The SAGE Handbook of Survey Methodology

Edited by
Christof Wolf, Dominique Joye,
Tom W Smith and Yang-chih Fu

The SAGE Handbook of
Survey Methodology

SAGE was founded in 1965 by Sara Miller McCune to support the dissemination of usable knowledge by publishing innovative and high-quality research and teaching content. Today, we publish over 900 journals, including those of more than 400 learned societies, more than 800 new books per year, and a growing range of library products including archives, data, case studies, reports, and video. SAGE remains majority-owned by our founder, and after Sara's lifetime will become owned by a charitable trust that secures our continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

The SAGE Handbook of
Survey Methodology

Edited by
Christof Wolf, Dominique Joye,
Tom W. Smith and Yang-chih Fu

SAGE reference
Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Mila Steele
Editorial Assistant: Mathew Oldfield
Production Editor: Sushant Nailwal
Copyeditor: David Hemsley
Proofreader: Sunrise Setting Ltd.
Indexer: Avril Ehrlich
Marketing Manager: Sally Ransom
Cover Design: Wendy Scott
Typeset by Cenveo Publisher Services
Printed and bound by CPI Group (UK) Ltd, Croydon, CR0 4YY

At SAGE we take sustainability seriously. Most of our products are printed in the UK using FSC papers and boards. When we print overseas we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

Library of Congress Control Number: 2015960279

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-4462-8266-3

Contents

List of Figures

List of Tables

Notes on the Editors and Contributors

THE EDITORS

Christof Wolf is acting president of GESIS – Leibniz Institute for the Social Sciences and Professor of Sociology at Mannheim University. He is currently representing the Secretariat of the International Social Survey Programme (ISSP) and is a member of the Executive Committee of the European Values Study (EVS). His main research interests include sociology of religion, social networks, sociology of health, and survey methodology. He is co-editor of the *SAGE Handbook of Regression Analysis and Causal Inference* (2015).

Dominique Joye is Professor of Sociology at the University of Lausanne and associated to FORS. He is involved in the analysis of inequality and life course, and is participating in NCCR LIVES in Switzerland; part of this handbook was also realized in this frame. He has published many papers in this area as well as defining the way that social-professional positions are measured in Switzerland by the Swiss Statistical Office. He is also interested in comparative surveying, and is a member of the methodological advisory board of the European Social Survey (ESS), of the executive and methodological committees of the European Values Study (EVS), and Chair of the methodological committee of the International Social Survey Program (ISSP).

Tom W. Smith is Senior Fellow and Director of the Center for the Study of Politics and Society of NORC at the University of Chicago. He is Principal Investigator and Director of the National Data Program for the Social Sciences which conducts the General Social Survey and collaborates with the International Social Survey Program. He studies survey methodology, societal trends, and cross-national, comparative research.

Yang-chih Fu is Research Fellow at the Institute of Sociology and former Director of the Center for Survey Research, Academia Sinica, Taiwan. He is Principal Investigator of the Taiwan Social Change Survey, a large-scale survey series launched in 1984. His recent research focuses on the social desirability effects that occur during face-to-face interviews, as well as multilevel analyses that use contacts as the building blocks of interpersonal ties and social networks.

THE CONTRIBUTORS

René Algesheimer is Professor of Marketing and Market Research and Director of the University Research Priority Program 'Social Networks' at the University of Zurich. His main research interests lie in social networks, social media and the consequences of the digital transformation on firms and individuals. He has authored several articles in leading journals of the field, such as *Marketing Science, Journal of Marketing Research, Journal of Marketing, Harvard Business Review or Sociological Methods and Research*.

Duane F. Alwin is the inaugural holder of the Tracy Winfree and Ted H. McCourtney Professorship in Sociology and Demography, and Director of the Center for Life Course and Longitudinal Studies, College of the Liberal Arts, Pennsylvania State University, University Park, PA. He is also Emeritus Research Professor at the Survey Research Center, Institute for Social Research, and Emeritus Professor of Sociology, University of Michigan, Ann Arbor. In addition to his interest in improving survey data quality, he specializes in the integration of demographic and developmental perspectives in the study of human lives. His work is guided by the life course perspective, and his current research focuses (among other things) on socioeconomic inequality and health, parental child-rearing values, children's use of time, and cognitive aging.

Jean-François Beaumont is Chief in statistical research in the International Cooperation and Corporate Statistical Methods Division at Statistics Canada. He is responsible for the Statistical Research and Innovation Section. His main research projects and publications are on issues related to missing data, estimation, including robust estimation and, more recently, small area estimation.

Dorothée Behr is a senior researcher at GESIS – Leibniz Institute for the Social Sciences, Mannheim. Her research focuses on questionnaire translation, translation and assessment methods, and item comparability as well as cross-cultural web probing. Besides publishing in these fields, she provides consultancy and training in the wider field of cross-cultural questionnaire design and translation.

