drawing whole blood requires phlebotomy training and licensing. And in addition to that, there’s sometimes handling issues about processing the samples quickly after they’re taken. So, that would require something different, closer to, say, the ELSA nurse visit, which is something we have piloted and actually does work. It's not inexpensive, but it can work.

Dried blood spot is feasible at much lower cost, and lower risk to both respondents and interviewers. And dried blood spots you get by using a small lancet to stick the finger. You manifest drops of blood and drop those onto filter paper, where it dries and then is storable in that form until you want to do your assay. The tradeoff for doing this much more convenient thing you can do in the home is that there's a limited range of assays available currently from dried blood spot, and there's some issues about the quality of the measures. So, to give you an example of the limited range of assays, Add Health and HRS, which are dealing with not quite opposite ends of the life cycle, but pretty different populations, turn out to do almost identical DBS markers, because those are the ones that are out there. HemoglobinA1c is a measure of blood glucose. C-reactive protein (CRP) is a measure of inflammation. Cholesterol is something we are all familiar with from standard physical exams. Two somewhat less-familiar ones are Epstein-Barr virus and cystatin-C. For the immune system there isn't a good marker of overall functioning. What tends to be done is you find something that the immune system generally keeps in check, and you see how well it's being kept in check. The Epstein-Barr virus, which is mononucleosis, almost everybody has had exposure to in adult populations. The amount of those titers that you find in the blood relates to how well the immune system is doing. How well it relates is kind of controversial. Cytomegalovirus is another one that you can do, which some people think is a better marker, and we're going to be doing some experiments with that in HRS.

And then finally, we do a marker called cystatin C, which is a relatively new marker of kidney function. It's actually a better marker than creatinine, which is the more commonly known one. And NHANES has now gone back and done cystatin C from some of their stored blood samples when that assay got developed. All right, so our experience with DBS is that this works well in the field, but in the labs not so much. We've had some issues with the labs. First we look at our consent rates – how many people agree to do these things in the home. It's over 90 percent for the physical measures; and for blood, it's in the mid-to-high-80s; and for saliva, it's in the mid-80s.

What I want to draw your attention to is that, particularly for blood, between 2006 and 2008, we did substantially better at getting cooperation with the blood biomarkers in 2008, and that's stayed pretty steady. Saliva consent shows what looks like a decline, but that's actually a mirage because it's about who we're asking. We repeat the request for blood; we don't repeat the request for saliva if you did it once so over time the ones who are left are non-cooperative. Now one of the issues that is real and that people do talk about is racial differences in cooperation with biomarkers. We found in 2006 that African-American populations were substantially less cooperative, particularly with blood requests, where there's an 11-point differential in cooperation. In 2008, as I mentioned, we did substantially better with all groups but in particular we did a lot better with African-Americans and reduced that racial gap in 2008. Now, one thing that people told us to look at was race-matching of interviewers, that African-Americans might be more cooperative if they were paired with an interviewer of the same race. And it turns out that's not true at all. If you look at the 2006 cells over here, the bottom left quadrant, you see that the worst group of all was a black respondent paired with a black interviewer. They had the lowest cooperation rate of all.

Generally speaking, we found lower cooperation for black interviewers, regardless of race of respondent and lower cooperation for black respondents, regardless of race of interviewer. And I have to emphasize that on almost no other metric of performance or cooperation do we see this. Our black interviewers are virtually identical to our white interviewers on all measures of performance, response rate on other things, and so forth. It's not a general performance issue. Similarly, our black respondents are retained in the sample at about the same rate as everybody else. It's not a cooperation issue on their side. It's pretty much unique to this, to the biomarker collection. So, what did we do that helped in 2008? There's no magic bullet, but my belief is that it's a lot about the interviewer training. In 2006, we thought we put a lot of effort into training interviewers, but basically none of them had ever done this before. And that appearance of a lack of familiarity is not encouraging to the respondent. If the interviewer seems like they're a little nervous about this, that kind of communicates itself. For 2008, we did better due to both the experience of 2006 and then further development of our training protocols in 2008.

For example, we have a set of videos that are on DVD that go out to prospective interviewers before they even come in for training, in which they see these procedures demonstrated, and they get past that queasiness factor of, you know, what it looks like to stick somebody's finger. So, you can't just add this on willy-nilly to a survey. You've got to train your interviewers on this. A final point about cooperation with these requests can be seen in the data in 2010 and 2012, when we're now going back to the same groups that we did before. We've also got some new respondents who were newly recruited. The data show, again, mid-80s for consent for both blood and saliva among new respondents, first time we're asking. For people who gave us consent in the past, it's in the 90s for blood, but not 100. So, some people who did it once said, “No, I don’t want to do it again.” And similarly for saliva, there are a few people who gave us a sample that wasn’t big enough to do what we needed with it and provide storage, so we went back and asked them for a second one, and they were pretty cooperative. But among people who refused us the time before, we’re getting half or more to do it the next time. So, it really does work to persist at these requests.

Lab validations became a kind of major preoccupation for us. There’s at least three ways people look at data quality from the assays. One is to compare the population distributions and their covariations with some important things, to a gold standard, which is usually NHANES. You can also look at test–retest reliability of the assays you’re doing. And then finally, you can look at external samples, where you’ve got whole blood paired with a dried blood sample and send those to labs and look at how those compare. And we’ve done all these.

For blood pressure, and not a blood draw, but our systolic blood pressures in 2006 and 2008 looked virtually identical to the NHANES systolic blood pressure for the same years of collection. We did this from 2006 for a range of things, and it is basically identical, except for diastolic blood pressure. So, I described systolic blood pressure, where we match almost exactly. But for diastolic, there’s a very substantial systematic difference. NSHAP, which used the same cuff as us, got virtually identical diastolics to us, also quite different from NHANES. NHANES’ doctors listen to you with a stethoscope, and diastolic blood pressure is when that sound goes away. And that’s actually a pretty hard thing to mark. I can remember Raynard Kington, when he first started working on NHANES saying probably his biggest nightmare was standardizing doctors and how they do blood pressure. And that’s the hard part. So, I think this is really just a result of having the standardized cuff. But again, for cut points it matters because if the mean level is different, you’re going to push more people to one side or the other of that cut point definition of who has hypertension. You’d expect A1c, a measure of blood glucose, to differ between diabetics and non-diabetics, and it does, and the levels are fairly similar between us, NSHAP, and NHANES.

For test–retest reliability, we did a very small sample in terms of a number of people, but we took 25 blood spots from these three people and sent them in periodically to the lab in 2010. And for HbA1c, the measurement error is very small. It’s about 5 percent of the population variance in the biomarker measure. For cholesterol – total cholesterol, it’s much higher, and for HDL it’s half signal half noise from what we get from the test–retest reliability. That pattern is pretty much confirmed by what we got from a large project comparing DBS to whole blood drawn from the same people at the same time. That's part of a project led by Eileen Crimmins looking at a variety of labs and a variety of assays. CRP and A1c show very high correlation between dried blood spot and whole blood. Cystatin C, our kidney marker, is pretty good, and the cholesterols are worse, and that seems to be something people have found generally. We thought we had a solution, using a different marker called apolipoprotein A and B. Turns out it's subject to basically the same problems that cholesterol is, and it's also difficult to do. The problem, if I have any theoretical insight into it at all, is that when you dry blood and then reconstitute it to measure the quantity of something per unit of blood, how much you reconstitute affects that measure. And it's difficult to make that be exact. That's a big part of the problem with things like cholesterol. The other problem with cholesterol is that your red blood cells in the cell wall contain cholesterol. When those dry, that leaches into it, and the assay can't tell the difference between what was in your serum and what was in the blood cells. So, that's a double whammy for our cholesterol measurement. A1c is a ratio. It's glycosylated hemoglobin. Your hemoglobin cells have a life cycle, and it's pretty predictable. And if they're exposed to a lot of glucose, they change at a certain rate in this process called glycosylation. And so, by measuring the ratio of changed hemoglobin cells to unchanged, you get a measure of how much glucose they've been exposed to. That's a ratio. That doesn't matter how you reconstitute. So, that's why that tends to be fairly accurate.

Which biomarkers are important? There's lots of ways to define what's important; lots of outcomes to look at. A common example is mortality, which is certainly one important outcome to compare on. And I bring this up because I like to emphasize to people that these non-invasive, non-blood-based biomarkers of grip strength, the puff test, walking speed, and waist circumference are as predictive of mortality as any of these blood-based tests are, including the ones that are pretty well measured. Of the blood-based tests, cystatin C is the most predictive in this older population. When you get into kidney problems, which tend to be the end stage of people with hypertension and diabetes, that's much more predictive of mortality. And wealth also has some power to predict mortality, and conscientiousness, which is a personality domain. And they're about on a par with the average blood test.

Next, I'll summarize what we know about the impact of adding biomarkers to survey performance. Interview length is up necessarily. We went from a 75-minute or so telephone interview, to about a two-hour, in-home interview to do the biomarkers. And in our baseline, which we just did in 2010, it was more like three hours to do the interview. And that gets to be a little concerning. Our panel response rates for the core survey have vacillated between 88 and 89 percent per wave pretty constantly, and we actually got a bump upwards in 2006 the first time we did the biomarkers. So, we don't see any negative effect on our response rates, but we do see an effect on the effort it takes to do the survey. We have more calls per case. As a result of the increase in effort, our response rates so far have not been affected. One question we looked at was dd it cause people to leave the study permanently? Attrition in the literal sense of being removed from the sample is relatively low in HRS both cumulatively and from wave to wave. That's evolving over time, but not much. And we haven't seen an effect attributable to biomarkers. Now, I'm not going to tell you it never happened – there are certain anecdotes, where people say, "I want out of this study. What do you mean, coming in and collecting blood?" But they say the same thing about collecting wealth. They say the same thing about other requests – you know, if somebody wants out of the study, there are lots of good reasons. So, we haven't seen a mass exodus as a result of this, but certainly there are people who will point to it as a concern.

I'm going to turn now to genetics, which, as I said, is kind of a different thing, although it's also a biomarker. Collecting DNA is easy and inexpensive, relatively speaking. Analyzing it is costly. And the benefits from that analysis are, at this point, kind of uncertain. The analysis costs are dropping very rapidly, and the benefits, I believe, are increasing. So, my advice, and my colleagues at the Panel Study of Income Dynamics (PSID) have heard this, is you'll wish you had. Do it sooner rather than later. This is my recommended device for doing it. Oragene is a Canadian company that makes collection kits, and there are a number of them to choose from. This is the kind that we've tended to use. You dribble saliva into this cup, and you do that until you fill it to a certain point. And if you do that, you're almost guaranteed a pretty good volume and concentration of DNA for doing most kinds of analyses you'd want to do.

In 2006, we used a different technology, which is a mouthwash that you just kind of rinse your mouth with a solution and spit back into the test tube, and that had more variable concentrations and volumes, which is why we had to go back and replace some of those samples to maintain the repository. The samples that we had from 2006 worked fine. We haven't had problems with analyzing them, but Oragene is definitely the preferred technology. It costs $20.00 to $25.00 a kit, depending on how many you're buying.

We started DNA collection in 2006, when we did our other biomarkers. We had no funding to do any analysis with it, just store it. We had a lot of discussion about what to do with it, and that led to a consensus in favor of the GWAS approach. That is, doing a broad genome-wide assessment of what's on each individual's DNA. Because it's centrally managed, we do it once. It uses less sample than sending samples to each individual researcher to do whatever they were interested in, and it supports a wide range of research interests with a single pass, as opposed to, you know, "These are the genes that the cardiovascular people care about. These are the genes that the savings people care about. These are the genes that the psychologists care about." You've got coverage of the entire genome. There are also drawbacks, which I'm not going to emphasize right now. The American Recovery and Reinvestment Act (ARRA), the stimulus funding to the National Institutes of Health (NIH), created a window of opportunity for us, which is really fortuitous. We'd collected the DNA; we'd thought about what we wanted to do with it. All of a sudden, there's this call for proposals, and so we were ready to go and be successful with a couple of them to do the genotyping. We've got 12,500 cases that are already available, and that'll be up to close to 20,000 when we finish off the 2010 and 2012 samples.

Consent and confidentiality are important issues here. HRS uses a pretty broad consent, which says you're donating this sample to be used in research. The research aims are fairly unspecified, and there's also no stated commitment or obligation to report back anything to you from the analysis. That doesn't mean we won't, at some future point, be obliged to report something. But right now, the state of genetic research is, particularly for older people, who survived the really strongly genetic diseases of childhood, there's very few genetic markers that are anything more than a modest elevation in the risk of something for which there's lots of other risk factors. So, there's just not a lot of reason to give people genetic information back at this point in time. The problems that you've seen about consents in DNA mostly come from studies that had a very narrow focus initially, collected DNA for that narrow focus, like studying diabetes, and then allowed people to use the DNA to study something completely different, and that's a problem. So, it's better just to ask for a very general consent when you do it. NIH initially took the position that genetic data was non-identifiable, which is sort of like saying fingerprints are non-identifiable, you know, but worse. Of course genetic information is identifiable. It's perfectly identifiable. So, we've had that concern about, "How do we protect our respondents' privacy?"

We work with the NIH repository dbGaP because if NIH funds your genotyping, you are required to deposit the results in dbGaP, which is a data-sharing mechanism that NIH has set up. They have a pretty strict set of rules about who gets access and how they have to maintain it. And we gave the genotyping data, with an ID structure that bears no relation to anything else you could have, unless you come to us and get our approval to have the link to match that to public ID. And part of the dbGaP agreement, and our part of our agreement, is that you agree not to try to identify anybody. That hopefully prevents, a law enforcement organization, or something like that, from taking the database it has of DNA matched to names, matching it to our DNA, and then knowing who in our sample corresponds to their identified group. dbGaP has a website you can go to, to get information about the genotyping and so on. The data were released in April. There are about 27 groups that have been approved so far to use the genetic data.

When we did our genotyping, we included, with our cases, cases from something called the HapMap, which was a study of targeted ethnically pure populations. And this includes a group from Nigeria, the Yoruban tribe; and a group from Japan that are a pure Asian population. This point is important for social scientists, because genetics and social science have kind of a bad history. And that's because it used to be equivalent to saying, "Race is determinant." And what contemporary genetics does is essentially to take race out of the picture, and this is the way it's done. You calculate the principal components, which basically describe the main axes of variance across the genome. There are a lot of interesting things about the origins of the American population that can be done with data like this. The field of human genetics is very short on power. We need vastly larger sample sizes to do this research. It makes even a large survey, like HRS, feel small, but it also means we're far away from the point of diminishing return. So, everybody who comes in is really adding a lot to the power. So, my conclusion is, we've really only scratched the surface of biomarkers. Key biosystems are not yet marked with good markers. Dynamics of systems are not really tracked at all, and that's an important part of physiological function. Technology's going to move this forward, and I think we've now demonstrated that it's feasible to combine these kinds of collections with population surveys where they're useful. Thank you.

David R. Weir is a Research Professor in the Survey Research Center at the Institute for Social Research at the University of Michigan and Director of the Health and Retirement Study (HRS). He received his PhD in Economics from Stanford University and held faculty positions at Yale and the University of Chicago before returning to Michigan in 1999. He has led the transformation of the HRS into a world-leading biosocial survey combining its traditional excellence as a longitudinal survey with direct biological measures of health, genetics, linked medical and long-term care records from the Medicare system, and enriched psychological measurement. His research increasingly includes comparative analyses from the international family of HRS studies that now cover more than half the world's population.

66

Specialized Tools for Measuring Past Events

Robert F. Belli

My goal for this chapter is to provide you with some information about some of the kinds of things that I've been doing over the past several years; things that have been, if you will, governing my life and an opportunity to inform the directions that I think this type of work should be moving in the future. Skip Lupia's chapters raise the important issues of legitimacy and when I read those and looked at what I want to write in this chapter, the first thing is that the type of interviewing methodology one uses is an issue of legitimacy. And if one is interested in collecting retrospective reports, an important issue of legitimacy concerns what kinds of methods best promote accuracy in reports.

If you look at the goals of conventional standardized interviewing, it really is set up as a system by which one minimizes interviewer effects; that is, the notion is by presenting a standard stimulus, all the variance that you get will be the variance that are due to the actual experiences of the respondents, and there will not be variance from interviewers by engaging in different techniques. And this is an ideal, it's a goal, it's a worthwhile goal in terms of trying to minimize interviewer effects. But, its main purpose is not to optimize the quality of respondents reporting on their autobiographical past. And, hence, there may be a better strategy that we can use. And if one looks at the structure of autobiographical memory, there are various cues in that structure

R.F. Belli (✉)
University of Nebraska-Lincoln, Lincoln, United States
e-mail: bbelli2@unl.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_66

that will assist people to remember, more fully, their past. And the strategies are based on a very simple principle. One of the best ways to get people to report and reconstruct their past more accurately is to use whatever they have told you about, what they already can remember, as cues to remember more difficult to retrieve events. And because any of those kinds of cues are idiosyncratic to individuals, that is, what I have as a memory is going to differ from what you have as a memory and if those memories serve as adequate cues, then standardization becomes a real impediment in terms of providing standard stimuli because everyone's situations are fairly unique, and unique cues have to be used to optimize autobiographical recall.

There's an also, perhaps, an added benefit that speaks to issues about conversation that we know a little bit about and which we need to know more; and that is, there are aspects of conversation that help clarify meanings and clarify intent and there has been a literature that talks about standardization as actually providing certain levels of impediments to the ordinary benefits of conversation. Both calendar and time diaries, in many ways, do allow a method by which flexible interviewing, which is needed to maximize more accurate recall of one's past, can occur. And they're quite different instruments. The calendar instrument can go seek retrospective reports for decades, for the past year, and for one's entire life course. As Art Stone indicated earlier, time diaries are essentially a kind of method in which one is seeking information on what happened yesterday.

There is a calendar, which is implemented to gather information on unpleasant things having to do with exposure to intimate partner or domestic violence; such things as whether or not an intimate partner threatened to hit you, threw anything at you, punished, grabbed or shoved you, slapped you, and so forth, many, many different items. In terms of how calendars work, what this calendar reveals is that there are various different domains of interest, such as places where people live, residence, where they went to school, places of work. And the thinking here in constructing a calendar like that is to also collect a very rich relationship history that can either work by asking respondents to report partners that they can remember most easily to help or assist people remember partners that they had later. And it would also help them remember partners they may have had earlier. I refer to that as sequential retrieval as moving earlier or later in chronological time is a sequential retrieval process.

There are also opportunities for parallel retrieval. That is in different domains, having lived at a certain place, could provide a good memory cue as to who was your partner at the same point in time. And then, there's top down retrieval, moving from more general to more specific; moving from the name of your intimate partner to whether or not that partner had engaged in any of these domestic violence incidents.

The calendar I just described was a paper and pencil instrument used in face-to-face interviewing. When I was involved in the Panel Study of Income Dynamics (PSID), one of the issues there is that it's a telephone interview. And the other issue is that we're moving more towards computerized interviewing, Computer-Assisted Telephone Interviewing (CATI) interviewing in this case. Another example is a calendar that was developed for purposes of collecting information over the entire life course. It was a methodological study and it compared a calendar against a conventional questionnaire. But, pretty much, the same principle holds here; that is, we collected residential history earlier during the interview. We then went into an employment history, via various tabs and then went to the data entry sections of that interview and could use sequential retrieval in terms of remembering who the employers were at various points in time and cross reference that or use a parallel retrieval to point back to people's residences.

But, in terms of looking at the basic idea is that calendar interviewing will produce better data quality in terms of retrospective reports than conventional, standardized questionnaires, there have been a number of studies that have actually done comparisons between the two. I just want to go over this fairly quickly in terms of just pointing out that in some of these studies, the calendar was used as an aid. Colleagues in the Netherlands, Wander van der Vaart, and Tina Glasner have led this work. In this research that they have used various modes, face-to-face, paper and pencil, CATI instruments, web based instruments. And in some of their comparisons, they use the calendars in aid to a conventional questionnaire, conventional alone and what they observed when they had validation data or going under the rubric that more reports is better, assuming that underreporting is the norm under these situations, and they observe that the calendar plus conventional condition had done better than a conventional condition all of itself.

If we look to other work that has looked at calendars, administered alone versus conventional alone, we still observe, if you will, that calendars do better; not always, but more often in terms of accuracy with different kinds of criteria. In another case, applying the more is better criteria, we have looked at a validation set that consisted of PSID data that was asked in earlier interviews. With regard to that calendar that I described on intimate partner violence, I want to talk about a more reasonable pattern of results here. We'll know things have been observed in these kinds of studies and other studies using conventional questionnaires is that there's an age cohort effect in the sense that older persons will report exposure at later ages than younger persons. The age cohort effect, apparently, was due to recall problems because with the calendar, the age cohort effect was eliminated. And, again, in these validation studies, you can also see various modes, telephone, face-to-face, and computerized instruments as well.

Moving on, then, to the notion of a time diary. One of the things in which both of these methods, calendars and time diaries are similar, is the extent to which they're based on people remembering temporal information and, hence, we have, with regard to the PSID child development supplement, there is a time diary that's asking parents to report on their children what they did during the day – time beginning, time ending. And people then put sequences of different episodes. They also have what's referred to as secondary activities. Moving forward and backward in time is a kind of sequential retrieval. Secondary activities, is somewhat similar to parallel cueing, parallel retrieval. With the American Time Use Survey (ATUS) they apply a very similar idea, but now it's computerized. It is a CATI instrument; whereas if you looked at the PSID child development supplement, this was self-administered. So, now we have a CATI instrument. And some of the differences, which I think are worthwhile to point out, are that the ATUS asks people to either provide durations and stop times to each activity. So, people can either provide information on when an event stopped or how long it took. Reporting duration versus timing, if you will, may be important. And if you look at the PSID instrument, it only allows response reports in terms of time, not in terms of duration. There's also in the ATUS, if you will, these precoded activities that interviewers can introduce. And, also, there then would be verbatims; that is, information about activities that are not precoded.

In terms of the of the validity with the time diaries, alternate sources of data, such as experience sampling, beepers, produce fairly high correlations with time diaries, looking at spouse's reports, fairly high correlations; looking at a source of validation that is actually looking at electrical output that was used and then looking at time and day patterns of energy use, high correlations and this notion of reduction in suspected over reporting, if you will, from standard question form estimates has been observed in time diaries as well. What is needed in terms of future work and some of this is already ongoing; some of it is not ongoing. One of the big areas, which needs a lot more concentration are areas with regard to visual design. Issues with regard to visual design, which has been explored much more fully with more conventional, ways of asking questions and that, in terms of these various kinds of instruments and how they look and how they appear and how useable they are, really have not been as fully explored as they should be with computerized calendars and time diaries.

We have to get a much better understanding with regard to the interviewer–respondent interactional processes and how language and memory interact. And one way to gain some insight on that, which I'm going to show you, is conducting more behavior coding studies. There's also now the emerging availability of paradata, which can, for example, provide information of the extent to which interviewers are interacting or interfacing with the instruments. And we can also begin to look at different kinds of interfaces and data quality, just as I have here with behavior coding as well. And we also have to look at mode issues. I'm really fairly convinced – I think most of us probably would agree with this sentiment; that self-administered, web based instruments are going to be the wave of the future. And, hence, we have to begin the process of actually understanding more fully how we can implement these kinds of benefits of cues in a self-administered instrument and that is, in some ways, perhaps, replacing interviewers and calendars with an interviewer that may be more virtual.

Next I'm going to discuss what is involved in behavior coding by writing out an example transcript of an interview interaction and the behavior codes with each conversational turn. Historically, it has looked at interviewer performance and identifying problematic questions. My interests are ones in which you want to examine the quality of verbal exchanges between interviewers and respondents. There is a theory that I provided to you in terms of why calendars ought to work better. Well, they ought to provide more retrieval cues in comparison to standardized interviewers. Do they, in fact, do so? And, if they do so, does it matter? That is, if those cues that are made more available in calendars, if they do occur, do they actually matter in terms of data quality? We're going to get into the nitty-gritty of behavior coding process and the different kinds of behaviors that were observed. So, in an excerpt example of labor history from the life course calendar that I described to you that was computerized. So, somewhere in the middle of the labor history and the interviewer says, "And how long did you stay there, please?" It's a bit interesting that it almost sounds like a residence question, but it's not. Talking about staying at a place of work. And it's also asking for a duration. How long? Curiously, the respondent provides a timing response, provides a stop time, "October of '92." "Okay. And then, around October of '92, did you take another job?", which is a sequential probe. "I took another job and it lasted for, like, a month." That's a duration response, a month. "And then, I went to work someplace else" which is a sequential response, "not that month" and then, we'll take the next. Now, this is going to be confusing to you. It's a point of clarification from the interviewer. The interviewer's following with the respondent in already understanding that spells of work less than three months were not to be recorded.

"So, okay. The next job was at?", sequential response, "Let's see." Explanation of that response in terms of the employer name. Verification of the employer. Verification agreement. "And when did you start working for them?" "In '92." You notice that '92 appears as a directive query. In standardized interviewing it is something that you shouldn't do, you shouldn't direct a respondent to making a particular response. But, then it makes sense with regard to this person only having been at a work place for a duration of a month, that started in October, just really gives you an indication that the interviewer's really listening to the respondent. "Yes, December of '92 until May of '93." And you can, sort of, get the idea that the respondent's picking up on the task in terms of now providing the entire spell in terms of its beginning and ending. "Alright. December of '92 until May of '93 and then, in May of '93, did you", sequential probe, and then you get this fairly rich history from the respondent. "I went to work and stayed three" – that's a parallel response, a residence domain, for employer 11 and a data element response. "But, I'm trying to think how long I worked there." "I came in state 3 in August." Parallel again. "I have gone – I must have gone for a job," sequential. "I guess I started in January of '94," timing, "and only worked for six weeks duration and then, I went on disability." Another parallel response.

So, you get a sense that in terms of this calendar format, the nature of the conversation is one in which the language used is in more of a story format. We talked about autobiographical memory as providing stories. It's more in a story format consisting of all these elements associated with cuing mechanisms. If we do look at the main question of whether or not it makes a difference between calendar and standardized interviewing, in terms of the prevalence of different kinds of behaviors, indeed, it does. Parallel and sequential retrieval probes by interviewers, spontaneous retrieval strategies by respondents are all more prevalent in calendars. Conversational behaviors seeking to clarify meanings are more prevalent, but also, then, we have these more prevalent potentially biasing behaviors, directive probing and what we term as unacceptable feedback.

If we look at associations with data quality, it's not as clear as we would have liked. One of the factors that is important to take into account is people's experiential difficulty. That is, how complicated are their histories? Did they have multiple jobs in their past? Did they have multiple marriages in their past? And if we look at experiential difficulty, which means a much more difficult retrieval task, the use of these cues leads to greater accuracy, is associated with greater accuracy. It's probably an overstatement to say it leads to greater accuracy. However, if the past is unremarkable, a higher number of these retrieval probes and strategies is actually associated with less accuracy. Conversational behaviors lead to mixed results and rapport behaviors are also lead to mixed results. And so, we probably have to get even further down into the nitty-gritty of understanding, perhaps, in a more qualitative sense what's going on in these interviews to make some sense out of that.

We have recently acquired paradata from the ATUS. There are a number of different variables from those audit trails as Paradata, which we think will be useful. One of them, as I talked about before, which we actually saw in the calendars, is that people, sometimes, prefer to report in elapsed times in terms of durations and, at other times, in timing, in beginning and stop times. That may be important. Other kinds of potentially important variables are whether or not a precode or verbatim had been entered; whether or not the number of activity entries in the paradata is greater than public use data. We have access to public use data as well, which means that there has been some editing going on. Some of those activities were, for some reason, deleted.

We also are able to look at data quality variables, such – and these exist in the public release file or can be extracted from them, such as answers that are too vague to categorize; respondent has an unfilled gap in time; they may provide an overabundance of rounding their answers in terms of timing or in terms of durations and missing key reports of things that you think that people would do every day, such as sleeping, grooming and eating. We have some preliminary results and, in terms of associating among interview entries, associations between entries and data quality and association between interviewer characteristics and data quality, it's kind of curious with regard to educated respondents having increased vagueness. You would think that they would always have more decreased vagueness. But there might be more sensitive information that these more highly educated respondents may not want to report on.

I want to illustrate to you now the big picture. And the big picture is moving towards a greater reliance on computerization. I've worked with a conceptual model that was developed by a colleague at the University of Nebraska, Computer Sciences, LeenKiat Soh, and I want to give him full credit for this conceptual model. So, for the sake of, perhaps, not being as sophisticated as one would like, in terms of interesting concepts, I'm going to do my best list the elements of the conceptual model.

We do have lots of source of potential information. Paradata, behavior coding, interviewer characteristics and respondent characteristics. I mentioned behavior coding with calendars, Paradata with the ATUS. Ideally, you could have both. I did mention respondent characteristics and interviewer characteristics as well. And you can then go through various steps in terms of developing more intelligent, smart instruments instead of reliance, if you will, on the dumb instruments that we have now. Data processing, doing such things with regard to, mundane things, perhaps, with regard to the Paradata of the ATUS, for confidential reasons, that the verbatims have to be scrubbed out. They have to be sanitized. Well, programs are being developed to do that process instead of relying on the human eye to try to get rid of those verbatims. Machine learning and pattern recognition. We engage in these kinds of pattern recognition processes and seek to clarify predicting data quality, through our own intuitions. Well, computer scientists are beginning to develop data mining techniques that may be able to do a better job than what our intuitions are able to provide. You have the sense of having a great amount of data with regard to behavior coding. Talk about the richness of transcription data with regard to the Paradata itself. Vast amounts of data that, certainly, are going to go beyond our individual abilities to be able to make sense out of them and there might be computer algorithms that actually can do a better job in terms of classifying these various patterns. And this can lead, if you will, to adaptive, assisted instruments that are instruments in which interviewers are still engaged in some way, but which various probes are provided to interviewers as to how to deal with the particular respondent. They can be fairly static, such as the need to fill on a gap in a timeline. That's fairly static. Or they may be much more adaptive in the sense that you may have a history of a different kind of pattern of behaviors that provided indication that a certain error is about to begin. And, hence, you may be able to prevent that error through some sort of intervention through the interviewer.

And, finally, the idea here is to develop an intelligent, self-administered instrument, a virtual interviewer instrument, if you will. That is an instrument that will replace interviewers by being able to examine the various data mining techniques and by looking at different classification of behaviors of data and how they predict data quality, to use an intelligent agent as an interviewer; in a sense, a virtual interviewer that can produce and tailor various ways of asking respondents questions in a manner in which the best data quality possible and the best memory possible can be derived from that respondent. And, hence, with regard to a self-administered, web based questionnaire, you really would have a virtual interviewer, if you will, guiding the respondent to maximize the cues that would be available in autobiographical memory. And so, both of those, we would think that we could engage in both of those types of activities in the future with both calendars and time diaries.

I hope that in this chapter I have been able to convey the importance of using temporal information to assist with remembering; the value of behavior coding and Paradata analyses. We can maximize these coming opportunities with self-administered questionnaires, especially by using smart instruments and eventually getting from just assisting interviewers to actually replacing them.

Robert F. Belli is Professor of Psychology at the University of Nebraska-Lincoln. He served as North American Editor of Applied Cognitive Psychology from 2004-2009. He received his PhD in experimental psychology from the University of New Hampshire in 1987. Dr. Belli's research interests focus on the role of memory in applied settings, and his published work includes research on autobiographical memory, eyewitness memory, and the role of memory processes in survey response. The content of this work focuses on false memories and methodologies that can improve memory accuracy. His current research is examiningSo, you get a sense that in terms of this calendar format, the nature of theconversation is one in which the language the electrophysiological correlates of suggestibility phenomena, and the conversational and memory processes that optimize the quality of retrospective survey reports. Teaching interests include courses on basic and applied cognitive psychology, and on the psychology of survey response.

67

Linking Survey Data to Official Government Records

Joseph W. Sakshaug

In this chapter I will summarize some research about linking survey data to official government records. First, I'll be reviewing the rationale for linkage. What are the advantages to it? What can we do with linked survey and official government records? And then, I'm going to briefly summarize some of the linkage techniques that I'm most aware of that have been used to link survey and government records. And I'll also discuss some of the practical issues involved in those techniques. And then, I'll conclude by identifying some research opportunities that I believe deserve scholarly attention and increased funding.

For those less familiar with the idea of record linkage it may be helpful to start with a broad definition of record linkage. At the most basic level we are joining together units from two or more data sources to produce a single merged data source. So, we're really linking rows and cases to increase the number of columns and variables. And the units that we may be linking could be persons or households. But, they could also be establishments and other types of units exist as well.

So, why do we link records? There are many advantages for linking records and I distinguish between methodological reasons and substantive reasons for why we link to official records. We do this methodologically to check the accuracy and reliability of survey self-reports, so we can quantify the

J.W. Sakshaug (✉)
Institute for Employment Research, Nuremberg, Germany
e-mail: joe.sakshaug@manchester.ac.uk

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_67

measurement error associated with these self-reports. At the University of Michigan, we've used administrative data and official government records to assess things like non-response bias and try to get a sense of the quality of the survey data we collect and how it can be affected by non-response. In the substantive domain, we also link these data sources. It permits us to conduct more comprehensive longitudinal analyses. Many of these official government databases are collected and assembled over a long period of time. So, we have quite a bit of information over a range of time for our survey respondents, which can be very valuable for particular types of analyses.

Also, it allows us to investigate and try to answer some very tough policy-oriented questions and if you do a literature review of studies that link survey and government records, you'll find examples of this. Healthcare spending among older populations is a very popular topic and linked data sources are commonly used to address that question. Also, we use these linked data sources to address things like lifetime earnings and retirement planning. These are two popular domains for using linked data. Some administrative databases that are quite popular include social security records and these records contain a detailed history of earnings and benefit receipts. Medicare claims is often used in the health domain where we look at things like Medicare enrollment and try to get a sense of the detailed healthcare expenditures that these Medicare beneficiaries have.

Another data source is the National Death Index, where death certificate records are collected from State Vital Statistic Offices and can be linked through the National Center for Health Statistics (NCHS). These records could be collected in different ways from the states and have different levels of quality associated with them. These data are available to link to particular types of survey data housed within the NCHS.

There are three main record linkage approaches that I'm most familiar with and that I've seen being used for linking to official government records. The first one is exact linkage and this is where we have a unique key or a unique identifier that we can use to directly link from the survey data to the administrative data. The next one is probabilistic linkage where we don't have a unique identifier, but we have other identifiers that could potentially help us to link these data sources. And the next record linkage approach is statistical matching and this is where we may not have high-quality personal identifiers that are common to both data sources. So, we try to predict the similarity between two records belonging to completely disparate datasets and try to merge records belonging to different units.

I'll go over each of these different approaches in more detail. The first is exact linkage. This is the method where we do have a unique identifier, such as a social security number or a Medicare ID number where we can use this unique key to bridge multiple data sources and records belonging to the same unit. And respondents usually provide this unique ID. We ask them to provide it in our surveys. Of course, it's optional and it's up to them whether or not they provide that. Another thing we do with exact linkage is we ask for consent from respondents. First, we ask them for consent and then we ask them for their unique ID, like a social security number, conditional on whether they consent. And interviewers in face-to-face surveys or telephone surveys usually administer the consent request. The interviewer will ask respondents whether or not they consent to a data linkage. And in a face to face survey, for example, this might consist of giving a paper form to the respondent and the paper form might provide a description of the purpose of the linkage and some of the potential risks and benefits involved in linking those data and also the safeguards that are put in place to protect the confidentiality of the official records.

As you can probably guess, there are some practical issues associated with exact linkage and the first one is that linkage consent is not universal. Not everybody consents to link their records for many reasons – confidentiality concerns, a distrust of the survey organization and so forth. I've done a little bit of work on looking at consent rates across different studies and across different administrative targets. And we find quite a bit of variability in the consent rates. Consent rates can be quite high for some studies and some administrative data types, but can be quite low for others. One of the implications of non-consent in the survey context is that it can lead to reduced sample sizes and efficiency loss. The linked analytic sample that we have is less than the total number of respondents in our sample because not all respondents consent to the linkage. And this can also lead to biased inferences, especially if those who consent to the data linkage are systematically different from those who don't consent based on key variables of interest. There's a concern in the literature that consent rates won't always be high and, in fact, they could experience the same drop as response rates have in surveys worldwide. And so, this is an important issue that's gotten more attention in the literature regarding looking at potential biases associated with non-consent and determining whether or not we should increase resources to study and try to correct that issue.

The second practical issue associated with exact linkage is that there could be mismatches in the matching variable between the survey and administrative records. For example, consent could be provided, but establishing a record link may not be possible because we have an incorrect, or a partial, or perhaps no identifier provided by the respondent. And so, that makes the linkage task more challenging. And some respondents, recent immigrants, for example, may not yet have an official record, so we have really nothing to match to it. I've examined consent rates for a couple of different studies in different countries, and the administrative data that they link to surveys range from social security records, Medicare records, health records, and hospital or tax records. The consent rate for these studies show a little bit of variability and some of them are quite low. The British Household Panel Survey, for example, a popular U.K. study, experienced a consent rate of 41 percent in a recent wave of data collection. Some of these consent rates can vary, depending on which administrative data type they're asking for and also, whether or not they're asking to link children's records or adult records and those can experience different consent rates as well. I wouldn't call these consent rates particularly high. Some of these consent rates, sort of, mirror response rates for many surveys and so that's a concern.

The next record linkage technique that I'm going to be talking about is probabilistic record linkage and this is the technique that we've been hearing about most often in the presentations so far. Not all databases contain a unique ID key and even in the cases where we obtain consent, maybe we don't get a social security number, we have to rely on another technique besides exact record linkage to try and link to their official record. In this case, we do have some identifiers, such as surnames, given names, date of birth and address information that we could, potentially, use to estimate the probability that two records belong to the same unit and that's how we implement this approach. The match status – whether we can determine whether a match is given or not – is determined by a pre-specified threshold. So, essentially, if the probability of two records belonging to the same unit is above a certain threshold, then we would call that a match. And we might have a lower bound; if that probability is below a certain lower threshold, we would call that a non-match. And any probabilities in between we would be uncertain about. They could be probable matches, but they would require further review. And there are many software packages available that do this for you such as LinkPlus and Big Match, which is a software package used for linkage that was developed by the Census Bureau.

There are a few practical issues associated with probabilistic linkage and one is that it's difficult to estimate the frequency of false matches and false non-matches. And it also really depends on what threshold we use to determine whether or not we can call it a match or whether we can call it a non-match. And, as we know, there are quality issues too. The matching variables themselves have varying levels of quality and missing data, which can adversely affect the quality of the match. Another issue with probabilistic linkage is that it can be a complex task to link more than two data sources. And there's no well-developed solution for linking many data sources together. Chaining is one common approach. We might order the datasets in terms of their reliability or quality and start with the highest quality datasets and then chain them together in that particular order. But, it's unclear whether or not this is the most optimal method of linking these data sources together.

Statistical matching is yet another technique that is used when exact linkage or even probabilistic linkage is not really possible due to confidentiality restrictions and/or lack of high-quality identifiers. And the basic idea behind statistical matching is to try to estimate relationships between variables that are never jointly observed. That is, we're essentially linking records belonging to entirely different units. The one thing needed to implement this approach is common variables between both datasets and some examples might be age, sex, race, ethnicity, and/or other socio-demographic variables. And we use these common variables to try to identify and link records that are statistically similar based on some kind of quantitative metric. And different metrics are used to identify similar records, Euclidean distance, predictive mean matching and propensity score algorithms are often used for this purpose. There are a lot of issues associated with statistical matching. One is that it makes very strong mathematical assumptions, particularly about the conditional independence of Y and Z. If you have Y from dataset one and Z from dataset two, it assumes that these variables are conditionally independent given the common variables X. And this assumption is very difficult to test with the observed data. Another issue is, sometimes, there are very few common variables between these datasets and/or weak associations with these common variables and the target variables that we're most interested in and this can create problems as well when we implement statistical matching.

There are other time consuming tasks like harmonizing common variables between the data sources. There could be qualitative and quantitative variables that need to be harmonized in order to implement this matching approach. And I would argue that there needs to be more evaluations to determine whether statistically matched records reflect the true relationships in the population. I think there's much more skepticism in terms of this statistical matching approach relative to the other approaches that I've been discussing.

Now, I'm going to switch to focus on some general areas of research that I think deserve funding in the context of data linkage. First, I think it's important to really dig deep and investigate the properties of these different linkage techniques and really try to get a sense of their strengths and weaknesses. One question is how linkage errors affect subsequent statistical analyses for these different methods and can the strengths of one approach be used to overcome weaknesses in another? I've already given examples where, in exact and probabilistic linkage, consent from respondents to link their administrative records is needed.

With statistical matching, consent is not really an issue since we're not attempting to link the respondent's record in the administrative data; we're linking them to a different record in the administrative data. And so, maybe we don't have to ask for consent to do that from respondents. Thus, if the future shows that consent rates are dismal, if they are really low so that exact linkage is less practical, perhaps, statistical matching may be a way to overcome that limitation. But, yet, statistical matching itself has its own limitations so we really need to assess the tradeoffs of these different methods.

And that leads to the next question, how low do consent rates really need to get before alternative and non-exact linkage approaches should be considered and used in practice? And how do we balance the tradeoffs between data utility and data confidentiality in the context of record linkage? We're in the business of trying to create higher quality datasets and I think record linkage is an important tool to try to do that. But, as you link more datasets together, the probability of re-identifying respondents increases, so how do we balance the tradeoff between utility and data confidentiality? I think there's really no well-developed solution yet for how to handle this tradeoff.

Next, I think it's important to really think deeply about the theoretical framework for linkage consent. What are the mechanisms that drive respondents to consent or not? There's been a lot of work done to identify mechanisms of survey participation and survey response. But, I think different mechanisms are at play here when we think about consent to link to administrative records. We're dealing with a different population of respondents: we're only dealing with the respondents here. And so, I think more attention needs to be turned toward identifying those mechanisms of consent for respondents. On the other hand, there could be theories that, perhaps, we can borrow from the survey participation literature, particularly leverage-salience theory or these foot-in-the-door techniques. Surveys provide a rich source of survey variables that can be used to study linkage consent. We have extensive information that can be used to study linkage consent, often more than we have to study survey response. And so, maybe we can leverage survey variables to try to get at the mechanisms of consent.

It would be useful to study the theory and framework of linkage consent, but then we have to ask ourselves, "Well, how can these known mechanisms of consent be operationalized to increase consent rates in practice?" Frauke Kreuter and I have a paper where we cited a lot of this foot-in-the-door literature and we thought, you know, maybe it's better to ask the consent question upfront at the beginning of the survey as opposed to the end. The overwhelming majority of studies that we've seen always ask the consent question at the end of the questionnaire. Right as they're wrapping up the survey, interviewers ask respondents for consent to link their records. And we found that actually asking upfront yields higher consent rates than asking at the end of the interview. We think there are opportunities to use this theory to help us to design more effective consent protocols and try to increase consent rates. And I think we really need to ask ourselves how we can effectively address confidentiality concerns because this is the most often reason cited in the literature for non-consents. And so, how we can address these confidentiality concerns is a big issue and there isn't an easy answer for it, but I think it deserves more attention and more research.

Interviewers play a very important role in all of this. They are the ones that are intimately involved in this consent process. They're the ones who administer the consent request to respondents. They address concerns that are raised by the respondents and they're the ones that ultimately collect the unique identifiers that the respondents provide. The interviewer effects literature is much more developed for survey response, as opposed to linkage consents. But, there's some really exciting work being done right now in terms of looking at interviewer characteristics and how those influence the propensity of respondents to provide consent to linkage. One question that I wonder about is "Are interviewers incentivized to obtain linkage consent in the same way they are to recruit respondents? Is linkage consent part of their performance criteria in the same way as response rates or cooperation rates would be when we evaluate these interviewers?" We find that interviewer-level consent rates exhibit large variability. Some interviewers are very good at obtaining consent from respondents and others not so much.

Another question is how do interviewer characteristics, attitudes and expectations influence respondent consents? And some of the studies we're working on in Germany right now, we've tried to collect detailed information about interviewers. We've tried to collect information about their attitudes toward privacy and data sharing and try to correlate these with their performance in the field in terms of getting consent from respondents.

And we find a correlation. We find that interviewers who are more likely to consent to hypothetical linkage requests and data sharing requests are more likely to obtain linkage consent from respondents. And I think this makes sense because interviewers who would have trouble consenting to the same data requests that the respondents are presented with are not going to be very good at selling this linkage idea to respondents. So, the question that we're trying to grapple with right now is, how can we use this information to enhance interviewer training, to improve interviewer performance in getting consent from respondents?

Another area for further research that I'm very interested in is identifying and adjusting for this consent bias. First off, we need to develop tools for how to assess bias in the first place and then, think about how we can correct for it, if it exists. Probably, the most common technique for identifying consent bias in surveys is to compare the characteristics of those who consent versus those who don't consent based on the survey data. And, as I said, we have a lot of information about our respondents based on the information that we've collected from them in the interview. Thus, we have a lot of information that we can compare among those who consent and those who don't consent to the linkage.

But, how can we assess bias for estimates obtained from the administrative data? Usually, we don't have administrative data for those who don't consent to the linkage. Typically, this information is not provided in the case of Social Security records. Social Security records are not provided to the data collection agency for those who don't consent to the linkage. So, it's really difficult to assess bias for estimates that are based on the administrative data themselves. And so, I think we need more sophisticated tools for bias analysis. We have gotten around this problem of estimating biases for administrative data by linking the consent indicator to the administrative data. We had administrative records of both respondents and non-respondents and, by linking the consent indicator to those data, we could determine which of those records were associated with consenters and non-consenters. Then, we can estimate different statistics based on the administrative data and assess consent bias. I don't think that study would be possible here in the States. But, one thing I think administrative agencies could help us with is to provide some information about the non-linked cases. For example, maybe they could provide aggregate administrative estimates for the linked cases and for the non-linked or the population units so we can see how these consenting respondents compare to the rest of the population that we're interested in. And, perhaps, we could relay such information to data users in order to give them a sense of the biases associated with the data that they plan on using. Additionally, an open question is, is it possible to incorporate such information into weighting adjustments and other bias adjustment procedures without compromising data confidentiality? I think that's an important area to look into.

There has been a lot of discussion in the literature and field about the quality of administrative data and I'm just as skeptical as everyone else in terms of whether these data should really be treated as a gold standard or not. I think we need more innovative methods of assessing the quality of administrative data and we need to develop better metrics and promote these metrics in order to give data users a sense of the quality associated with these data. And we could do this with quantitative metrics, as well as qualitative metrics, missing data rates, measurement discrepancies, the timeliness and the transparency regarding the data collection methods and the methods that were used to assemble the administrative data are all quality metrics that I think would be useful for data users to know about. And how can these quality profiles be relayed to users to prevent adverse data uses? In some cases, we might want to inform data users that the survey reports, in fact, may be more accurate than administrative reports.

Another area that I think is really important, and I touched on this briefly in the section about probabilistic linkage, is linking many data sources together, linking three or more data sources simultaneously is an imperfect art. And I said the most common method is this chaining approach based on starting with the most reliable dataset and working your way down, but I'm not sure this is the most optimal method of linking many data sources together. So, I think there needs to be more evaluations on how multi-way linkage errors affect subsequent statistical analyses.

I'll conclude with an idea that was mentioned by my colleague, Frauke Kreuter, which is a different way of thinking about the use of administrative data. Currently, we use administrative data to supplement our survey data and the administrative data acquisition is treated as secondary to the main survey data collection. But, maybe, there would be benefits to reversing this approach. So, the idea would be, starting with the administrative data and designing our survey around it. And I think this approach could be promising in terms reducing data collection costs and respondent burden. And this approach may be advantageous in terms of creating more efficient survey designs and using administrative data more effectively. I also think this approach could post a greater spotlight on administrative data and promote greater transparency in how these data are collected, which I think is important for all of us to know. Lastly, this approach could, potentially, expand opportunities for scientific research.

Joseph W. Sakshaug is a Senior Lecturer in Social Statistics in the School of Social Sciences at the University of Manchester (UK), a Senior Researcher in the Department of Statistical Methods at the Institute for Employment Research (Germany), an Adjunct Research Assistant Professor in the Survey Research Center at the Institute for Social Research at the University of Michigan, and a faculty member in the International Program in Survey and Data Science offered through the University of Mannheim and the Joint Program in Survey Methodology. He received his PhD and MS degrees in Survey Methodology from the University of Michigan and a BA in Mathematics from the University of Washington. He conducts research on record linkage, non-response and measurement errors in surveys, small area estimation, and statistical disclosure control.

68

Linking Knowledge Networks Web Panel Data with External Data

Josh Pasek

Unsurprisingly, given the small extent to which the field of survey methodology has engaged the issue of linking datasets, there hasn't been a ton of work done trying to assess exactly what we end up with when we merge together these big datasets and survey data and start looking at data quality. What I'm going to discuss in this chapter is an initial cut at assessing accuracy and bias in the relations between consumer file marketing data and what we end up getting from individuals recruited from the GfK KnowledgePanel. But first I want to review a couple of the challenges and opportunities I think we're dealing with as a field because they're a good context in which to consider how new data sources might help us move forward.

Challenges and Opportunities

One of the central challenges for contemporary survey research is declining response rates. In Scott Keeter's chapters in this volume you can read about the numbers from the Pew Research Center that illustrate declines in survey response they have observed over the last few years. We have also seen a number of other reports that note that response rates have been going down over time. As a field, we have experienced increasing costs across a variety of

J. Pasek (✉)
University of Michigan, Ann Arbor, United States
e-mail: jpasek@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_68

different modes for various reasons; these include refusals and the increasing difficulty of reaching particular populations. Further, for some survey modes we are encountering additional coverage challenges; that is, it is becoming increasingly difficult to reach Hispanic and young Americans, in particular. Of course, if we want to have truly representative samples, we need to ensure that surveys capture these populations. And finally, methodological issues such as accounting for who has telephone access and determining how best to combine landline and cell phone responses have also raised costs and introduced additional challenges into what was, for a while at least, a pretty steady paradigm of Random Digit Dial surveying.

Because of these various challenges, scholars are finding it increasingly difficult to translate from the respondents that we reach at the end of the day back to the population of interest. And this is exacerbated by the lack of a strong statistical theory that links our sampled population to society when response rates are as low as nine percent. Those who have examined the implications of these trends note that these developments have not yet resulted in large biases, but there are many reasons to worry that such biases will eventually emerge. If and when this happens, it is unclear what should be done to maintain the set of inferences that surveys offer us about society.

At the same time, as a discipline, we are encountering a number of opportunities that could serve to mitigate the gathering storm. In particular, there are opportunities emerging from the new forms of data discussed in this volume: things like social media data, mobile phone data, tracking data, and marketing data. These new kinds of data sources are often collected at the individual level and might, therefore, tell us about some of the people that we are trying to reach with our studies. Simultaneously, we have also developed a reportoire of new modes of data collection and more sophisticated analytical tools that could facilitate linking novel data with the survey measures we collect. In particular, researchers have highlighted the possibility of using Bayesian statistics and machine learning tools to link up datasets, but suffice it to say that there are a bunch of new tools for matching the people that we end up with at the end of our data collection process with the society we hope to describe.

Given these new tools and new sources of data, the question becomes one of what model are we going to use to make the link between the data that we are gathering and the population we hope to describe. So the big question in this environment becomes one of whether the opportunities that we are seeing can in some way offset the challenges induced by the increasing difficulty to translating between our respondents and the population. That is, can we use new methods for data collection and analysis to help yield an understanding of how the data we have collected relates to the population or is some other sort of data necessary for us to make those kinds of links.

The Potential Value of Consumer File Marketing Data

The study presented here compares survey responses and consumer file marketing data on the same individuals. By consumer file marketing data, I refer to the types of data that are generated by commercial firms and are typically purchased by companies for use in marketing products. The data for this study was purchased from Marketing Systems Group by GfK and it originally sourced from Experian, Axium, and InfoUSA; finding out where individual measures come from before that seems to be an impossible task. In this study, I assess what you can do with demographic information from consumer file marketing companies and how the distributions of these variables relate to those of survey responses. Although they are somewhat expensive, consumer file data can be easily purchased and readily matched to survey samples. The matching process is relatively straightforward because we have addresses linked to the postal service's *Computerized Delivery Sequence File* for both the marketing data and for our respondents. Because of this, Consumer file data provide a rich source of information about all individuals in our sample, not just for our respondents. So we are dealing not just with the individuals who responded to our survey, but we can also claim to know something about the full set of individuals in the sampling frame.

If these consumer file data are high quality, they should provide us with a number of valuable capabilities. First of all, they could be used to improve the efficiency of sampling. With these data, we could target hard to reach groups, for example, by oversampling Hispanic and young populations. If we conduct targeted oversamples based on these ancillary data, what we might end up spending far less money per respondent. This would allow us to produce less expensive surveys and we would be able to get questions in the field in a more efficient way. Another potential boon from these kinds of data stems from the ability to learn information about who survey non-respondents are and what features they share. If we can say something substantive about the individuals who aren't responding to our survey, we now have a much better sense of how much non-response bias we have.

Additionally, these data might allow for corrections for non-response and, potentially might enable us to identify, and thereby correct for, differences between a general population sample and perhaps some kind of biased sample such as an online non-probability sample. That is, we may even be able to correct for problems in the sampling frame. So the potentials here are enormous if we have consumer file data that correspond with traditional survey measures.

It is valuable to note that these techniques are not new. There is a long history of trying to connect auxiliary sources of data with survey measures. Much of our weighting toolkit comes out of this idea that we are linking our respondents to Current Population Survey (CPS) data or to some other known population benchmark. There is also an emerging literature on using individual-level non-survey data – things like the paradata that Frauke Kreuter's chapters in this volume cover – to correct for non-response. So there are a number of emerging tools that move in this direction and provide methods for connecting new sources of data with surveys and using these new data for a number of purposes.

Before we get too excited by these possibilities, however, I want to suggest that there are a bunch of preliminary questions we must ask. The first is: "What are we actually planning on doing with these data?" Are we planning on analyzing the data in and of itself, which is what the people are doing who are examining whether and when tweets seem to track an election? Are we looking at it as a way to supplement survey data, which is what Iexamine in this context? Or are we thinking of potentially other uses, such as for imputations or ways solving more complex problems? Our ability to garner insights from these data will depend in part on what we hope to achieve.

As a preliminary question for each of these goals, we must determine how accurate the data are in describing the units they claim to describe. Our first quest, then, concerns whether what we find out about from a piece of ancillary information is true, valid, and matches what we would learn from other sources of inference. Second, how complete are the data of interest? One of the things that I'm going to focus on in this chapter is the extent of missingness in ancillary data. Quite frequently we don't have full ancillary data on every individual. We should therefore determine whether the processes that produce missing data are ignorable or non-ignorable and thus how much we should worry about the completeness of the ancillary data.

The third component that we really need to consider is the model we are using to link our data with the world. Specifically, we need to conceptualize the new data that we're acquiring, whether it be a tweet, some piece of information we're buying from a company, or whatever else. Then we need to model the connections between the individuals whose information we have collected and the actual state of those individuals as well as the connections between sampled individuals and the population parameters we wish to describe. That is a difficult challenge, but one that we, as a discipline, need to address.

And finally, how do different models perform across different types of inference? Even once we've gone through the data, we've said, "Okay, we're comfortable using this data, we understand how they relate to our respondents to the world, we get this process," we still need to think about whether what we're concluding is indeed true and whether the procedures can be used for more than just point estimates or predicting an election. For example, if we are interested in understanding how variables relate to one-another, we need to know whether the new source of data we are using and the model we are using to connect these data to society will yield accurate inferences for these types of questions as well. Similarly, we want to know if the data and model can be used to understand trends over time or the influence of experimental interventions. So we have to really seriously consider the types of inference we hope to make.

Linking Consumer File and Survey Data

In the current project, I first describe some assessments of the correspondence between ancillary data from marketing databases and self-reports. This will reveal the extent to which we end up getting the same and different answers about individual demographics from these two types of data. Where they differ, I provide some data on how different those answers are. Second, I evaluate the nature of missingness in ancillary data and assess whether the ancillary data that was purchased seems to be missing at random in some way that we could understand or is instead non-ignorable. And third I explore whether correctives using ancillary data might give us a better sense of the entire sample, not just our respondents. To accomplish this last goal, I use a combination of self-reports and the ancillary data to impute information about the entire sample using multiple imputation and thereby estimate demographic information about non-respondents. Presumably, if the ancillary data do a good job at least of mapping the differences between our respondents and the full rest of the sample, this should generate an accurate portrait of the population (and can do so even if the ancillary data themselves aren't all that accurate). This analysis tells us how well the consumer file data perform at reconciling differences between respondents and the population.

The data that we are going to be using in this case is comprised of 25,000 households sampled by GfK from the USPS *Computerized Delivery Sequence*

File (CDSF). The CDSF has over 95 percent coverage of the US population. GfK uses an address-based sample recruited from the CDSF, sample recruitment was conducted by mail in January 2011. Respondents who needed Internet access were provided with Internet access and the results I'm going to be discussing are from the core adult profile, which is the first survey respondents completed upon admission to the panel. At the end of this process, we have self-report data from 4,472 individuals in 2,498 households that were successfully recruited to KnowledgePanel. That's an AAPOR RR1 of 10 percent. Ten percent of the households that were sought were actually recruited into the panel in some way, shape or form. Though we have more individuals from each of those households because GfK allows multiple individuals to be recruited into the panel per household.

The consumer file data come from Marketing Systems Group (MSG). It has been merged with all sampled households and there is a 100 percent match on the addresses. So because the ancillary data companies do indeed use the same computerized delivery sequence file data, we have matching results for all of the households in the sample, so we're not dealing with a case where we're having trouble matching the records. Instead, we are using the same frame to collect data with both methods.

We produced eight different sets of weights, comprised of four different weighting types for two different comparison groups. The data come from what I will call the "best ancillary match weight." For most intents and purposes this isn't too relevant. All sets of weights present the same story, but we're actually biasing these weights toward finding a match between the ancillary data and the respondents by selecting the respondent in each household who most closely matches the age recorded in the ancillary data.

One note about these ancillary data, and I'll talk about them in a moment, is that when they're purchased from MSG they are purchased on a household level. This means that you end up only with age and education information for a mysterious individual referred to as the "head of household." And we don't quite know what the "head of household" is, but what we did to try to find the head of household in the survey responses by choosing the individuals in the household who were closest in age to the number reflected in the ancillary data. This could also include non-respondents within a responding household, because all individuals gave proxy reports on additional household members. Matching to these individuals was our best available strategy for targeting the individuals the ancillary data may have been referencing. We also defined the weights either to match respondents or to match all sampled individuals depending on the outcome of interest. Weighting to match respondents was used for assessments of correspondence and missingness and matching all sampled individuals was used for all the analysis I'm going to use with multiple imputation.

Notably, none of what I just told you actually ends up mattering for any of the analyses. As for the measures we are going to use, there were only six variables in this case. Not that many were purchased here because purchasing them was somewhat expensive. In particular, we are looking at three household-level measures: whether you own or rent your home, household income, and household size, and three individual level measures about this mysterious "head of household:" marital status, education, and age.

Results

Correspondence. As a first metric, we considered correspondence between the survey and ancillary data for the same individuals. Our basic strategy was to assess the proportion of matches between ancillary data and self-reported data and also to look at how many pieces of data were relatively far off. In our comparison, the results indicate that the ancillary owners tended to match self-reported owners a little bit better than ancillary renters tended to match self-reported renters, but overall when it comes to home ownership we do pretty well. We have about 90 percent agreement between ancillary home ownership and self-reported home ownership.

The correspondence is not necessarily as good when we look at something like household income; in this case we looked at income categories and the difference between them. We took the income categories and subtracted the ancillary data category from the self-report category. What we found is that only 23 percent of households occupy in the same income category in the ancillary data as they do in the self-report data. And fully 44 percent of households are more than one category off. So weren't just off by a little bit, but were actually off by at least $10,000.00 (the smallest category) from their self-report. These discrepancies were relatively large. We end up finding similar discrepancies for a lot of our other variables. When we look at household size there's about a 30 percent agreement on the number of individuals in the household and about 32 percent who are off by not just one individual, but two or more.

For marital status, interestingly, in the ancillary data more of the people who are unmarried come across in the self-reports as married than come across as unmarried. So the ancillary data actually provides the wrong answer for most people who self-report as unmarried. Overwhelmingly in the ancillary data people seem to come across as married for some reason. But again since these are a black box we don't quite know why. Overall the variable has pretty decent agreement, but this is driven principally by married individuals.

Looking at education we again see around 40 percent agreement. This is not miserable, but we still have 20 percent of cases that are off by more than one education category. For age, recall that this is the variable where we bias toward agreement through our selection of household members and we further bias things toward agreement by considering it a match if respondents are within one year because maybe there might be a lag in the amount of time it takes to get the ancillary data. Age estimates agree between the ancillary data and the self-report 70 percent of the time, but we still have almost 20 percent of our data where no one in the household is within 5 years of the age provided in the consumer file data.

Across these measures, there seem to be varying levels of correspondence between the ancillary data and the self-report data but considerable discrepancies were apparent for all of the variables we considered. If we are thinking about consumer file data as something that should give us the same answer as what we' would get using other methods, this is probably not going to be the case.

It is important to note that we did not get a name from the ancillary data. In this case we're aggregating at the household level so this is all address matched. When we are estimating values for individuals it is for the individual we think is most likely to be the head of household. Although it is possible to get a name from the list of telephone numbers, we did not do so. This was not done because including the name changes the sampling units in problematic ways (we would be sampling individuals rather than households) and it may bias the process when multiple individuals are recruited. As noted before, we attempt to circumvent this problem by seeking the person that most resembled the ancillary information we had for the household. We are biasing the individuals we select as best we can toward a match given the data we have. We might be able to improve on individual matches by purchasing additional consumer file data; and with more resources, we could get tons of additional data from these sources. Right now we are working with the kind of data that a survey organization might reasonably acquire and use.

Missingness

The second thing we wanted to do was to evaluate the nature missingness in the ancillary data. To begin an assessment of missingness, we wanted to know how often we did not have ancillary data for a particular variable about a household. I told you we had 100 percent match, so it wasn't that we were missing a record in the file for some individual, but what we ended up with was a sizable amount of item-level missing data for many variables.

In terms of missingness by ancillary data variable, we are missing a little bit less data for the household-level variables than we are for the individual-level variables. There's no variable for which we are not missing at least 6 percent of respondents and on ancillary age we are missing 28 percent of households. That means that 28 percent of households do not have an ancillary age that was attached to them. Again, we could probably buy other datasets that would have another ancillary age measure that we could merge into the file, but this is what we end up finding using an already aggregated dataset across a number of different variables. Looking at the distribution of these variables, it is not the case that a small subset of individuals are missing data on many measures. Instead, for about half of cases, none of the ancillary information is missing and for some cases most of it is missing. Notably, many cases are missing one or two ancillary variables.

If we start looking at when ancillary measures are missing, we find that missingness is likely non-ignorable and is often related to self-reported status for those same variable. If we look at missingness on homeownership, renters have almost a 30 percent chance of missing ancillary home ownership information. Owners, in contrast, only have a 7 percent chance of missing ancillary home ownership. For household income, we find that individuals with incomes lower on the income bracket are missing around a little more than 10 percent of ancillary income status on average, whereas under 3 percent of ancillary measures are missing for those at the high end of the income bracket. Ancillary household size information is more often missing for households that self-report a smaller number of residents.

Among individual variables, ancillary marital status data were missing far more frequently for individuals who self-reported as not married than for those who identified as married. For education there was not much differential missingness across self-report categories. Although it is not significant, there is some suggestion that we might actually be missing a little bit more ancillary data for respondents who report the highest levels of education. Finally, when examining respondent age, we are overwhelmingly missing ancillary data for younger people.

How well can we predict when ancillary data will be missing? Well if we take the self-reported measures together and predict the number of missing variables we end up finding a handful of significant predictors: non-homeowners seem to be more likely to be missing information along with households with fewer people, unmarried, and younger individuals. Overall, however, we are not doing a great job of predicting when ancillary data might be missing. At this point, it does not seem to understand when these data will be missing although the pattern seems to be non-ignorable.

Imputation

From what I covered so far, ancillary data do not appear to be reliable given these assessments of correspondence and missingness. But perhaps if we use a Bayesian brute force tool we can get around these inaccuracies and can still use the data to understand our non-respondents. Perhaps the cross-cutting nature of the ancillary data still tell us something substantive. So, for the final analyses, we imputed the distribution of self-reports for all sampled individuals – not just our respondents – based on the ancillary data. We wanted to see sort if this approach would provide a substantive improvement over the self-reports of respondents alone and how much of an improvement we would see.

In terms of the basic procedure, we imputed self-report answers for the whole sample (all 25,000 cases) and then compared the results of these imputations with raw, unweighted, self-reports, the ancillary values, and the CPS. We included a series of extra measures in the imputations because adding additional variables should generally improve the quality of the imputed results and should not bias them.

The results come from 100 imputations using multiple imputation via chained equations (mice), which is one of the best imputation methods available. We use these to produce point estimates for what all sampled individuals might look like. We don't want to use weights because we are looking at how the demographics in our sample match what would be expected from society.

Looking at home ownership, which was one of the variables that was pretty accurate in the earlier analyses, we find that we do about equally well using the raw GfK estimate as we did with the imputed data but that the ancillary estimate was far from the CPS. Looking across a number of the analyses, we frequently find that the raw GfK and the imputations tend to be the closest datasets to the CPS estimates. The imputations end up doing a little bit better on average than other methods and the ancillary data themselves, if treated as an estimate of the population, are frequently further from the population benchmarks. This is true with regard to household size and with regard to marital status. For marital status, the imputations were particularly close to the CPS estimates. When we look at education, we find that the imputations were sometimes very close to the benchmarks and other times were further off. Age was the same story, across all of these.

The raw GfK data compared to the imputations were adjusted to correct for variations in sampling likelihood due to household size as well as oversamples of projected hispanicss, and young people, but no poststratification was applied. There was a differential probability of sampling whereby GfK used some of the ancillary data on Hispanics and young people to supplement its initial pool with additional members of these target groups, and individuals who were predicted to be in these groups using the ancillary data were down-weighted as a corrective. We also weighted all data to the household level. Those two weights are on the data that I am referring to as "raw" right now, but nothing has been done in an attempt to match the population, just an attempt to correct for different sampling probabilities.

Across all of the analyses we find that the imputations tended to perform the best in matching benchmarks, but not by an enormous margin. Raw estimates were sometimes moderately off but never enormously off. The ancillary estimates were often far off, particularly in the case of marital status, and were almost always much further from benchmarks than the other variables. On average, we find that imputations performed the best, though the difference between the imputations and the CPS was still more than half of the discrepancies in the raw data that used no correctives to try to address differential and non-response across groups. How this should be interpreted is an open question, but it appears that the six variables we used were not able to eliminate most of the error that distinguished respondents from the rest of the sample. This result is perhaps most clear when looking at the total. On average, the raw self-reports, which were not corrected to match the population, differed by about five percentage points from the CPS. The imputed data differed by about three percentage points on average. So we get rid of some of the error but less than half of that difference. I should note that the self-reported variables were collected in a manner that is almost identical to the CPS, using similar household-level sampling and allowing for proxy reporting by other household members.

So what should we make of this? First of all, estimates from the raw self-report data were not all that far from the CPS. Second, imputations based on the ancillary data did eliminate some portion of the discrepancies between the self-reports and the CPS. And finally, the ancillary data themselves are just not something you'd want to rely on as an estimate of the population. Taking all these analyses together, we find that there are frequently discrepancies between the estimates of the ancillary data and self-reports. Missing ancillary data seems to be systematic and it appears to be non-ignorable. And a standard set of Bayesian imputation algorithms don't fully correct for the biases even though they do seem to improve things somewhat.

Discussion

What does this suggest about using consumer file data and where we are at the moment? Well these data may be useful for accessing hard-to-reach populations. It was definitely true that there were more young people that were identified as young people in the ancillary data than young people that were identified as older people. This type of approach might thereby provide an improvement in sampling efficiency. Notably, there's a distinct bias-variance tradeoff with any given variable that researchers might want to approach with this strategy. The ancillary data also do not seem particularly efficient at correcting for non-response. This, in turn, makes it unlikely that we can use consumer file data to correct for a bad sampling frame.

Overall, leveraging this sort of ancillary data seems of limited utility; at least with the number of variables we have here. The question, then, is one of why these data did not seem particularly accurate or useful. We lack the information we would need to assess this given that the ancillary data themselves are a black box. We don't know what is actually being measured, what these firms are imputing, or how data are linked across various sources. This is the critical problem we're dealing with. Because we are looking at what comes out of the black box and not what goes into it, we don't really know where biases emerge or how they shape what we end up measuring by the end of this process.

Moving forward from here, I think there are a lot of reasons that we do want to encourage good ancillary and auxiliary data of these sorts as both an improvement to survey sampling and as an additional tool we can use. To date, however, the kinds of demographic ancillary data that we used in this study were not sufficient for many of these purposes. They didn't seem to be high enough quality and, given the quality of the data we had, we would have needed many more variables to improve our estimates of non-response and to address coverage errors.

With the current data, we should seek to understand how the results of correctives using these data compare with those of traditional weighting techniques. We also want to be able to assess whether using a larger set of ancillary measures might allow us to use brute force methods to reach more accurate conclusions. Perhaps we can think about linking even more types of data together by seeking additional data sources. In my view, the most valuable contribution is that it gets us to think about how to get data we can trust and evaluate. To do that we need more transparently generated ancillary data.

The process of linking data sources to one another needs to be more systematically addressed. It's something we have to deal with if we want to start incorporating these additional sources. Perhaps academicians want to be focusing on an in-house option, on something that we can do ourselves instead of purchasing these data from corporations that aren't going to tell us quite how they produced the data. We should start thinking about whether university researchers can build a dataset that allows us to examine how different sources of data can be linked so we can understand what we can do with them. Funding agencies like the NSF can play a pivotal role in putting a dataset like that together, which would be my strong suggestion. When we are considering additional sources of data there are some central questions that we need to consider regarding what we plan to do with the data, how accurate they are, how complete they are, the models that we are thinking about using to link the data with society, and how those models will perform for different types of inference. These will be key questions moving forward.

Josh Pasek is Assistant Professor of Communication Studies, Faculty Associate in the Center for Political Studies, and Core Faculty for the Michigan Institute for Data Science at the University of Michigan. His research explores how new media and psychological processes each shape political attitudes, public opinion, and political behaviors. Josh also examines issues in the measurement of public opinion including techniques for reducing measurement error and improving population inferences. Current research explores the origins and effects of both accurate and inaccurate political beliefs and assesses the conditions under which non-probability samples, such as those obtained from big data methods or samples of Internet volunteers can lead to conclusions similar to those of traditional probability samples. His work has been published in *Public Opinion Quarterly, Communication Research*, and the *Journal of Communication* among other outlets. He also maintains R packages for producing survey weights (anesrake) and analyzing weighted survey data (weights).

69

History and Promise and Blending Survey Data with Government Records on Turnout

Michael P. McDonald

Although the 2008 election was the public watershed moment for data analytics, campaign consultants quietly began building their national voter registration data infrastructure following the passage of the 2002 Help America Vote Act (HAVA), which mandated states create statewide voter registration databases to assist managing their voter rolls. Historically, local election officials managed voter registration through a highly-decentralized system. In the mid-1990s, when I was engaged as a California campaign consultant, consultants were dispatched to local election offices to obtain copies of voter files to build a statewide file. Collecting these data on a national scale was prohibitively expensive, until concurrent information technology and government policy innovations in the early 2000s launched the era of big election data. In this information era, commercial vendors – such as Aristotle, Catalist, L2, NationBuilder, and TargetSmart – collect and sell their data, which they may enhance with proprietary algorithms.

There are many purposes for voter registration data. Perhaps the most traditional is to provide election survey samples. Most states require eligible individuals to register some period of days in advance of an election if they wish to vote. (This is not strictly true since some states allow unregistered voters to register and vote on the same day.) Registered voters thus reasonably represent the universe of eligible voters in many states. Survey

M.P. McDonald (✉)
Department of Political Science, University of Florida, Gainesville, USA
e-mail: michael.mcdonald@ufl.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_69

organizations seeking to understand campaign dynamics often draw samples from voter registration lists, what is known as a registration-based sample (RBS). YouGov has innovated RBS further through propensity matching of non-probability internet samples to samples of registered voters. Survey organizations that conduct random digit dialing surveys now have a similar capacity as RBS to match respondents with their voter registration record, following the administration of the survey.

A piece of voter registration data of critical interest to scholars and practitioners is voters' prior history of voting. Election officials track this information primarily for compliance with federal and state voter list management requirements regarding how long voters remain registered, if they do not vote. A greater share of survey respondents typically report voting than what actual election results indicate; this phenomenon is so well-known that a word has been coined for it: "over-report bias." Many researchers generally consider voter registration vote history superior to self-reported vote since the former is an administrative record of an individual's validated vote that is not contaminated by over-report bias. Academic surveys, such as the Cooperative Congressional Election Study (CCES), integrate voter registration data with their survey data so that researchers can produce estimates for self-reported voters, and for vote-validated voters. Survey organizations employing RBS samples have innovated on the concept of vote validation further by developing likely voter models that weigh how often registrants have voted in past elections.

Scholars and practitioners also use voter file big election data to analyze voting behavior, entirely independent of surveys. One vein of research involves simply analyzing characteristics of registered voters and non-voters with correlates that may be found on voter registration files, such as age and gender, or modeled from the geographic location and surnames, such as ethnicity and race (in most states). Another vein involves experiments. A typical experiment involves randomly assigning registered voters into control and treatment groups; providing a stimulus to the treatment group, such as a voter mobilization contact; and observing their recorded vote history. This approach was popularized in a seminal 1998 New Haven election study by Donald Green and Alan Gerber, but its roots extend back to at least a study of the 1924 Chicago election by Harold Gosnell.

Scholars are attracted to voter file data due to their increasing accessibility and a growing body of research that validates the efficacy of their academic worth. Based on my thirty years of working with voter registration data and speaking with election officials about their data management practices, the rapid diffusion of research on voter files has outpaced critical assessment of their efficacy. Potential biases lurk in how election officials maintain election management systems, how campaign companies enhance voter file data, and by how voter registration data are matched to external data sources.

The 2002 HAVA required states to produce statewide voter registration file, but it did not require states to create a centrally-administered election management system. In the election administration world, states with centralized election management systems are known as "top-down" states, while states with decentralized election management systems as known as "bottom-up" states. In top-down states, the state provides a statewide election management system that local election officials use. Voter registration information is keyed directly into these systems. In bottom-up states, local election officials have locally managed systems, which may have been purchased from a commercial vendor or are home-grown. Some states have hybrid systems, where some localities participate in a state-provided system and some have their own. To be HAVA-compliant in bottom-up or hybrid states, localities transmit their voter data to a state-administered repository.

Why these election management details matter is that, surprisingly, there exists no canonical national voter file. A national file is an aggregate of state and local voter files. Timing of the data collection matters: I have shown by the time election officials have entered vote history for a recent election, some voters have moved and already been purged from the voter file. A result is that the total number of ballot tallied for an election in an election typically exceeds the total votes recorded on a voter registration file. This is not true everywhere, as some states maintain snapshots of their voter file as it existed for a given election, but this is not standard operating procedure everywhere. Furthermore, I am aware that in at least one state the voter registration records – and associated vote history – can be duplicated when voters move across counties within a bottom-up state between the time the first and second county transmits their data to the HAVA-compliant statewide database, leading to more registered voters with a vote history than ballots counted in a county. Thus, the theoretical canonical snapshot of the electorate is in practice more blurred than one may think. This blurriness increases with time as more people move and election officials purge their voter registration records, or transfer them to a new county.

Researchers must further assume that voter registration data and associated vote history are entered without error, or at least that the errors are random. Anecdotal stories drawn from allegations of vote fraud abound of a cast ballot being associated with the wrong voter, often as a consequence of a voter signing the wrong line on a poll book when they check in to vote, or similar error. Errors can also occur from simple fat-finger data entry errors by temporary workers tasked with the data entry. Election officials in supposed top-down states have shown me their desktop spreadsheets where they track their voters, eschewing their state's system because they did not like how it works. Data entry into election management systems for these election officials is a very low priority as it is considered duplicative busywork. Perhaps these errors are random, but no one has established, for example, that these errors are uncorrelated with the size or capacity of a local election office. Indeed, the recognition of these issues is not necessarily new. Until 1990, the American National Election Survey asked its face-to-face interviewers to visit local election offices to examine and record respondents' vote history. The ANES discontinued this exercise because they were uncertain as to the quality of the administrative records. Yet, these lessons appear to have been lost from our institutional memory, only to be rediscovered when researchers analyzing validated vote data produce surprising estimates.

Once voter registration data are collected from election officials, scholars and practitioners who wish to work with these data are confronted by non-standard data formats and availability across states, and within localities among bottom-up and hybrid states. For example, some jurisdictions may provide a voter's exact birthdate, others the birth year only, others voters' ages, and still others do not release this information. In another example, a handful of Southern states that held White-only primaries in the past have race and Hispanic ethnicity on their voter files, but the voter registration form response items are not the same list as those asked by the Census Bureau. For states without race or ethnicity, consultants estimate this information from Census Bureau lists of the race and ethnicity frequency of last names, perhaps further adjusted for the racial and ethnic composition of a voter's neighborhood. To do this latter task, a consultant must be able to geocode voter file addresses, which is deeply challenged in rural America where rural mail routes are used in place of numbered addresses. As with vote history, no scholar to my knowledge has investigated the scope, direction, and magnitude of any errors that arise from these data practices.

Voter registration rolls are notoriously inflated with records of people who are no longer living at their listed address. Election officials routinely purge this "deadwood" from their voter registration lists. Two notable consortiums of states – the Electronic Registration Information Center and Crosscheck – match members' voter files across states to identify movers, and also match against states' driver's license databases. However, the National Voter Registration Act of 1993 requires election officials to keep registrants' on voter rolls unless a voter notifies election officials they have moved, or if the voter has not participated in two consecutive federal general elections.

Political consulting firms do their own purging since their primary purpose is to have the most up-to-date voter registration file so that pollsters and campaigns can be assured that they are not wasting resources by contacting registrants. These commercial vendors match voter files with other data sources: the Post Office's National Change of Address database, commercial credit databases, other states' voter files, and through contacts that canvassers have with voters.

List matching is challenged on three accounts. The first is that false positive and negative matches happen with surprising frequency when millions of records are being assessed. By chance, it is possible for two people to have the exact same name and birth date, leading to a false match. The second arises from the quality of the data: most matching algorithms require the specific birth date to assure the lowest frequency of false positive and negative matches. If a state does not provide exact birthdates, these data may be obtained from commercial credit vendors, which increases the risk of false matches since obtaining these data involves list matching, too. Finally, there is the threat of data entry errors on the matching criteria: names, addresses, age, etc. The challenges of matching have been known since at least the 1950s, and while sophisticated fuzzy algorithms have been devised to account for name variants and misspelling, the issue remains that there is little guarantee two records necessary identify the same person, or that the same person in two databases can be identified.

There is a key piece of information that is required for phone surveys that deserves special attention: phone numbers. Some states request a registered voter to provide their phone number, but this may not be a requirement: response rates tend to be low, there is no guarantee the phone numbers are valid, and there is no guarantee numbers are missing completely at random. A survey organization conducting an RBS survey typically obtains an RBS sample from a list vendor that enhances their voter file with phone numbers from other data sources. The errors present in these external data sources and the biases that may arise from matching algorithms are unknown.

There are many potential pitfalls when working with voter registration data. To my knowledge no one has ever exhaustively determined the scope, magnitude, and direction of bias of the potential errors that I have outlined in this chapter. I believe a more critical assessment of the value of these data is warranted and researchers should exercise caution when drawing inferences from research based on voter registration data. Unfortunately, commercial voter registration list vendors treat their data and enhancement algorithms as proprietary. And of further unfortunate import is that many use these data uncritically in their research, and become uncritical proponents for their use.

From an academic perspective, these data are not replicable and therefore cannot reliably contribute to our body of knowledge. More work, I believe, is needed by academics to develop sandboxes to explore the reliability of voter registration data, enhancement algorithms, and matching algorithms; and for commercial vendors to be more transparent when their data are consumed by academics and the public. Academics and practitioners can thereby come to a better understanding of the potential magnitude and direction of bias of the various threats to inference when analyzing voter registration data.

Dr. Michael P. McDonald is Associate Professor of Political Science at the University of Florida, where he is affiliated with the University of Florida's Informatics Institute. His scholarship has appeared in top scholarly journals and he produces United States voter turnout statistics used widely by academia, the media, and policymakers. He has provided election-related consulting to ABC, NBC, the Associated Press, the United States Election Assistance Commission, the Federal Voting Assistance Program, and the media's national exit poll organization.

70

Improving Information Quality and Availability Through Interactions Between Government and Academic, and Industry Survey Research Sectors

Robert M. Groves

My chapter is much more general than many of the others in this volume. I want to begin by discussing about what I believe is unique to surveys within the social science context, and focusing about where we are right now and then what the implication for whatever we – maybe we won't even call them surveys but what survey like inquiries should be in my own opinions. My hope is that some of this will be provocative.

So what are the unique attributes of surveys first? I believe within the social science context they offer a couple of important attributes. One is uniform measurement, consistent measurement over the objects being measured. That's true of other forms of social science inquiry but not all forms of social science inquiry. They have traditionally been gloriously multivariate and that's really important I think because we have fed at the trough of multivariate nature of surveys for a long time. That multivariate nature has the wonderful offshoot that the existence of data archives in the social sciences based on surveys actually are useful for a long time by thousands and thousands of people and are useful years later when new combinations of variables become interesting to a sub group. So I'm a big fan of the multivariate nature. It is an environment where researchers control the data collection as opposed to someone else. We invent the measures. When

R.M. Groves (✉)
Georgetown University, Washington, D.C, USA
e-mail: bgroves@georgetown.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_70

we're good the measures are invented to measure the constructs that the research is all about and I like that.

And then I think the last thing which is often the first thing that people mention is that they enjoy a rather tight inferential paradigm statistically. We got the basic theory of this down for inference from the observations to a larger population. That's based on the existence of universal frames from which we sample the subset we measure and also the measurability of the missingness. So the fact that samples from universal frames also allow us some insight into what we're not observing that we'd like to observe. That together produces useful inferential frameworks.

And if you go back on this last point, historically the explosion of surveys and the use of surveys took place when you got the inferential paradigm right. Before that there were tons of things that sort of looked like surveys but their impact on societies weren't as large as they are now. The present state that we find ourselves in is a little chaotic. There's a lot of noise, intellectual noise around surveys going on now and threats. My hunch is if surveys are able to continue producing accurate estimates and predictions, the commentary will not going to be as loud as it could have been if things had turned out differently. But we have new resources. Some of the new resources are in the measurement side. We have devices and things and processes that are measuring phenomena that are kind of cool. The researcher in generally doesn't control them, they're controlled by somebody else. And they have a really cool attribute to them in contrast to surveys, and that is they're fast, some of them are near real time.

Anything Internet connected, anything process connected that's generating transaction level data generally are really timely. So this is a new world for us because the one beef about surveys is they're slow. They're slow mainly because of the inferential paradigm. We have to assemble all the measures on the sample before we can do the estimation. It's hard to do minute by minute surveys or second by second surveys. So I label these kinds of new data as organic data under the metaphor that we now have an information echo system and that ecological system is producing data about itself as part of itself and it's all over the place. Now one subset of that isn't real-time data but is valuable and that is all of the data that used to be written on forms and stored in file cabinets are now digitized. And if we could get our hands on those that might be neat. So that's like a special case I think of the organic data. It's always been there but it's been unusable by us because it hasn't been digitized enough.

Part of this world then forces attention at various things. One enormous problem that we have in this new world is that those new datasets aren't ours.

They're not in the public domain; private sector firms disproportionately hold them. Some of the firms have realized that if they're smart the data themselves become a revenue stream.

So I find us at a really interesting point in history in that what we thought of as social science might be completely privatized to the extent that you can make money on the kinds of statistical information that we generated out of the social sciences in our lifetimes. Most people are pretty old in this room. There are a few young people but we made our living as it were, we found our careers by this older paradigm where we invented the data, we held the data and then we gave it away to the public mainly because it was federally funded. It's a whole system that we built our careers on.

So what are the ingredients for change and what should National Science Foundation (NSF) do in this world? I have opinions on this. We're at the point where our current framework or what we do is undergoing unsustainable cost inflation given how we fund our work. I don't see a solution on that. At the same time the quantitative side of the social sciences or quantification of information if you will, has won the day societally. There are very few institutions in the society that aren't running on quantitative data now. So that kind of basic fight that the older ones of us in this field of survey research were part of fighting is just over. So that's a good thing. People are accustomed to informing their decisions based on quantitative information. That's good. But we have money squeezes all over the place.

So, I think there are some comments that we could make. First, let's comment on these new data resources that are intriguing and some of us are playing with. How are they different? Well I've said one thing already, they're near real time. This is really cool. It forces us to do the critical longitudinal thinking that we didn't have to do before and really forces us to ask questions about what is the value of timeliness in the whole equation of quality of information I think. That's a healthy thing for us to do. They are horribly univariate or near univariate. They're not rich data sources in general that's a real weakness.

So we can do interesting little single variable tracking which is kind of intriguing but we I know as soon as we see stuff like that say gee, I wonder how that varies by X, Y and so on and you can't do that very well. They are in general traces of behaviors, not internalized states, unobservable internalized states. So they're really weak on attitudes and opinions that may not manifest themselves in behaviors. So you know our sector of thought runs a lot on those kind of measures. They tend not to be there. And as is trivial and easy for survey people to say, they don't cover the population we're used to covering. They're inadequate, obviously inadequate on various dimensions. Okay, so real timely, lean in variables, coverage problems.

And I guess I didn't say one other thing that's obvious, we don't control the measures. They don't necessarily measure the things that all of us would want. They measure often we think, correlates of the things we're really interested in but since we don't control the measures they're only proxies for a lot of what social scientists would be interested in. Okay so how do we, how should we react? We have an old paradigm that's getting chipped away by societal changes. We know that weaknesses and surveys more and more – in fact, most of our work in survey methodology especially is focused on those weaknesses more than the strengths. It's kind of an odd state of affairs.

So how do we get from here to there? The kind of funding that I think the U.S. needs right now on this would address access of new data sources. It would address how to use them jointly in combination. And then there's a bunch of statistical programs that I think we need. So let me go down this list. Before I left Census I was spending a lot of my time talking to folks who hold so-called big data. So these are the CEO's of the companies that are sitting on vast stores of data and the purpose of those discussions was really to see whether the federal government might not enter into agreements with those companies so that those data might be used for official statistics as auxiliary measures in a combined way. And the purpose of the conversation was to say what would it take for you guys to put your data or give access to your data into what you would label a safe environment for common good purposes. And the answers were pretty consistent across these leaders. One is they're worried about liability, that they want legislation that says if our data are used and a widow in Lincoln Nebraska thinks that she's been harmed by some data analysis, I pick on you guys, we don't want to be subject to million dollar claims. We're worried about violations of confidentiality because they know their business depends on a modicum of attention to that. They're worried about getting scooped on a product if you will, developed off their data that if they thought about a little more carefully they would've thought of and they can make money on. And then there are a bunch of practical things that I didn't know about that they taught me. One is the whole notion of moving data is over. You don't move data anymore. I guess it was eBay that changed their datacenters. They just moved their datacenter. Took them six months to move the data.

So it's not like an Inter-University Consortium for Political and Social Research (ICPSR) archive that they're talking about. You're not going to have someplace where all the data reside. Wasn't some one place, you have to have some virtual place. So the idea that hasn't happened is something that I think only a public–private partnership can achieve and that is you could call it a big data consortium or something but there is a need for fuller access. There's one other thing they told me by the way. Some of these companies are making money on statistics off of their data but they make it off of what happened yesterday or what happened last week. And no one wants to pay them money for what happened last month or last year but you know, many social scientists could put up with last week or last year I think that would be a quantum leap in our ability. So that could be an area of negotiation.

So the very first idea is getting serious, assembling the right stuff for a big data consortium and that's all about access I think. And we could talk about what ought to be in there and what would be hard to get in there and so on. And then, so that is more like administration or research administrative stuff, right? There's no science there at all. It's just setting up the infrastructure that permits the science. There is a bunch of work to be done on data integration.

So if I was at the NSF and had Myron Gutmann's job or if I had Cheryl Eavey's job I would announce a data integration research program. We're all doing various bits of this. It's happening in a variety of ways in a variety of disciplines. Computer scientists are doing it, statisticians are doing it, social sciences are doing it. There are a set of ubiquitous problems when you're faced with two datasets that might have shared unit observations in them, may not. They measure sort of similar things on kind of the same population. And you as a scientist look at this and say, "Gee, I would probably know more about the phenomenon I'm interested in if I could somehow put these together. And putting them together is the issue. What does that actually mean technically? How would you do that? So this means support much more fully of matching algorithms that ought to be interdisciplinary between computer science and statisticians.

But thinking about how you practically used probabilistic matching techniques to impact the estimation and the inference. That hard work could really pay off relatively quickly I think if you dropped some money into it because people are working on it but they're working in their silos and we could, I think you could do a lot of good work relatively quickly if you threw money at it.

There's the traditional stuff that survey researchers think about and that is the measurement error properties of different media of data collection. That needs nurturance right now. There's a long history of this. The history however if you think about – I mean go over every article you read on measurement techniques or mode effects. They are often asking the question what's better. They're not asking the question how could you combine optimally for inference if you're stuck with different modes of data collection and two different datasets, how do you factor that into your estimation process. And there isn't a lot of literature on this as it turns out. There's much more on I can show you the weaknesses of one mode or the other.

The missingness side is a huge literature. We need I think a little funding, a little work on moving all the stuff that has focused on nonresponse into the coverage domain a little more.

So, the summary of this is – I don't think NSF can do the first one by itself frankly. The big data consortium I believe probably needs a little national academy work to set the groundwork. It needs legislation; it needs convening of the private sector and the social science sector. It's multiple years of work, high risk, high payoff, low odds I think. I admit that but gigantic payoff for the country. I think you could sell it on Capitol Hill by saying the country that figures out how to use all these data resources in a coordinated way is going to win. I actually believe that. So you could do a little nationalism to sell it on Capitol Hill, but you have to face the private sector side on this. Let me just do one more pitch on this. Many of these large companies are throwing away data when they get old. When they can't sell them anymore and they don't need them to run their business they're throwing them away. And from those of us who analyze data from Roper or ICPSR, you know it makes you cringe inside to realize what we're not going to know about yesterday ten years from now. So there are real stakes here. But I admit that's heavy lifting as they say. I believe fully that the NSF capability is the data integration stuff and I think it needs infrastructure funding because if we just fund it through the normal programs, you know political science will do a little stuff with voting records and sociology will do a little thing on birth records or something like that and we'll have these little things. We will reinvent a whole bunch of linkage and estimation and scrutiny techniques that will eventually get us there but it's going to take a lot longer.

This is unsexy infrastructure funding I admit but I think it will take us a lot longer unless we fund it that way so that would be my pitch to the powers that be on this. So let me end by going back to the beginning. I wrote a piece that ends with saying I don't think surveys are dying, I think they're changing. I believe that. There will be something that offers the wonderful attributes or we need something. This universal coverage is really important for Democratic societies. We need to know that we're studying all of the people in a society and we're making inference in some sense equitably and fairly. This is not just social science, it's a bit of political philosophy too but I think that's important for the U.S. And the rigor of the survey inferential paradigm is a desirable thing. So I was with someone who's analyzing Facebook data at some meeting who began the conversation after I said, at that time I was at Census, began the conversation by saying, "Do you know all surveys are biased?" And I said, "Yeah I spend a lot of my time worrying about that." But he meant it in a completely different way. He meant it that since we don't measure social networks our inference in individual attributes is perverted. Now that's an arguable point I think. But when I did my survey bit of the dialog which is "yeah, but is everybody on Facebook?" He said "sort of" what I remember early sociologists saying about their city studies which were popular when I was a young man, and that is his answer was, "Yeah but they're 9 million people. I have 9 million observations so what do you have?"

But we are headed for a time I think where inferential statements will get cloudy. And I think surveys have a unique perspective first of all and an obligation to keep that part of the conversation because we're going to get more and more data off of these organic data sources where the inferential population will be unstated and we need to keep talking about that. I think it can be talked about only if we do these sorts of combinations that I'm proposing. So to sum up I think we need work on a big data consortium. This is a new institution in the country and I think at a much finer grain level NSF and federal agencies need to get really serious on data integration problems and there's a bunch of basic questions that can be answered with the right funding and the right assembly of interdisciplinary teams but I don't see it happening right now.

Robert M. Groves is the Gerard J. Campbell, S.J. Professor in the math and statistics department as well as the sociology department at Georgetown University where he has served as the Executive Vice President and Provost since 2012. He is a social statistician who studies the impact of social cognitive and behavioral influences on the quality of statistical information. His research has focused on the impact of mode of data collection on responses in sample surveys, the social and political influences on survey participation, the use of adaptive research designs to improve the cost and error properties of statistics, and public concerns about privacy affecting attitudes toward statistical agencies.

Prior to joining Georgetown as provost he was director of the U.S. Census Bureau, a position he assumed after being director of the University of Michigan Survey Research Center, professor of sociology, and research professor at the Joint Program in Survey Methodology at the University of Maryland.

Dr. Groves is an elected member of the US National Academy of Sciences, the National Academy of Medicine of the US National Academies, the American Academy of Arts and Sciences, the American Statistical Association, and the International Statistical Institute.

Dr. Groves has a bachelor's degree from Dartmouth College and master's degrees in statistics and sociology from the University of Michigan, where he also earned his doctorate.

71

Metadata and Preservation

Steven Ruggles

I didn't go to graduate school in information science. I'm a historian by training, but I've been hanging out with information scientists for many years, so I'm beginning to pick up the lingo. Historians have actually been at the center of data curation activities for many years. Both the current and immediate past directors of the two biggest data archives in the world, the U. K. Data Archive and the Inter-University Consortium for Political and Social Research (ICPSR), are historians. And I think the reason for that is that being an archivist is an honorable thing for a historian to do. For an economist, becoming a data archivist would be a real step down, but for historians, it's got a lot of status. I'm going to focus how we got here and where we're going and I will discuss four big issues – data integration, electronic dissemination, sustainability, and metadata.

I'm going to start by giving you some background on the development of census microdata because that's how I got involved in this. We're going to start by introducing the idea of data integration. The reason why we need to integrate data is to make information compatible across data sources or over time or between countries. The need for data integration was evident from the outset. The first microdata was the 1960 public use sample, which was distributed either as a 1 in 10,000 sample on 18,000 punch cards or a 1 in 1,000 sample on 13 UNIVAC tapes. One inch could hold more information

S. Ruggles (✉)
University of Minnesota, Minneapolis, United States
e-mail: ruggles@umn.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_71

than a punched card. Then, in 1970, the Census Bureau greatly expanded the public use sample and added a lot more detail. But, the most important thing they did was they went back to the 1960 sample and they made all the fields compatible, so they coded relationship the same and occupation and everything and they lined up the record layouts so it was just the same as 1970. So, that led to an explosion of research on change over that decade from 1960 to 1970. This meant that everybody got the idea to push this back in time. Hal Winsborough got a huge grant from National Science Foundation (NSF) to make samples for 1940 and 1950 and that was done in 1982. Then, Sam Preston, completely independently, at almost the identical moment, had the same idea. He wanted to make a sample of 1900 census and he did that by the digitizing microfilm of the original census, which had just been released in 1972. So, he used a dumb terminal and brought it the National Archives and made a sample. When that was finished, he made a sample of 1910. And then, both Winsborough and Preston got tired of making census samples, so we started in the late 1980's, making samples at the University of Minnesota, first of the 1880 census, then the 1850 census. By 1991, we had 9 samples and a 10th one on the way – 1850, 1880, 1900, 1910 and 1940 through 1980 and then, 1990 in preparation. There were ten different codebooks, 3,000 pages of documentation, eight incompatible systems for sub-state geography, nine different coding systems. My favorite variable is 'Relationship' which had 72 categories in 1900; 113 in 1910; 23 in 1940; 12 in two different variables for 1960; 20 in 1980. When the Census Bureau did the 1980 census, they forgot about the idea of being compatible with 1960 and 1970 and made a completely different system. Even worse, the three people who made two samples each, Winsborough, Preston, and me, each of the two samples we made was completely incompatible with the other one. I mean, even – and I'm embarrassed to say that was true of ours as well; we didn't make our samples of 1850 and 1880 compatible with each other.

In 1991 I proposed to create an integrated dataset that would cover all of the censuses with harmonized codes, consistent record layout and no loss of information and that was the initial version that came out in 1995. And it allowed things like this, which is one of my favorite graphs. This is the percentage of elderly residing with their adult children. It declined from 72 percent in 1850 to its low point in 1990 of 14 percent. We developed new metadata for data integration. The structured metadata provides the column locations for each census year and where the data quality flags were and what the original codes for each category were. And then a composite coding system with integrated codes has two parts. The first two digits are the lowest common denominator that's compatible across all census years and the second two digits give additional detail available in some census years. And then, finally, we gave it a standardized label. So, our current version of that metadata is much richer, has much more information in it, but it's really the same basic idea. And you can see the results here. This is the relationship variable. This is the integrated version available from 1850 to 2010. And then, we have, if you click there, you get the detailed version of the extra categories that are available only in some years, like polygamous spouse. So, or you could flip over and see how many cases are available for analysis.

Around 2000, Bob McCaa, my colleague, convinced me that we ought to do this for the whole world and so far, he's managed to convince 100 national statistical agencies to go along. And so, we have been integrating a lot of data. And then, we started working with genealogists and national archives and genealogical companies to develop additional historical census data. And so, we now have over 500 censuses and surveys that are all harmonized from 75 countries and almost a billion person records and that's probably going to double over the next five years or so. Now we're going to the next level with data integration to try to make data that has drastically different formats and it comes from different scientific domains, easily interoperable and we're starting with population microdata, small-area data, land-use statistics, land cover data and historical climate data. And these are in three basic formats – microdata, like the census data; area-level data which could be any other type of data that pertains to a polygon, typically a political unit, but also could be a watershed or something like that; and then, raster data, which is data based on remote sensing and modeling. And our idea is, what we're doing is we're attaching the characteristics of communities to the microdata, whether those are derived from rosters or from aerial statistics, then we're converting area-level statistics. Then, we're converting area-level data into raster data and raster data into area-level data by summarizing it. And then, we're going to work on making microdata easily converted in area-level and master data through rapid tabulation. So, that's the general idea of Terra Populus.

So, how does all of this apply to surveys? Well, the big three surveys supported by NSF have pretty good temporal integration. The American National Election Studies (ANES) and Panel Study of Income Dynamics (PSID) have combined files that cover all the years. PSID has a mechanism in the extract system that can merge data from multiple waves. So, that's pretty good, while they could be better when it comes to documenting issues of cross temporal comparability, none of them have, at least, very visible documentation of those concerns and I think that that's very important.

But, maybe the relative success here is that they've been very careful not to make changes. As Barbara Entwistle put it, "If the goal is to measure change, then it is obviously important not to change the measures." And I think that that's true, but there is always a tension in the surveys between the need for continuity and the need to make innovations, and that tension, I think, is always going to be there. There's also the issue of integration across data sources; that is, integration across the surveys, the big three, and other surveys and this could be agent if all of them could agree on standardized classifications across all variables. That would just simplify things. Very importantly, it would be valuable to try to agree on standard geography, especially low-level geography. And NSF could contribute here by funding data integration projects, retrofitting older datasets and helping to make crosswalks. And, particularly, doing the type of stuff that Terra Populus is interested in to make it easy to get contextual variables of all types or for survey data, including information from rasters, which would include things like tree cover, air pollution, all kinds of interesting variables that could be attached to individual records.

The second topic I want to cover is dissemination. Dissemination is critical. The data have to be widely used to justify the investment in their creation. All three of the big three are widely used, but it could be better. This is an excerpt from the budget justification of the original Integrated Public Use Microdata Series (IPUMS) proposal to NSF submitted in 1991. We requested funds for 450 2,400 foot, 6,250 dpi 9 track tapes. And the reason why we made it 450 was the dataset would fit on 150 tapes, roughly speaking. We figured we could get about 160 megabytes on a tape and the whole thing was going to be about 16 GB. And so, we needed 150 tapes for that. And we needed one copy to import the data, one copy to create a version of the integrated datasets for us to keep in Minnesota and then, one copy to ship off to ICPSR and we were going to have ICPSR do all the dissemination for us. As you can see, we also had funding in there for 300 diskettes. But, anyway, we never bought a single tape and the reason why was that the Internet got capable enough just in time. We were able to acquire all of the datasets we needed over the Internet. In 1993 with our preliminary version of IPUMS, we started disseminating data over the Internet on our anonymous FTP site. You can see there Joel Perlman, on August 3 at 10:00 in the evening, was the first person to download a dataset.

And then, in April of 1993, the big innovation came along: the Mosaic web browser made the World Wide Web feasible. As you can see there, it's explaining to the reader, "You click on these things and you'll follow the link." And so, this is our first website. In 1994, it didn't do much. They all looked like this then. We made it fancier pretty soon and, very quickly, we developed an interactive data access system, which would merge data from all the different samples so you could conduct pooled analyses and allowed one to subset by variables and by subpopulations. And then, we made the system dynamic and we had online registration. Maybe the most important thing is that we converted all of our documentation into hypertext format and made it so that it was linked. Every place we mentioned a variable name, you could get directly to the documentation for that variable. So, we're currently on our fourth generation of dissemination tools and they're a lot fancier and most sophisticated and they have lots of bells and whistles. We provide Data Documentation Initiative (DDI) metadata with every extract. We also have online analysis tools and other nice things.

So, how does this relate to the surveys? We needed to have dissemination tools early on because the scale of the data we were dealing with was bigger than most people could handle. And so, we needed to make it feasible for people to slice the data and get just a little piece of it so that they could analyze it. But, surveys, generally, pose a little bit different problem. It's not the number of rows. The number of rows is very trivial. The number of cases in all of the surveys, they're small. It's the number of columns, the number of different variables. And, in a lot of cases, it's also the complexity of the survey design that also complicates access, either because it's longitudinal or because it's multilevel.

One of the most important issues with all of the surveys that is unmet is tools for variable discovery and exploration. It's very difficult to figure out variable availability across time or across samples and a lot of the key metadata that is necessary to use most survey data is locked in PDF files and it is not actionable, which means that it's expensive to make data access tools.

Here's an example. This is the main page for PSID questionnaires and important documentation. There are 130 files on this page and I estimate that it's about 50,000 pages of documents. This is not unique to PSID. I think that the other two surveys, the General Social Survey (GSS) and the ANES, have approximately the same number of PDF files and they have slightly smaller number of pages. My page estimates were based on just sampling a few documents and figuring out what the average length was and so, they could easily be off quite a bit, but these are in the right order of magnitude, I think.

And these are very important documents. This is a couple of examples from PSID again. The survey instrument for 1968 is on the left. The study design for 1972 on the right, describing how to split the sample for ransacking and testing. Right now, this stuff is very difficult to track down. There's many documentation files for each physical file and it's not easy to find out where the piece of information you're interested in is.

There are other key issues. There's online analysis. GSS is the main one of the big three that uses online analysis. We do a lot of online analysis. Two of the tools that are widely used for online analysis are survey documentation analysis, which was developed by Merrill Shanks and Tom Piazza, political scientists at Berkeley and which organized like a co-op. ICPSR relies on them, we rely on them, GSS relies on them. And, you know, it's not being well maintained; the new version is, maybe, never coming out. And so, that's a problem. We're up to about half a million Survey Documentation and Analysis analyses and it's doubling every year. I think GSS does more than we do and ICPSR does more. And so, this is a concern. The NESSTAR software started out as cooperation between European data archives. Now, it's private. But, it also is not being maintained and you'd be crazy to adopt it at this point. So, there's not a good generic online data analysis tool that's readily available. So, we also need data access tools for complex data structures. This is especially true for very complex survey designs like The National Longitudinal Study of Adolescent to Adult Health (Add Health) and the Survey of Income and Program Participation, which are very difficult to use, but there were good tools, that would not be the case.

Another issue is virtual data enclaves. We all have restricted versions of data, including IPUMS and the virtual data enclaves have been set up by the National Opinion Research Center (NORC) and by ICPSR. I think this is an area where NSF might consider building a shared resource that could be used across surveys. This seems like it might be something that everyone should not have to replicate.

I think that the biggest dissemination issue with the current surveys is metadata discovery and browsing. The metadata that do exist are not accessible on the Internet for other resources to discover. And so, it's very difficult to locate variables and that sort of thing. So, I think that in the case of the big three in particular, there is a great need for metadata browsing functionality that would allow people to understand variable availability across time and quickly get to the underlying documentation without going through a million PDF files. And we need better dissemination software. It still would be nice to have a way to just select the subset of variables you need very easily and it would be so easy to build. They all need to deliver machine understandable metadata and they just don't. PSID is set up for

SAS, SPSS, and Stata, but they just provide variable names. There's no value labels. There's no weights, no missing values. It's very, very sparse metadata.

The third topic is sustainability. Why do we care about sustainability? Well, we, as a nation, have invested hundreds of millions of dollars in these datasets and the reason why they're important is the power of cross-temporal analysis. The older the data, the rarer it is, the more valuable. And it is vital that we have stewardship for this. So, there are two issues. One is just preservation, but the other one is maintaining access and maintaining access means that you've got to have some kind of organizational sustainability to whatever organization is in charge of the data. And they've got to have some long run financial sustainability plan and they've got to be able to migrate the data as technology changes and as the software needed to access it changes.

The key question we have to answer, especially when it comes to the big three, but more broadly, any sort of investment in survey data is are these surveys data archives? And do they want to be data archives? If they want to be data archives, they need a formal data preservation plan and that includes some sort of redundant storage. They could join Data-PASS, for example, or they could find some other way to do lots of copies of the data to make sure it doesn't go away in a disaster. They need to work on getting compliant with the Open Archival Information System reference model and some sort of plan for longer organizational and financial sustainability.

If no, if they don't want to do that, then they have to make a deal with an archive that agrees to and they have to get the data there. But then they still have the issues of if the software and metadata are not in standard formats; they've got to do something about that even if it's going to ICPSR. GSS has a little less than half of the files at ICPSR. PSID has very few. ANES has most of them. So, basically, right now, it looks like the big three are regarding themselves as archives and assuming the responsibility for preservation. But, I don't see a lot of sign that they have taken responsibility for preservation as seriously as they should.

Persistent identifiers are closely related to sustainability. Data citation has been very, very casual and that's one reason why it's so difficult to replicate social science research or scientific research for that matter because, often, it's difficult to figure out exactly what data did they use and what version did they use and that sort of thing. And so, it's really critical that we have a system of citation and persistent identifiers for data. And it's going to happen. And it would be good to get in front of it. And, last year, the American Sociological review began a requirement that manuscript submissions should include citations and specified that they should include a persistent identifier, such as a Digital Object Identifier (DOI) system. And I think DOI, for the social sciences, is the standard that we should just all agree that we're going to use. It would be not unreasonable for NSF to simply make this a requirement. It's not that big of a deal, but it really would make the world a better place.

Okay, finally, metadata. So, why worry about metadata? Well, metadata underlie the other three issues. You can't really do data integration or dissemination in an efficient way unless you've got good metadata. And for preservation, you need metadata. If everybody starts to have their ad hoc PDF code books, the work required to do preservation is multiplied. All of the major archives have adopted the data documentation issue initiative standard. DDI is just a – it's an metadata standard that specifies the structure for documentation. It can accommodate virtually anything that we know about a dataset. It's a little clunky, but the point is that it's a standard.

Okay, I thought I'd go briefly over what NSF has been doing to promote curation of survey data and these are just a few links I happen to know about because I was involved in them and I'm probably missing a lot of other activities that I wasn't involved in. But, in any case, there was a 2007 workshop on the general social survey and it made some very nice recommendations. For example, adopting the DDI standard; developing integrated web dissemination system that was driven by the metadata; and I don't know if it was a result or not, but the PSS and PSID's solicitations did specify – both of them specified for the renewal of those projects that they had to have data dissemination with a cutting edge, web-based data archive and ability to maintain cyber infrastructure, and documentation with expansions in innovations and data sharing tools as technology develops. So, there was, sort of, mandate in the proposal renewals for more attention to dissemination, although not metadata.

Then, in 2010, there was a workshop on future investments and large-scale survey data access and dissemination that had representatives from all three projects, had a bunch of outside people, including me, who they thought was from the University of Michigan, but there you go. And, here, the recommendations were excellent, I think. This is the recommendations, all data and metadata be presented or documented according with well-defined protocol – such as DDI – and retrofit all of the legacy data so it becomes readable and develop a federated portal. That hasn't happened for the three datasets. And supports the development of common online tools for search, downloads and basic analysis. That has not happened either. But, there was a solicitation last year for metadata for long-standing, large-scale social science surveys. And so, a few projects emerged from that big call, have just begun in the last month or so, but the one for PSID is going to scan all of the paper forms that were used to actually collect the data. So, the individual responses, and there's 15 million of these forms, and I think this is an excellent thing to do for preservation. I don't think it has much of anything to do with what the solicitation had in mind. But it will be good to have it done.

The other two projects are a cooperative agreement between ICPSR and NORC and they are going to develop DDI metadata. They call it a pilot project, so it's a little bit hard to know how much they're going to do. Obviously, $500,000.00 is not nearly enough money to do all 100,000 pages of documents that are locked up in these PDFs. But, it would be enough to get a start on some of the basics.

In conclusion, everything depends on the metadata. Without having structured metadata, we can't do anything. And I think NSF already knows this and has been taking steps to try to fix the problem. And it's not fixed and it's, maybe, I think, it could be a bigger problem than NSF has yet estimated. Maybe there will be a new solicitation that comes out following on the initial one that will include enough funding to actually make a dent in the backlog of legacy documentation and that would be great.

Steven Ruggles is Regents Professor of History and Population Studies and Director of the Institute for Social Research and Data Innovation at the University of Minnesota. He received his Ph.D. from the University of Pennsylvania in 1984, followed by a postdoctoral National Research Service Award at the Center for Demography and Ecology of the University of Wisconsin. He has published extensively in historical demography, focusing especially on long-run changes in multigenerational families, single parenthood, divorce, and marriage, and on methods for analysis of historical populations. Over the past 25 years, Ruggles has developed large-scale data infrastructure for economic, demographic, and health research. He is best known as the creator of the world's largest population database, the Integrated Public Use Microdata Series (IPUMS), which provides data on billions of individuals spanning two centuries and 100 countries.

72

Usability of Survey Project Websites

David L. Vannette

When I agreed to write this chapter I decided to collect some a little bit of original usability research about survey project websites. I decided the best approach was to actually bring in a bunch of people and have them do some fairly simple and some more complex tasks with the big three National Science Foundation (NSF)-sponsored survey websites, the American National Election Studies (ANES), Panel Study of Income Dynamics (PSID) and the General Social Survey (GSS). And we did this in our lab with screen capture technology going so that we could observe what kinds of patterns emerged in terms of the kinds of problems that people ran into. The questions we wanted to answer were: What works on these websites? What's not working on these websites? And what kinds of principles about usability can we extract from these analyses?

So, what I'm going to describe is the outcome of what we observed in that regard. And so, I'm going to try to use, mostly, examples from the big three, but I'm also going to deviate and show some examples from the Inter-University Consortium for Political Science Research (ICPSR) website, which I think is a much better example of a usable website. So, I think you'll notice, there's going to be a bit of dovetailing, maybe even some overlap between this chapter and the things that Steven Ruggles wrote in his chapters in this volume about the importance of making different types of

D.L. Vannette (✉)
Department of Communication, Stanford University, CA, USA
e-mail: vannette@stanford.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_72

data and documentation easily accessible and usable by people. My focus is on the usability of these websites.

So, first, why should we care? I think it's important to motivate this discussion with why we should care. Many researchers may not have direct influence over survey websites. Maybe some sit on their boards or things like that and can have influence. But, I think we all should care about this for a couple of important reasons. The first, which Steven Ruggles wrote about very convincingly in this volume, is dissemination. And from my perspective, another important aspect of this that he didn't really touch on quite as much is how it's really important for the field of survey research to see an expansion of our user-base in terms of the people who are using survey data. And usable websites can really influence the ease of access of data and documentation and, hopefully, we can bring new users into these fields so that we see more use of our data and then, hopefully, more funding of data collection. My second point is a little bit self-serving. I actually use these websites a bit and download data sets and find documentation and it's not easy to track all of these things down. These tasks are harder than they need to be so, making secondary research easier to conduct, should also be an important motivation for all of us.

So, what do I mean when I talk about usability? There are a number of principles of usability that I was operating under when we were having the users come in and test these websites. The first thing, I think, is pretty obvious and we want to provide relevant and easy accessible information to the users. And I think it's important that those things go together and that they're both relevant and easy to access. Second, we want to enable learnable routines through usability and so, this refers to having similar tasks on these websites follow similar routines so that you can apply learning from the browsing or searching that you did for variables to the browsing or searching that you do when you go to the documentation and trying to identify things in code books and things like that. You want those similar browsing and searching tasks to follow very similar routines.

You want to design efficient paths and this refers to minimizing the amount of – the number of clicks, the number of search terms that people have to use before they find what they want. You really just want to minimize the distance between people when they arrive at the site and when they get to the data or documentation that they came looking for. We want to create memorable patterns and this is, I think, referring to having elements not moving around on web pages, like, disappearing and coming back and I'll show some examples of this later when I walk through some of the principles of usability.

And we want to minimize user errors. We want to make it really hard for users to make mistakes when they come to these websites. We don't want people to get 10 minutes into a task and realize that they should have caught a mistake that they made 10 minutes ago. We want it to be very clear. We want these to be intuitive websites. And, ultimately, I think what this comes down to is we want users to come to the websites and we want them to leave these websites having had a satisfying experience. We want them to come. We want to find what they want and need and then, we want them to be able to make use of that information.

Alright, so what do survey website users want? I think it's pretty clear that they want data and it's pretty clear that the big three know that they want data. These are the headers that are on the home pages of all three websites. So, you can see that PSID has a data tab; the ANES has a data center tab and a tables and graphs tab, so you can go look at some cross tabs and things like that that have already been created; and then, the GSS has a data download tab, a browse GSS variables tab and a data analysis tab. So it's pretty clear that survey users want data and the surveys are acknowledging that and trying to provide that. Now, how well they do that seems to vary substantially across organizations.

The next thing they want is documentation. So, PSID has a documentation tab also on their home page. The ANES has made what I think is, maybe, a debatable decision to bury their documentation in the data center. I think once you get into the data center, it's implemented reasonably well, but it's not immediately clear when you get to the home page where you need to go. It's entirely reasonable to think maybe you need to go to the help center or the library to find documentation. But, really, it's all here in the data center. And the GSS also has a documentation tab.

The third thing I think that website users want is to come and find examples of ways that these data are used in the literature and I think the websites are all pretty good about patting themselves on the back in terms of publications that have come out of their projects and everybody provides those pretty uniformly. So, some principles of usability that came out of these studies that we did where we had people come in, and I'm going to walk through each of these with some examples from the websites.

First, standardizing task sequences. We want similar tasks to be structured very similarly. So, consider the PSID website for a minute. In their data center looking at the cross-year index. On their website, you will see their variables are listed and then across the top, is the listing of years. And then, this grid down here corresponds, so you can select which year and which variable and you can select that and then you can add it to your cart, right?

So that's one routine. What we want is, at least if this is going to be routine, that PSID chooses, we want it to be the same across the different aspect of the website. Well, that's not really the case.

So, then I went to the PSID variable search tool and the year is here where you can select which year you want and the data type, section of the code book. And then, when you've done your variable search: for example, I've searched for income, the results come. Then, on the right of the page, is where you would select which variables you want and then, you can add them to your cart. So, it's clearly a very, very different task structure. The routine is very different. So if you're new to the website, first, you learn how to do it in the data center and the cross-year index and then you come to the variable search and you have to learn an entirely new routine. This is just really not an efficient way to use your user's time.

The next thing is reducing user workload. In researching for this chapter, I was at the GSS website at one point made five clicks with no discernable progress. Five clicks of unfolding, single option radio buttons. There's just absolutely no reason that I should have to click five times to finally get to a point where I have to make a decision about what I want. So, we really want to try to eliminate these blatant wastes of time whenever we can.

We want to design for working memory limitations. So, Steven Ruggles wasn't terribly optimistic about the NESSTAR system in his chapters, and after having played around with it a little bit on the GSS website, I'm not really optimistic either, and neither were any of the people that we had do usability testing on the GSS website. And I think part of that is that they have a 51-page guide just to how to get started with using this tool. And this isn't a tool that you can use elsewhere. It's really specific to the GSS. You have to do it on their website. It's just really, kind of, clunky and it's a lot to learn and try to remember in order to, like, get a cross tab for something. It's just too much to ask of people.

So, I think you also want to display directly usable information to people. You don't want them to have to make conversions or summarize things or go digging for things that are very, very important and that should be presented up-front. So, here the ANES does, actually, a pretty good job with this. The study page for the panel study is a good example. One thing I will point out before I get into this is that some elements of pages on the ANES seem to disappear and reappear without any rhyme or reason. So, in order to navigate back, if you wanted to go somewhere other than the data center they've removed that element from the webpage. And so, that's one of the principles that I mentioned briefly earlier. You don't want that to happen. You don't want things to disappear, especially when they're really crucial to being able to navigate through the site.

Anyway, what they do well is provide a really nice summary of some of the really important features of the data set so that things like the number of completions and the mode and information about the way it's making sure that people know to use the weights and where to get information about how to use the weights, things that people may not know initially, but that are really important for people to know. And this is directly usable information that people should have before they start doing data analysis.

So, the next thing is guiding users through documentation and this was one of the biggest issues that we had with the usability testing that we did. I looked at the ANES variable code book and didn't even make it past the very top of the first page of the ANES cumulative data file variable codebook. What I found was there is not really much meaningful, especially to a new user, that's going on here. I didn't know what was in the document and I couldn't tell where anything was in this document. It didn't orient me at all to what was happening. And I think this is really unfortunate because you could open this and it takes a while of scrolling before you start figuring out, kind of, what the pattern is and what's going on and where to locate things.

I think a better example is from the GSS codebook. They have a landing page. When you click on the codebook information, you can open the entire GSS code book and get it in PDF form, which is problematic given the issues with having everything locked in PDFs and not linked. But I'm not going to belabor that. But, you can choose which section of the codebook you want to go to. And so, you know, it kind of breaks it down. And then, they provide this really nice option to go to technical appendices. So, if you're looking for study designs or variables and things like that, you can find that and it's summarized here and you can just directly go to that portion of the code book, which is great because it's a 3,500-page code book. So, when you do go to the code book, they further layout the information for you by actually providing a table of contents, which isn't really a new innovation, but it's, kind of, an important aspect of orienting your users to the data that you're hoping that they'll find. And so, I think that's a couple of things that GSS has done somewhat right, at least if they're going to take this approach of having everything locked in PDFs.

There are a few more principles of usability that I'm going to walk through. I want to start by directing attention to the ICPSR website. It has a fluid design, and this is a minor point, but as somebody who uses these websites, it's important. For most of the other survey project sites, there are bars on the sides of the websites, just empty space, and that's because the websites are not using a fluid design in maximizing how they render to each user's screen size and resolution. They've standardized the way that the website will be displayed on a user's monitor. So, it doesn't actually fill the whole screen. ICPSR is using a fluid design and so, this actually just automatically resizes to take into account the resolution of your screen and your screen size, which is really nice especially if you're looking at, massive amounts of data or huge documents. You don't want to be limited to half of your screen.

So, the other thing that I wanted to point out about the ICPSR that is nice is the fact that the elements do actually move and that was one of the things that I said earlier shouldn't happen. But they've compensated for that flaw by the fact that they've actually color coded everything really nice in terms of their tabs. So if you click on the 'find and analyze data' tab you can remain oriented to where you are in the website because in the sub-header, they've also color coded that. No matter where you are in the website, it's pretty easy to tell where you actually are. If you start browsing and wonder which part of the website you're in you can just make that correspondence and it's pretty easy to keep track of where you are.

Another good thing that ICPSR has done is they've invested a bit in developing their search capabilities. And I think this is important because a lot of the survey websites right now are designed to be browsed and that just involves a lot of clicking. As we moved forward with the Internet, we've definitely moved towards much more of search based information retrieval model. ICPSR has recognized this and they've actually built a pretty nice search tool. You can enter a literal research question, so they give the example here, "Do children of Asian immigrants speak English in the home more often than children of Latino immigrants?" So, you can actually type that into the search bar and it will come up with the relevant datasets and variables and things like that that you're looking for. That's really cool, especially if you're a new user and you don't know the lingo, you don't know the keywords, maybe you don't know exactly what you're looking for. You can just type your research question in and it'll find it for you and I think that this is the kind of thing that would be really helpful for the big survey project websites. ICPSR definitely still makes browsing available and you can actually browse by a number of different criteria, which is great. But improving search algorithms and search models is definitely going to be the way that we should be moving with some of these web design projects.

One thing that we did as part of the usability study was have some users do the same task on the three different websites. So, for example, we had them locate the variable for income. It's an entirely reasonable task. It's probably one of the things that people do most often, since income a key variable that a lot of people use in their research and models.

None of the survey project websites that I evaluated gave any indication of which are the most important variables and I think that that this is a key oversight and it's one that's true across the different websites. I think it would be extremely helpful if these websites could give some sort of indication of which of these variables is downloaded most often, published most often, just any kind of indication of what the most important variables are because scrolling through hundreds or thousands of variables is really not efficient. In some cases, it is possible to do a search using a browser function, not a website function. But I read a statistic recently that only about 20 percent of Internet users actually know the trick that you can hit Control F and search the entire web page. The average user is not going to know how to do that. But at the websites are currently designed it's just not an efficient process.

So, the other two websites, the ANES and the GSS, had very, very similar experiences in our testing as the PSID, with just a lot of difficulty finding and identifying the actual variables that our testers were looking for. It is incredibly frustrating because this should be one of the simplest tasks and yet it was taking our users nearly 10 minutes just to simply locate a common variable.

So, in conclusion, you know, I think we do some things well. We do provide a lot of data and I think Steven Ruggles made a great point in his chapters that web has really transformed the way that data can be disseminated and the same with documentation. We provide documentation as much as we can and I think a lot of the newer surveys are, the newer waves, are doing a much better job of putting everything online right away, but there's still gaps in the documentation. I know, for example, with the ANES, there are some show cards from, like, the 1950's and 1960's that just aren't available. They've been lost forever. And so, if you happen to navigate to that page on the website, it says, "Do you have one of these? Because we need it." Just in case you happened to have found one lying around your house from when you're cleaning. These websites provide publications pretty well. The ANES actually has a nice thing that rotates through on a sidebar showing a bunch of their recent publications and I think that's kind of cool.

We definitely have room for improvement. We can make sure that tasks are structured in similar ways within and between websites. Common tasks shouldn't be terribly difficult to do. If you've implemented it once, you should be able to implement again. You don't need to reinvent the wheel twice on the same webpage. In terms of design, we can design much more efficient and intuitive web pages and there's no reason for us to not do that. And then, the codebooks full of unlinked PDFs, we're just locking away these resources that are absolutely crucial and making it really hard for people to find the documentation that they need in order to make proper use of the data and the variables that we've made available to them. The structure of the data sets and the fact that they aren't really compatible with each other over time also provides an opportunity to improve. It's currently really hard to link them. This volume contains multiple chapters about data linkage issues and linking with government records when we have a hard time linking our own surveys together. So that's also an area where we could use some improvement.

And so, there are some easy steps that can be taken. The first one is learn from users of the websites, get feedback from them. The second one is learn from each other and look at ICPSR and emulate the kinds of functionality that they've built. And, if you're not going to do this, you should push more of their data to ICPSR in order to make it more usable and more available to a broader range of users.

Conducting usability testing when making changes is also best practice. I think it's really important to conduct testing as you go so that you know what works and what doesn't work from a user's perspective. Don't just rely on your web design people or the intuitions of the researchers because there's a pretty good chance that they're wrong, which is what we can see from the current state in many respects of the big three websites.

And there's this great resource, so after I had done a lot of the usability testing, having people come in and watching a lot of these screen capture videos, you know, I had a lot of things written down and ideas and observations that I made. And then, I happened to come across: usability.gov and they have fantastic resources for how to make useable websites and they provided a lot of the terminology that I was able to put to the descriptions and observations that I had made. So extremely, extremely useful.

The last thing, I wanted to come back to one more principle of usability and this is, you know, we have a lot of expertise in doing surveys. We should actually apply some of this expertise to designing our websites and get feedback from our users and we should get feedback often. We can design good surveys. We can get the information that we need from our users and there's no reason not to do that. And we can also make use of things like pretesting with users and making use of paradata. I don't know if any of the websites are making use of the paradata that they have to find out where users are getting hung up. Where are the pages that people are dropping off and just leaving the website? How long is it taking before people – from when someone enters the website to when they get to what they're looking for? And are there ways that we can design more efficient paths to help users get to what they want in a much faster way?

I'll conclude with an anecdote. When I was preparing the research for this chapter, I was meeting with Paul Sniderman, who was one of my graduate instructors at Stanford. And he asked if I would please go to the ANES website and find the last year that the economic individualism questions were asked and send him the question wordings because he wanted to use them in a study that he's fielding. So, I said, "Sure, absolutely. I can do this." Fortunately, I actually had a Ph.D. student in political science at Stanford who had worked as an employee of the ANES for three years coming in to do some of the usability testing for this chapter that afternoon. My plan was to have her do the GSS and PSID, just because I figured she would know the ANES really well and I was curious to see how well that skill set will transfer to GSS or PSID. But, since I had to find these variables quickly, I thought, "Hey, great, I'll kill two birds with one stone. I'll see how efficiently an expert can actually use this website and I'll do a screen capture of that."

So, to make a long story short, we got to the end of our half hour period and she hadn't found the variables. She was just completely unable to find that and she knew the exact concept that she was looking for, that it was economic individualism and that they had been asked on the ANES. And using all of her experience, knowledge, and ability, and she knew the Control-F search trick, so it's not like she was scrolling forever and ever, wasting all of her time. She was completely unable to find these variables. Fortunately, after she left, I was able to go to Google and actually find the question wording using Google, looking for economic individualism and ANES. But, the fact that an extremely experienced user who actually works for the ANES, can't find a simple battery of questions that appeared in multiple years on the ANES, I think, says a lot about the current states of usability of these websites, in general. So, I think there's a lot of room for improvement.

David Vannette is a Ph.D. Candidate in the Department of Communication at Stanford University Stanford, CA, USA, and Principal Research Scientist at Qualtrics LLC, Provo, UT, USA.Vannette received a Masters degree in Survey Methodology from the University of Michigan and is finishing a PhD at Stanford University in 2017. His research has been published in academic journals, book chapters, case studies, technical reports and he is a frequent speaker on survey methodology for academic and commercial applications. He has also been affiliated in research roles with both the Ross School of Business and the Institute for Social

Research at The University of Michigan, the U.S. Census Bureau, and Stanford University. Vannette actively serves as a reviewer for academic research articles and volunteers on committees for the American Association for Public Opinion Research (AAPOR) and the Pacific Chapter of AAPOR.

73

Research Transparency and the Credibility of Survey-Based Social Science

Arthur Lupia

As survey researchers, when we are having conversations with others about the value of what we do, I think there are two questions that come up fairly regularly. One is what is the value of what you're doing? And by value, it's not just our ability to tell people why we like this work or why it's important. It's our ability to tell people who have the option to fund breast cancer research or research on post-traumatic stress disorder, why funds ought to be allocated to what we do or to pursue methodological questions that we have, as opposed to other things that they can do the resources. And so, for what it's worth, I think we can win those arguments if we give them correctly, but we actually have to engage them.

Value isn't on our terms. Value is relative to the other options that people have. So, when people start to think about the value of survey research, what is it that we provide to the federal government? What is it that survey researchers provide to society? And I would argue that what we provide is meaning, right? Events happen, people do things and people ask the question, "Why?" The Twitterverse and the net is such that lots of explanations float. But, there is a demand to try and evaluate a number of these explanations with respect to data, with respect to evidence and logic. And so, what we do, I think where our value comes to the federal government and other agencies is when there are competing claims about meaning. That's an

A. Lupia (✉)
University of Michigan, Ann Arbor, United States
e-mail: lupia@umich.edu

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_73

opportunity for us with our methods and our integrity to step in and say, "Here are some meanings that you ought to bank your strategy on, bank your view of society on, and here are things that you should set aside." If that's our value, providing credible meaning, then a primary focus for us should be to invest our credibility; to invest in our ability at those moments to give meaning to things that people ought to believe.

Within the social sciences, generally, and within the survey research field, in particular, we do have large pockets of limited introspection and documentation of how we know what we know. As a general matter, we have some vulnerabilities. Those vulnerabilities go all the way up to nearly the top of the food chain. I won't talk about the US Census, but I will refer to the major National Science Foundation (NSF) studies, such as the American National Election Studies (ANES), which Jon Krosnick and I have some familiarity with as former co-Principal Investigators, and other comparable studies. We have some vulnerabilities in terms of our ability to give credible meaning because we haven't asked questions.

Limited introspection and documentation threatens our legitimacy and credibility. I'll talk about why in a second, but it is a real threat. When you run an election study, people can ask you all kinds of questions not just about your methods, but about your incentives, whether you're really trying to skew things and whether you have a point of view – that is, whether you're just trying to do is give a numerical version of an ideological story. You've got to be able to protect against that. If you don't know why you're doing what you're doing, you can't explain it. When questions about your credibility and legitimacy show up, it's not a great day to be you. So, it's better to have had the introspection. That's the main vulnerability.

As we think about what to do next as a field, I want to give you some definitions and then, focus on what can we do to help ourselves make more credible and legitimate arguments. So, there are two definitions: credibility and legitimacy. A key thing when I use these terms is these are not inherent or organic properties of objects, right? A study isn't inherently credible or legitimate. Credibility and legitimacy are judgments that are formed and generated between the ears of people who observe what you do, who ask questions about what you do. To the extent that we want people to regard what we do as credible and legitimate, we have to earn that.

Credibility is "the quality of being believable or trustworthy." When there are questions about how we know what we know or whether our meaning should count, do people have a reason to believe us. A good social scientist seeks to offer a credible explanation, a credible view of what's going on, as opposed to things that might be less credible.

By legitimacy, I mean that something is done in accordance with a set of principles that one can articulate. With legitimacy, you might not agree with what I'm doing, but I can say, "Look, I've done it in accordance with a set of principles." And you might say, "Well, with respect to those principles, what you're doing is legitimate." If you think about where our leverage comes in in science, it's the idea that people choose to believe us for some reason and maybe a lot of them understand that we are acting in accordance with certain rules. But, of course, it helps if we know what standards we're operating with respect to and it helps if we can document that we're doing as we say.

So, in terms of what we can do and in terms of our response, I think here's the question. It's, "How much effort?" Many researchers put effort into documenting and thinking through how we know what we know. My underlying argument is that we could all do more. We could all do more about *data access*, about not just making the data available, but making it easier for people to use and to understand what we're actually doing. By *analytic transparency*, I mean making it easier for people to take our data and draw inferences from it. *Procedural transparency*: when we deliver data, part of the meaning of that data comes not just from the respondents, but from decisions that we've made about how to code their answers. Being transparent about that can help survey producers with their own credibility. Finally, documentational rigor. But, the tension here is, "How much effort?" and the question for each of us is, "Should we do more or could we do more?"

So, I want to talk a little bit about just threats to credibility and then, I'll talk about some responses we might have. I want to break threats to credibility down into two domains. One is in the *analysis of survey data* and then, one is the *production of survey data*. Now, this is a setup. It's just a little secret. For the analytic things, I'm going to show you some common things that people do with surveys that none of us would be caught dead doing and we're going to think, "Oh, I would never do that." And then, the trick is I'm going to go to the production of survey data and show some parallel moves that, maybe will make us a little more uncomfortable. But, for now pretend that you didn't read that.

So, here is a challenge for a lot of scholars. What they want to do is to make credible scientific contributions – and I'm talking about science, not the punditry. The problem is scholars have limited time. There's pressure to push things out the door and get things published, particularly if you're young. So, the temptation is to ignore some of the assumptions that you're making and then, you might have to document your study, but maybe that's something you can do later. You know, publish it now and then, later on, if it gets published, then you'll prepare your notes and your documents and all that. So, that's the temptation. Ignoring assumptions and incomplete documentation seems like a bad idea. Seeing these matters as a secondary activity, rather than something one ought to be doing from the very first stages of design, turns out to be a critical problem.

Here is a simple example from my home discipline in political science. Somebody will make a claim, and the reliability of this claim depends on the truth value of the number of assumptions. So, some of the assumptions that lead to this are in the domain of statistics. Some assumptions are about various statistical models. Others are based on some decisions that were made before the data was released. The decisions made by people who turned dollars into data points. So, every survey based claim about elections, for example, depends on some assumptions about some things that matter. Here's an example. A scholar wants to claim that presidential approval is 67 ± 4; they want to say why this happened and so, they run a regression. It's a common approach. You take secondary data like ANES data and you run a regression on it. The regression could be something of the form Y equals alpha plus beta X plus epsilon. Maybe it's a linear model, ordinary least squares, where what you're assuming is that you have an additive structure and you add this and you add that and you add this little error term. That's a way of representing a system of conceptual relationships. Maybe the researcher wants to know the effect of race on whether people voted for Obama or not. And so, what I want to ask you about this regression equation is: what are we assuming about the process? Because when you think about this, if this were ordinary least squares, you're assuming an additive relationship between a constant term, a set of variables, and an error structure that you've made some pretty strong assumptions about.

One question that we could ask about a lot of this work is, "Well, is there any actual neuro-scientific or experimental evidence that the relationship between the independent variables and the dependent variable is linear or additive? You know, is there any evidence of that?" If such evidence exists, I've never seen it. For decades, scholars in my discipline and in the field of public opinion presented these models, but we did not ask each other those questions. It's like, "you know, regression is how we do things." "The additivity, it really doesn't matter." But it does. When we do these things, we're actually buying into a set of structural assumptions about the relationships between variables, assumptions for which there is little to no evidence, okay?

Here's another thing; happens in my discipline. Again, I don't know if it happens in yours, *stargazing*. So, I'll begin with a hunch that a particular variable has some underappreciated relationship to a particular dependent variable. It could be voting behavior. And so, I'll run a regression and I'll look for stars. I'll look to see if on the variables that I'm focusing on, whether I can get a relationship between the coefficient and the standard error that gives us, say, a conventional measure of significance, you know, maybe 5 percent, 10 percent, whatever it is. But, what actually happens a lot of times is that a scholar runs a first regression, looks for stars and if the stars support the scholar's hunch, he stops and writes a paper. But, if the original version doesn't work, then he runs more regressions and redefines variables. Maybe he'll take the log of this or square that or interact this. And he'll keep doing that until "the stars align." And so, on the one hand, there's an exploration that goes on there and I'll talk about a way to do that exploration that I think is defensible. But, right now, just given what I told you about that process, there's no easily defensible theory that typically guides these decisions except for the search for the stars. And when you're choosing which coefficients to present based on whether or not they produce stars, the standard errors don't mean what you think they mean. And so, the typical interpretation of what the stars mean, of statistical significance, right, the conventional meaning doesn't apply when you've selected on the basis of a star search.

So, what are we assuming about this process when we make these claims? I'll show you one other example. Again, it's very common. There's the one cause. Now, I know, in terms of human nature, certain things about psychology, limited time in the media, we love single factor explanations. "Why did Obama win the elections? A racial shift." Okay, that's it. One factor. We love those simple stories; and in academics, we love them too. In the papers that I get to review, sometimes, focuses on a single factor. Maybe it was an income disparity or an economic inequality. The claim is that it had no effect in 2008. So, that would be the point of a paper, to use a survey to show that that had no effect. So, how do we know that had no effect? "Well, I ran a regression, I put a bunch of things in it and the coefficient on economic inequality wasn't significant. So, therefore, I know that it didn't have any effect."

Of course, as long as a bunch of other assumptions are satisfied, the only thing we know from that demonstration is that that particular variable does not have a statistically significant relationship to the dependent variable *in that model*. And so, a lot of the interpretation of the claim depends on how robust do you think the result is to other variations of the model. Now, again, for a lot of folks reading this chapter, this point is rather elementary. We all know about this and maybe we're saying, "What if the model doesn't represent the true data generating function?" So, the point here is not to say this is a new criticism. The point here is to say these are probably things that none of us would be either caught dead doing or, if we have done it now, we'd say, "Okay, well, that was in the past and I wouldn't do it again." Great. Now let's turn our attention to things that we are doing that may be having comparable consequences.

So now let's turn to the production of survey data, some claims that we make and some assumptions that we make. So, in every survey, the numbers just don't appear, as we all know. A complex decision sequence turns dollars and human energy into a grid, into a set of data points. For many users of many surveys, including the gold standard surveys, there are many elements of this path that are not public. In the case of the ANES, when Jon Krosnick and I started to run it, we tried to find out how many things were done. We tried to identify documentation about ANES' decisions. In most cases, we couldn't find it. There was an explanation internally that they had some staff people who'd been there a long time and ANES relied on their institutional memory. And so, a common theme in our phone calls was if a particular ANES staff person gets hit by a bus, we're screwed. But, there was a bigger problem in that we were relying on her memory for some very intricate decisions, such that when we found the problems with the open-ended responses I write about in my other chapters in this volume. In that case we had no written evaluations or written instructions to go back to. So, when people then say, "Well, can you go back and fix the past? Can you take the new data and make it continue the time series of the old data?", part of the answer is, we can't with any certainty. I mean, we can pretend to, but we actually don't have documentation or instructions from the past. So, what are we assuming about this process?

So I think our problem can get down to this. Let's think about the ANES. Pick a year, you know, 2008. Take the 1,738th column and the 425th row in the grid. There's a 3 in it. So, here's our moment. What does that three mean? What does it mean? What is the proper interpretation of the number three in that part of the grid? Part of the answer has to do respondent actions. We have this wonderful, blooming field, the psychology of the survey response, that helps us think a lot about what the respondent is thinking about at a moment that might lead them to take an action that leads to a three at that point in the grid. But, the thing that a lot of our users can't access, have to make assumptions about, have to guess about is that three isn't a blood draw. The three isn't from saliva for us. The three is a product of actions that the respondents that took and actions that a lot of other people took before and after the respondent responded. The meaning of that three depends on some sequence of all of these actions. So, the question is, if our users, if people who depend on us for, you know, to try and derive meaning, you know, if they're trying to figure out what three means, what information are we giving them? So, there are a lot of steps in this process that affects the meaning of the number three. There's sample selection, respondent recruitment, how we do that. What kind of consent we got, consent for what? Question selection, what we put on, what we don't put on, what sequence, right? Question wording. There are response options. What instructions do we give to the interviewer? We had a very complex instrument. We had this theory about how our interviewers are going to deliver our survey. Then we sat in on one or two of them. We saw them acting in unexpected ways and had to ask each other, "what is happening?"

Shouldn't we be able to describe that process in great detail so that when somebody asks what a three means, we can tell him or her how it depends on directions to an interviewer. There's post-interviewer processing.There are evaluations. What evaluations did you do along the way to defend the story that you're telling me about why my data means what you say it means? So, limited introspection and documentation of procedures has the potential, and I think the reality we've experienced, to undermine the credibility and legitimacy of even the most famous surveys. So, that's the threat to us. What do we do?

The bigger question is, how can we help people? How can we help our analysts draw credible inferences? My proposal is to make more of it public. But, let me be more specific now. When I'm looking for guidance about, "How public should I be? How rigorous and detailed should I be?", I go back and ask Richard Feynman who, I met once at Cal Tech in the 1980s. And, Feynman talked about when we're in a battle with people, other scientists or people in the public sphere about meaning, our bet as scientists always has to be to double down on transparency because that's the source of our legitimacy. If we need people to trust us or to buy in on faith, then there's no, kind of, separation between us and many other people in the public sphere who are making claims about meaning. Our credibility depends on our willingness to be more transparent.

The other source that I love and go back to, I actually carry this book around with me, is a book by Santiago Ramón y Cajal, who, just for, you know, a lot of the basic theoretical work that led to things like fMRIs, he did about 100 years ago. He has this book that I show my students. It's called *Advice for a Young Investigator* and it's really about, procedurally, how to be honest about what you've done. And one of the key things is about how, when we get to a certain career stage or maybe when we're trying to build our reputations, we want people to see our brilliance. So, we focus on the end points and the wonderful things we've done. But, for science to advance, for people to really understand the meaning of what you've done, wouldn't it be better for science, maybe not for your reputation, but for science, to show them the errors; show them the steps; show them the things you thought would work that didn't work? It can be hard to do. We had a lot of debates with people in the building about what would happen if we showed the ANES user community that, at some point, we'd had a mistaken belief about what would work, or that we've ended up being very critical of past practices? Wouldn't that undermine the reputation of the project? Jon and I decided that we were always going to double down on being transparent. There was no question about it. Because the value of this study comes from people being able to believe it's meaning and I think this is the route to greater crediibility and legitimacy.

So, when I was at Cal Tech, there weren't many social scientists and I like to have friends and so, most of my friends were chemists. Now, in chemistry, I don't know if you might know about chemistry, but most of the papers, even the ones that end up in the main journals, they're three pages long, first of all, and they have a lot of authors on them. A reason for this is that you write papers *as a lab.* The reason you write paper as a lab is because you have people coming in and out, doing parallel activities to try and help the whole group come up with an insight. A key thing that makes that work is a lab book. Now, lab books are electronic now, but, you know, in the 1980s when I was at Cal Tech, there was actually a physical book where, when you were preparing an experiment, when you were writing it or so forth, you'd actually write in the book what you were doing and you would tell the next group, "Here's what we're doing." And when you came in the lab, the first thing you did, you didn't go to the bench. The first thing you did is go to the book, so you could figure out what was being done and why. I got to read the book, you know, because I did some studying in the lab because that's where my friends were.

So, what do you do in a lab book? You state the theory. You talk about how specific hypotheses are derived from a theory, so a theory is a grand idea and then, a hypothesis is something you want to test. And then, you lay out some criteria for the things you could observe and how they're going to effect the truth value of the hypotheses you've thrown out there. And then, you state an empirical model in advance. You say in advance what the analysis is going to be, so at least, in chemistry at least, the group I worked with, you don't get to change the analytic model afterwards because you don't get the result you want. What you do is you state it beforehand.

One of the great things that's happening now, I think, in psychology and political science as well, is the evolution of experimental registries where the idea is if you're going to run an experiment and you want a higher degree of credibility, what you agree to do is before you run the experiment, you put the design in the registry so that then, afterwards when you release a piece of data or a result, we can go back and look at, "Was this the first design or was it the nth design?" And if it was the nth design, what can we learn, not just from the nth experiment that you ran, but also from one through N minus one. Are there things there that might help us understand the robustness of your claim or things like that?

So, finally, if you see something disconfirming, you write down what you think's going on so that the rest of the group can see your idea of why it went wrong. Somebody else might read it. You have a group meeting. You document what happened there. And you just keep doing this for every subsequent observation so that months and years later, people can go back and there's, you know, usually a series of lab books. Now, you don't have to wonder, "Why did we do that?" or, if somebody asks you, "What does this experiment mean? Did you ever think about another variation?", we can go back to the book and say, "Yeah, we ran that and the result was the same" or "We ran that and the result was different and here's how we changed our understanding."

This is something that we can do, right? If we believe that our ability to go to Congress, to go to the outside world and say, "Our meaning is the one that you should spend money on, the one that you should rely on", then we need to be able to explain and defend our processes and meaning. The more we do this, the more credibility, legitimacy, and the more leverage we have.

So, I can think about our experience, as a producer of an election survey. It starts with an ideal of what we're going to do. And then, reality hits. We look at our budget and time happens and our interviewers aren't doing what we want and so forth, so we have to change. We ought to let people know how we made those changes. The problem is if you decide to do it at the end, it's too late and then it seems very threatening. We have to do it from the beginning. So, in addition to producing the data, we produce lots of documentation. There are some people who are annoyed with us for the amount of documentation we've produced on the ANES. And maybe some of it'll never be read. But, there are so many decisions and so many things that came out of that dataset that people want to know, "What does it mean? Does it really mean what you said it means?" Now, they have thousands of pages of documents where they can go through and find out how we got to where we ended up.

So, that's our issue. Can analysts draw credible inferences from our data and analyses? Can government do it? It's not magic and we should make sure that, to the extent that other people understand this, they can see the value of what we do. There is a more general phenomenon about transparency. One of the initiatives I've been involved with, DART, is data access and research transparency. A group of us, Colin Elman, principally, and George Alter and I, we just went to the American Political Science Association to rewrite its ethics guide. And so, the status quo has changed in political science such that now if you make a claim based on evidence, the expectation is that you'll provide the data and the documentation that got you from the data to your claims. That is the expectation. That's the status quo expectation – or explain why you can't. Up to now, the situation has been Request:"Will you please give me your data?" Response:"You know, well, I don't really have it," you know? Request:"Will you show me the Stata code that got from the data?" Response:"Well, I don't really have it. " As you may know, in political science, there have been graduate courses run around the U.S. and in Western Europe where people try to replicate the claims made in the top journals and the success rate of that is about the same as I understand it's becoming in social psychology. It's pretty bad. Don't read into this as saying that I think people are doing fraudulent work because I don't, but I think they're really bad at documentation. They can't remember three years earlier that they took the log of this or that they squared that or that, you know, something of that nature. It's bad documentation.

One of the things we're trying to do next is figure out how do we change the incentives of individual investigators to give them an incentive to document from the beginning and here's the basic idea. A lot of journals use editorial manager, the last stage before publication, an editorial manager. As you press a button, it goes to the publisher. You could rewrite that program and have the document go to an archivist. Your editorial team, your journal, could have sent to this archivist a set of its requirements for data access, whether you need to produce the analytic code and so forth, like, what your requirements are with lags built in, whatever. And the archivist says, "The data are on file. The explanation is on file. This data are on file." And when they press that button, then it goes to the publisher. So, if everybody knows that that's the sequence – now, if you're starting an investigation and you want to publish one of the top journals in political science, you know that if you don't have your documentation in order, you're not going to get published. So, that's a very short story, but we're trying to do that and the main reason we're trying to centralize and automate transparency-increasing practices is that when you're editing a typical journal, it takes 25 hours a day and 8 days a week. So, you know, doing this type of thing is something editors would like to get to, but they just can't. But, since so many journals have the same problem and since so many journals want to be part of a more transparent movement, we're now trying to build something for them that automates it, that has a central archivist. There are a lot of issues there.

Procedural transparency is something that I think benefits us. At the ANES, we did a lot of the stuff to make things very transparent. One of the things we did is open the ANES Online Commons, which allowed anyone to propose questions. And we and our board work with people to refine those questions and all the new questions that we put in the ANES, the time series and the panel study, originated in our Online Commons. We then wanted to tell the user community and the scientific community about that process. With the help of John Aldrich, Kathleen McGraw, they wrote a book that describes many of the questions that we derived through the Online Commons and, basically, gives their histories of how questions were developed, how they were evaluated and why they ultimately were or were not included on the production studies, just to give, again, users a sense of what they meant, what this data means.

In conclusion, there are real threats to our credibility and legitimacy. Congress occasionally asks questions about the value of social science, in general, or specific projects. We always have to be prepared to engage the question of, "Why should the government pay for social science rather than the other things it can buy?" We need outcomes that are relevant and that's clear. But, ultimately, if there are contests about whether our meaning has value and is different from others, if we give up – if we don't pay enough attention to investing in our own credibility and our legitimacy, we're really pulling the rug out from under ourselves and hurting the people at NSF who are trying to make the case for us. So, I think that these investments and our own credibility and transparency, I think it really matters. It allows us to talk about meaning. It allows us to tell stories about value.

Arthur Lupia is the Hal R. Varian Professor of Political Science at the University of Michigan and Research Professor at its Institute for Social Research. He examines how people learn about politics and policy and how to improve science communication. His books include *Uninformed: Why Citizens Know So Little About Politics and What We Can Do About It*.

He has been a Guggenheim fellow, a Carnegie Fellow, is an American Association for the Advancement of Science fellow, and is an elected member of the American Academy of Arts and Sciences. His awards include the National Academy of

Sciences Award for Initiatives in Research and the American Association for Public Opinion's Innovators Award. He is Chair of the National Academy of Sciences Roundtable on the Application of the Social and Behavioral Science and is Chairman of the Board of Directors for the Center for Open Science.

Index

A
Accuracy
  bias, 20, 38, 116, 206, 221, 222, 228, 229, 257, 258, 264, 290, 316, 349, 377, 379, 523, 565, 607
  record check, 112, 139, 467, 468
  reliability, 65, 99, 257, 544, 597
  validation, 267, 269, 270, 294, 347, 543, 544, 589
  validity, 11, 213, 214
  variance, 38, 214, 468, 543, 587
Address-based sampling (ABS), 11, 13–17, 49, 67, 202, 363, 365, 370, 375, 413, 549
Advertising Research Foundation (ARF), 9, 357
Amazon Mechanical Turk (Mturk)/ Mechanical Turk, 358
American Association for Public Opinion Research (AAPOR), 9, 29, 30, 34–36, 182, 331, 353, 358–359, 361, 375, 546, 612
American Community Survey (ACS), 53, 72, 73, 155, 369, 398, 400, 417, 419–420, 426, 431–432, 509, 524
American National Election Studies (ANES)/National Election Studies (NES), 19, 63, 99, 121–126, 135, 155, 161, 162, 290, 293, 295, 298, 304–307, 311–313, 446, 447, 473–483, 485, 486, 499, 509–511, 520–523, 526, 527, 574–575, 624, 637, 639, 641, 647, 648, 649, 651, 653, 656, 658, 660, 662, 663, 665
Anonymity/Anonymous/ Anonymously, 136, 137, 184, 270–271, 277, 336, 425, 501–506, 638
Assumptions, 38, 76, 98, 113, 122, 171, 246, 259, 264, 268, 282, 315, 331, 339, 383, 428, 432, 448, 578, 601, 657–660
Attention, 3, 26, 48, 49, 55, 58, 66, 103, 115, 122, 137, 143, 144, 194, 203, 231, 245, 260, 282, 299, 312, 327, 365, 373, 377, 447, 458, 462, 466, 470, 476, 522, 524, 543, 580, 597, 599, 602, 603, 625, 628, 630, 642, 649, 660, 665

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6

Attrition, 65, 143–150, 155–157, 201, 336, 408, 509–516, 584
Audio-computer-assisted self-interview (ACASI), 218, 400, 552

B
Behavioral data/Organic data, 176, 177, 186, 222, 628, 633
Behavioral Risk Factor Surveillance System (BRFSS), 369, 397, 398
Behavior coding, 335, 591, 594–595
Big data, 59, 175–189, 298–299, 607, 631–633
Biomarkers
  blood, 205
  DNA, 228, 230, 231, 574, 584, 586
  dried-blood spot (DBS), 230, 579, 580, 583
  saliva, 205
Bogus pipeline, 138, 505
British Household Panel Survey (BHPS), 156, 510, 511, 600
Bureau of Labor Statistics (BLS), 459, 462, 555, 569

C
Calendar, 112, 115, 156, 235–239, 338, 470, 472, 555, 588–595
Centers for Disease Control (CDC), 258, 295, 369, 556, 557
Cluster sample/Cluster sampling/ Cluster, 45, 278, 376, 396, 469, 519, 521
Codebook, 304, 360, 636, 649, 652
Cognitive interviewing/Cognitive testing, 103–106, 457, 458, 463
Cognitive response process, 47
  optimal response process, 532
Commercial database/Consumer database
  Aristotle, 177, 290, 505, 621
  Catalist, 177, 275–287, 290, 621
  Experian, 16, 177, 609
  InfoUSA, 16, 263, 609
Computer-assisted personal interview (CAPI), 44, 46, 73, 110, 236, 246, 247, 394, 400, 419, 420, 465, 466, 543, 551, 552
Computer-assisted self interview (CASI), 244, 246, 247, 323, 394, 400
Computer-assisted telephone interview (CATI), 43, 45, 247, 324, 395, 539, 589, 590
Computer/Computerized
  laptop, 52, 56, 402, 558
  PC, 52, 53, 56, 57, 326, 558
Confidentiality/Confidential, 135–138, 184, 196, 229, 231, 259, 299, 336, 490, 499, 501–506, 526, 542, 585, 594, 599, 601–603, 605, 630
Consumer Expenditure Survey (CES), 183, 555
Contacts/Contact attempts
  contact rate, 146, 375, 378
  non-contact, 30, 65, 143, 146, 533
Conversational interviewing/ Conversational norms/ Conversational, 96, 97, 110–112, 116, 139, 236, 333, 334, 441–443, 466, 467, 471, 472, 591–593
  rapport, 593
Cost(s), 8, 9, 15, 20, 23, 26, 27, 29, 30, 35, 36–39, 43, 45, 47, 48, 49, 51, 54, 58, 63, 64, 67, 81, 83, 88, 97, 105, 109, 132, 146, 147, 155, 157, 160, 166, 167, 168, 171, 183, 214, 215, 224, 225, 230, 231, 237, 238, 260, 263, 276, 286, 290, 297, 298, 303, 307, 321, 332, 333, 337, 338, 356, 360, 363, 369, 371, 383, 384,

396, 407, 408, 409, 412, 414, 463, 483, 491, 494, 495, 499, 511, 514, 515, 516, 520, 529, 541, 546, 568, 576, 580, 584, 605, 607, 608, 629

Coverage, 3, 8, 11, 13–16, 33, 43–46, 51–54, 76, 170, 179, 183, 186, 207, 268, 276, 319, 332, 353, 362, 365–368, 370, 377, 388, 393, 395–398, 417, 423, 457, 479, 520, 566, 585, 608, 612, 630, 632

Cross-cultural/Cross-national/International, 29, 53, 55, 79, 83, 85, 105, 106, 167, 199, 200, 201, 204, 268, 270, 272, 463

Current Population Survey (CPS), 72, 87–88, 155, 306–307, 312, 378–379, 397–398, 417, 419, 422, 428, 509, 520, 524, 610, 616, 617

D

Database/Data repository, 15–17, 20, 166, 176–177, 179, 184, 205, 213, 214, 244, 258–259, 263–265, 268, 275–287, 289–291, 294, 305, 320, 367–370, 378–379, 381, 586, 598, 600, 611, 621, 623–625

Data Documentation Initiative (DDI), 305, 307, 312, 339, 639, 642, 643

Data quality, 9–11, 23, 30, 38, 45, 51, 57, 58, 59, 66, 74–76, 88, 99, 110, 115, 126, 137, 170, 176, 178, 212–214, 218–219, 231, 238, 246, 248, 259, 276, 277, 316, 332, 351, 356, 424, 446, 452, 469, 540–542, 546, 553, 556, 560, 582, 589, 591–594, 607, 636

Delivery Sequence File (CDSF/DSF), 14–17, 363, 366–368, 370, 609, 612

Demographic/Demographics
- age, 524, 601
- education, 66, 79, 80, 353, 361, 409, 428, 616
- employment, 428
- gender, 259, 361
- income, 16, 25, 79
- marital status, 616
- sex, 524, 601

Design effect, 525

Digital Object Identifier (DOI), 307, 339, 641

Disclosure, 11, 88, 111, 137, 162, 167, 265, 355, 360, 362, 503, 523, 525–527

Documentation, 124–126, 194, 270, 276, 303–307, 309–312, 315–317, 339, 175, 178, 179, 482, 483, 486, 522, 636, 639–640, 642, 646, 647, 649, 651, 652, 656–658, 660–661, 663–664

Document retrieval, 489

E

Ecological Momentary Assessment (EMA), 206, 221–225, 237, 337, 563–570

Educational Longitudinal Survey (ELS), 156, 157, 510, 511

Ethics/Ethical, 167, 168, 169, 171, 185, 231, 500, 526, 576, 664

European Social Survey (ESS), 30, 33, 37, 80, 213, 533, 543

Experience sampling, 206, 221, 237, 337, 563–570

Experiment, 20, 26–27, 36, 57, 58, 59, 63, 64, 67, 68, 76, 83, 157, 178, 200–203, 206, 207, 264, 319–320, 322–323, 325, 326, 358, 375, 385, 394, 406–411, 425, 443, 451, 453, 469–470, 472, 494, 497, 505, 514–515, 532, 569, 574, 580, 611, 622, 658, 662, 663

F

Face-to-face (FTF)/In-person, 14–16, 30, 33, 35, 39, 43, 45, 51, 53–54, 64, 105, 109, 136, 144, 146, 181, 195, 196, 201, 212, 213, 218–219, 246, 247, 321, 323–325, 332, 333, 337, 369–371, 375, 384, 393–396, 400, 401, 419, 503, 519–521, 539, 544, 547, 555, 556, 576, 589–590, 599, 624

G

Gallup, George, 7, 567, 570
General Social Survey (GSS), 19, 33, 63, 68, 88, 155, 242, 244, 298, 304, 305, 306, 307, 311, 312, 361, 375, 413, 509, 510, 639, 640, 641, 642, 645, 647, 648, 649, 651, 653
GfK Knowledge Panel (GfK)/ Knowledge panel, 199, 557, 607, 609, 611, 612, 616–617
GPS, 58, 59, 165, 166, 205, 219, 558
Grid question type/Matrix, 55–57

H

Health and Retirement Study (HRS), 202–203, 227–228, 397, 554, 569, 573–574, 576, 580, 584, 585, 586

I

Imputation, 177, 229, 264, 356, 540, 578, 610, 611, 613, 616–618
Incentive(s)
  differential incentives, 66–67, 411
  monetary incentive, 24, 63, 64, 67, 68, 157, 169, 332, 385, 406
  non-monetary incentive, 406
  prepaid incentive, 34, 63–67, 325, 327, 333, 406, 407, 410–411, 413
  promised incentives, 64, 327, 333, 406, 407, 410
  token incentive, 24, 385
Informed consent/Consent, 167, 244, 258, 259, 269, 270, 500
Institute for Social Research (ISR), 441, 510, 569
Institutional Review Board (IRB), 244, 414, 499, 526
  informed consent, 244
Integrated Public Use Microdata Series (IPUMS), 155, 304, 305, 509, 638, 640
International Social Survey Programme (ISSP), 80
Internet survey/Web survey/Online survey, 8, 24, 29, 34, 35, 39, 44, 46–48, 51–53, 55–58, 105, 178, 181–182, 194, 218, 244, 246, 247, 324–327, 335, 337, 357, 359, 384, 394, 395, 396, 402, 410–411, 431, 433, 530, 532, 552
Inter-University Consortium for Political Science Research (ICPSR), 306–307, 312, 630, 632, 638, 640–641, 643, 645, 649, 650, 652

Interviewer
  interviewer effects, 45, 116, 235, 246, 247, 394, 396, 468, 587, 603
  interviewer observations, 30, 178, 194–195, 211–215, 337, 534, 537–547
  interviewer training, 24, 97, 116, 213, 229, 231, 338, 412, 472, 540, 543, 581, 604
Item count technique (ICT), 138, 505
Item nonresponse, 57, 65, 66, 178, 247, 321, 322, 325, 326, 399, 408–409, 411
  missing data, 179

L

Linking/Linked
  ancillary data, 263, 611, 618, 619
  auxilary data, 26, 58, 148, 212–214, 338, 530, 532, 539, 543, 544, 545, 546, 547, 610, 618, 630
  external data, 16, 148, 251, 368–369, 607–619, 625
  linkage, 115, 170, 257–260, 268–272, 300, 545, 597–605, 652
  records, 260, 597, 601
Longitudinal Internet Studies for the Social Sciences (LISS), 53, 54, 56, 170, 199, 200–205
Los Angeles Family and Neighborhood Survey (L.A. FANS), 211, 538, 574

M

Mapping, 99, 140, 316, 363, 365, 367, 368, 432, 446, 611
Market research, 57, 64, 181–182, 184, 185, 257, 263, 264, 275, 354, 406
Matching/Matched/Matches
  exact matching, 259
  probabalistic matching, 258, 259, 598, 600, 601, 602, 605
  statistical matching, 259, 260, 269, 598, 601–602
Measurement error/Response error, 16, 45, 46, 48–50, 52, 54, 55, 57–59, 103, 138, 139, 178, 194, 219, 243, 319, 332, 378, 396, 399, 401, 402, 457, 458, 505, 524, 529–530, 560, 582, 598, 631
Memory/Recall/Remembering, 58, 90, 96, 104, 122, 124, 131, 155, 165, 168, 183, 184, 218, 221–223, 235–238, 245, 267, 294, 311, 337, 338, 431, 434, 440, 441, 459, 471, 472, 474, 476, 477, 479, 481, 492, 500, 512, 514, 515, 523, 532, 553, 555, 557, 559, 560, 564–567, 579, 587–592, 594–595, 614, 624, 648, 660
  retrieval, 236
Meta-analysis, 3, 20, 36, 39, 47, 64, 89, 320, 323, 326–327, 387, 400, 407, 429, 502, 505
Mixed-mode/Multi-mode/Mixed-device
  concurrent, 53, 56
  sequential, 53–54
Mobile/Cell phone
  mobile device, 49–50, 52, 56–58, 77, 207, 420, 423, 424, 558
  mobile survey, 34, 56–59
  smartphone, 52, 165, 167, 168, 170, 177, 181, 185, 205, 218, 219, 402
  tablet, 35, 50, 52, 56, 57, 170, 202, 218, 354, 372, 402, 471, 556, 558

Mode/Administration mode/Survey mode
face-to-face, 39, 53, 547
mail, 64, 326
mobile, 49, 50, 52, 55, 420
self-administered, 47, 105, 324, 325, 326, 396, 466
telephone, 49, 324, 401
unimode, 48, 49, 54, 332, 401
web/internet, 203, 410, 424
Monitoring the Future, 136, 503
Motivation, 65, 68, 88, 90, 96, 111, 115–116, 126, 147, 194, 284, 333, 366, 425, 451, 470, 471, 646

N
National Center for Health Statistics (NCHS), 258, 457, 459–461, 556, 598
National Health Interview Survey (NHIS), 212, 542, 577
National Health and Nutrition Survey (NHANES), 227, 556, 573, 575–577, 580, 582
National Institutes of Health (NIH), 585
National Longitudinal Study of Adolescent to Adult Health (Add Health), 88, 243, 574, 640
National Science Foundation (NSF), 19, 20, 45, 49, 63, 139, 155, 162, 204, 300, 305, 306, 307, 317, 374, 375, 381, 427, 431, 445, 455, 463, 478, 486, 509, 512–513, 529, 530, 534, 535, 629, 631–633, 636–638, 640, 642–643, 645, 656, 665
National Social Like, Health, and Aging Project (NSHAP), 574, 582
National Survey of Family Growth (NSFG), 211–213, 532, 538–541, 543–546
Non-probability sampling/Nonprobability
online sampling, 9, 11
opt-in, 8, 9, 186, 354–359, 396, 397
self-selection, 7, 58
Non-response/Nonresponse
biased nonresponse, 20, 349, 376
breakoff, 24, 34, 115, 325, 326
non-cooperation, 35, 144
nonrespondent, 20, 21, 25–27, 36, 67, 212, 321, 322, 325, 370, 371, 378, 387–390, 412
nonresponse bias, 19–21, 23–27, 36–38, 63, 66–68, 135, 145, 160, 184, 195, 229, 230, 231, 257, 264, 319, 324, 331, 332, 333, 337, 376, 380, 383, 386–391, 410, 413
refusal conversion, 20, 24, 30, 37, 64, 65, 66, 67, 332, 377, 384, 385, 386, 407, 411, 413
refusal, 63

O
Official records/Government records/Administrative records, 49, 213, 257–260, 263, 267–272, 289, 294, 295, 299, 338, 347, 349, 399, 524, 530, 539, 543, 545, 597–605, 621–626, 652
Open-ended questions/Open ends/Open-ended data
text data, 176, 195, 533
coding open-ended, 126, 130, 335–336, 473

P

Panel Study of Income Dynamics (PSID), 19, 63, 155, 156, 212, 304–307, 311–313, 509–511, 513, 516, 539, 541, 542, 584, 589, 590, 637, 639–642, 645, 647, 648, 651, 653

Panel/Survey panel/Longitudinal study
- panel attrition, 143–150, 156, 157, 336, 509–516
- panel conditioning, 206

Paper-and-pencil interview (PAPI)
- mail, 14–16, 24, 34, 35, 39, 43–49, 51–54, 57, 64, 66, 75, 144, 146, 181, 201, 202, 218, 247, 276, 277, 279, 283, 285, 287, 321, 323–327, 337, 365–371, 384, 386, 393, 394, 396, 400, 402, 407, 409, 413, 421, 504, 551, 552, 554, 556, 557, 612, 624

Paradata, 146, 148, 150, 178, 193–196, 211, 336–337, 529–535, 538, 540, 591, 593–595, 610, 652

Participation, 24, 33, 35, 65, 67, 68, 72, 87, 88, 144–148, 150, 168, 206, 223, 229, 295, 298, 299, 336, 359, 412, 417, 497, 505, 555, 577, 602, 640

Pew Research Center/Pew, 199, 356, 373, 375, 376, 388, 557, 607

Pretesting, 67, 104, 105, 335, 412, 452, 463, 652

Privacy, 46, 59, 121, 135–138, 167, 169, 184–185, 218, 399, 476, 490, 499, 501–506, 553, 560, 603

Probability sampling/Probability sample/Probability-based
- nationally representative, 227, 297, 542, 573, 576
- random sample, 170, 171, 353, 360, 388, 433
- representative sample, 8, 89, 227, 263, 359, 361, 429, 608
- simple random sample, 511

Probe/Probing, 55, 97, 103–105, 112, 236, 245, 420, 422, 423, 458, 459, 461–463, 468, 469, 591, 592, 593, 594

Proxy report/Proxy, 87–91, 212, 222, 242, 245, 327, 333, 387, 427–435, 510, 542, 612, 617

Public Opinion Quarterly (POQ), 104, 353, 376, 462, 544

Q

Quality/Response quality, 8–11, 14, 19, 23, 30, 38, 45, 51, 56–59, 63, 65, 66, 85, 88, 110–112, 115–116, 126, 132, 137, 165, 167, 168, 170, 176, 178, 182, 185–186, 195, 199, 207, 212–215, 218, 219, 230, 231, 235, 238, 245, 246, 248, 259–260, 264, 276–277, 290–291, 294, 297, 300, 316, 319–325, 327, 332, 351, 355–360, 362, 364, 366, 367–368, 370, 381, 386, 390–391, 397, 406, 408, 409, 412, 414, 423, 424, 429, 446, 448, 449, 452, 467–468, 478–479, 483, 492, 503, 532, 535, 539–547, 553, 556, 560, 580, 582, 587, 589, 591–594, 598, 600–602, 605, 607, 609, 618, 624, 625, 627, 636, 656

Questionnaire, 8, 17, 24, 26, 34, 43, 44, 48, 53–55, 58, 59, 82, 95, 96, 100, 103–105, 117, 136, 146, 165, 178, 183, 199, 201–203, 206, 213, 222, 223, 237, 239, 244, 247, 248, 250, 286, 323, 332, 333, 335, 337, 338, 352, 355, 362, 384, 394, 413, 433, 434, 439–455, 457–463, 503, 504, 531, 542, 554–559, 565, 566, 589, 594, 595, 603, 639

R

Random digit dial (RDD), 45, 49, 359, 364, 375, 385, 395, 397, 402, 407, 413, 539, 567, 608, 622
  telephone survey, 364
Random error, 160, 245, 527
Randomized response technique (RRT), 138, 140, 504, 505
Random sampling error, 45, 396
Recruitment/Recruited/Recruiting, 25, 34, 35, 47, 48, 53, 54, 66, 68, 103, 109, 156, 170, 200, 201, 202, 207, 211, 221, 230, 316, 379, 465, 523, 555, 556, 574, 576, 581, 607, 612, 614, 661
Regression, 55, 398, 474, 475, 490, 493, 545, 658, 659
Repeated cross-section, 511
Replication/Reproducibility/Replicability, 104, 126, 162, 167, 200, 316, 388, 483, 496, 498, 499, 525, 526
Respondent burden/Burden, 24, 44, 47, 54–56, 58, 76, 117, 144, 146, 166–171, 184, 218, 219, 223, 225, 230, 237, 244, 246–248, 250, 260, 270, 375, 385, 399, 422, 452, 552–554, 557, 559, 560, 567, 568, 605
Respondent characteristics/Characteristics, 4, 9, 20, 21, 25, 26, 36, 56, 57, 66, 81, 90, 91, 110, 136, 138, 145, 149, 169, 175, 176, 193–194, 200, 217, 241, 245, 246, 269, 315, 323, 325, 337, 361, 365, 366, 367, 377, 381, 387, 395, 399, 409, 413, 431, 466, 467, 516, 520, 521, 526, 530, 531, 543, 550, 555, 569, 593, 594, 603, 604, 622, 637
Response bias, 39
Response rate(s) (RR)
  completion rate, 9, 156, 324, 357, 512–513, 515
  contact rate, 146, 375, 378
  cooperation rate, 87, 146, 321, 375, 378, 386, 428, 554, 555, 557, 581, 603
  response propensity, 20, 25, 212, 213, 294, 332, 387, 539, 541
Response scale/Rating scale
  number of scale points, 367, 510, 559
  scale labels, 99, 100, 334, 440, 447–450
RTI, 367, 510, 559

S

Sample size, 8, 45, 67, 104, 183, 258, 381, 413, 459, 461, 521, 541, 586, 599
Sampling
  address-based sampling, 11, 13–17, 49, 67, 202, 363, 365, 375, 413, 549
  non-probability based sampling, 354
  probability-based sampling, 7–11

quota sampling, 7, 353, 359
random sample, 170, 353, 360, 388, 433, 511
representative sample, 8, 89, 227, 263, 359, 361, 429, 608
sampling bias, 45, 139, 397
sampling frame, 11, 13, 14, 17, 45, 46, 53, 178, 223, 242, 263, 264, 270, 353, 364–368, 370, 395, 529, 539, 545, 609, 610, 618
simple random sample, 511
Sampling frame/Frame
coverage, 46, 53, 395, 396
coverage error, 46, 53, 395, 396
overcoverage, 13–15
undercoverage, 13, 15
Satisficing
acquiescence, 440, 504
don't know, 96
primacy, 45
random responding, 441, 504, 505
recency, 45
response order, 45, 396
straightlining, 55
strong satisficing, 96
weak satisficing, 96, 441
Scientific, 7–9, 25, 76, 85, 126, 131, 147, 162, 200, 260, 271, 304, 315, 317, 319, 320, 352, 428, 459, 473, 478, 491, 497, 524, 525, 605, 637, 641, 657, 658, 665
Show card, 48, 49, 194, 307, 338, 401, 651
Social desirability
misreporting, 139, 322
sensitive questions/topics/information, 218, 503, 553
Social media/Social network
facebook, 10, 170, 176, 182, 184, 247, 249, 250, 360, 361, 499, 633
twitter, 10, 170, 180, 184, 188, 249, 360–362, 499, 558, 559, 655
Socio-economic status (SES), 79, 84, 409, 575
Software, 56, 170, 182, 184, 193, 247, 248, 258, 306, 307, 320, 365, 367, 368, 467, 496, 525, 530, 600, 640, 641
Standardization
conversational interviewing, 112, 236
interviewer behavior, 109
interviewer deviations, 335
survey script, 111
Stratification, 16, 21, 45, 87, 359, 367, 520, 521, 524, 617
Survey of Consumers, 349, 376, 390
Survey of Health Ageing and Retirement (SHARE), 268

T

Theory
leverage salience, 24, 386
satisficing, 441, 446, 574
Total Survey Error (TSE), 38, 54, 58, 103, 139, 178, 183, 185, 194, 380, 395, 457, 529, 530
Transcript/transcribe/transcription, 97, 124, 257, 335, 468, 476–478, 481, 490, 591, 594
Translation/Translational/Translating, 79, 80, 105, 129, 130, 206, 489, 490, 494, 497, 608
Transparency, 10, 11, 85, 125, 126, 161, 162, 205, 260, 264, 265, 276, 277, 294, 312, 315–317, 339, 475, 481–483, 486, 523, 525, 526, 605, 655, 657, 661, 664, 665
Turnout, 3, 4, 88, 275, 276, 283, 284, 287, 289, 290, 293–295, 347, 349, 574, 578, 621

U

Understanding America Study (UAS), 199, 200, 202–204

University of Michigan/Michigan, 180, 184, 358, 376, 394, 397, 441, 473, 510, 525, 530, 534, 542, 569, 576, 598, 642

U.S. Census Bureau/Census Bureau/Census, 7, 10, 47, 48, 71, 72, 73, 76, 80, 82, 83, 87, 155, 177, 178, 180, 183, 184, 258, 281, 304, 352, 359, 363, 364, 367, 370, 411, 417–426, 428, 431, 459, 461, 463, 501, 503, 509, 510, 512, 513, 515, 526, 529, 532, 534, 555, 558, 600, 624, 630, 633, 635–637, 656

V

Validation, 89, 186, 213, 219, 230, 267, 269, 270, 276, 289, 290, 294, 295, 347, 348, 355, 484, 486, 543, 544, 582, 589, 590, 622

Volunteering, 20, 379, 380

Voter file, 276, 278–281, 289–291, 294, 621–625

W

Web panel/Online sample, 8, 9, 394, 396, 397, 607–619

Weighting/Survey weight
- post-stratification, 21, 521
- post-stratifying, 160
- rake weighting, 525

Westat, 367, 460