Nejc Berzelak is a researcher in the field of survey methodology at the Faculty of Social Sciences, University of Ljubljana. The main topics of his research include questionnaire design, measurement errors, mode effects, analysis of survey response process, and cost-error optimization of mixed-mode surveys. He is participating in several research projects related to the development of survey methods and works as a methodological consultant for surveys conducted by academic, governmental, and private organizations.

Jelke Bethlehem is Professor of Survey-methodology at the Leiden University in The Netherlands. Until recently he was also senior survey methodologist at Statistics Netherlands. His research interests are nonresponse in surveys, online surveys, and polls and media. He is author or co-author of several books about surveys.

Paul P. Biemer is Distinguished Fellow of Statistics at RTI International and Associate Director for Survey R&D in the Odum Institute for Research in Social Science at University of North Carolina. His main interests lie in survey statistics and methodology, survey quality

evaluation and the analysis of complex data. He is the author, co-author and editor of a number of books including *Introduction to Survey Quality* (Wiley, 2003) and *Latent Class Analysis of Survey Data* (Wiley, 2011).

Jaak Billiet is Emeritus Professor of Social Methodology, Centre of Sociological Research, University of Leuven. He combines methodological research with substantial longitudinal and comparative research on ethnocentrism, political attitudes and religious orientations. He is author or co-author of many published book chapters, articles in academic journals, and several co-authored books and edited volumes including *Cross-Cultural Analysis* (Routledge, 2011).

Jörg Blasius is a Professor of Sociology at the Institute for Political Science and Sociology, University of Bonn, Germany. His research interests are mainly in explorative data analysis, especially correspondence analysis and related methods, data collection methods, sociology of lifestyles and urban sociology. Together with Simona Balbi (Naples), Anne Ryen (Kristiansand) and Cor van Dijkum (Utrecht) he is editor of the SAGE series 'Survey Research Methods in the Social Sciences'.

Annelies G. Blom is Assistant Professor at the University of Mannheim and Principal Investigator of the German Internet Panel (GIP). Her research looks into sources of survey error during fieldwork, in particular interviewer effects and nonresponse bias. She is author and co-author of many peer-reviewed articles that appeared in scholarly journals such as *Public Opinion Quarterly*, *Journal of the Royal Statistical Society: Series A*, *International Journal of Public Opinion Research*, *Journal of Official Statistics*, and *Field Methods*.

Michael Braun is Senior Project Consultant at GESIS – Leibniz Institute for the Social Sciences at Mannheim and Adjunct Professor at the University of Mannheim. His main research interests include cross-cultural survey methodology and intercultural comparative research in the areas of migration and the family. He is co-editor of *Survey Methods in Multinational, Multiregional and Multicultural Contexts*.

Kristen Cibelli Hibben is a PhD Candidate at the University of Michigan Program in Survey Methodology and Research Assistant in the International Unit at the Institute for Social Research's Survey Research Operations. Her research interests include respondent motivation and data quality, cross-cultural survey research, and the application of survey methods in challenging contexts such as post-conflict or in countries with little survey research tradition. She has co-authored book chapters in the present volume as well as *Hard-to-Survey Populations* (Tourangeau, Edwards, Johnson, Wolter, and Bates, 2014) and *Total Survey Error in Practice* (Biemer, de Leeuw, Eckman, Edwards, Kreuter, Lyberg, Tucker, and West, 2017).

Jan Cieciuch is Project Leader of the University Research Priority Program 'Social Networks' at the University of Zurich. His interests are applications of structural equation modeling especially in psychology, with focus on the investigation of human values and personality traits. Recent publications appeared in leading journals such as the *Journal of Personality and Social Psychology*, *Journal of Cross-Cultural Psychology*, *Annual Review of Sociology*, and *Public Opinion Quarterly*.

Eldad Davidov is Professor of Sociology at the University of Zurich and president of the European Survey Research Association (ESRA). His research interests are applications of

structural equation modeling to survey data, especially in cross-cultural and longitudinal research on which he has published many papers. Applications include human values, national identity, and attitudes toward immigrants and other minorities.

Don A. Dillman is Regents Professor of Sociology and Deputy Director for Research in the Social and Economic Sciences Research Center at Washington State University in Pullman, Washington. His research emphasizes methods for improving response to sample surveys in ways that reduce coverage, measurement and nonresponse errors. He has authored more than 250 publications including the 4th edition of his book, *Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method* (Wiley, 2014), coauthored with Jolene Smyth and Leah Christian.

Claire Durand is Professor of Survey Methods and Quantitative Analysis, Department of Sociology, Université de Montréal. Her main research interests pertain to the quality of electoral polls, the historical analysis of survey data and the statistics related to the situation of aboriginal people. She is currently vice-president/ president elect of WAPOR. She is author of numerous articles, book chapters and blog posts on the performance of electoral polls in various elections and referendums.

Stephanie Eckman is a Senior Survey Research Methodologist at RTI International in Washington, DC. She has published on coverage errors in face-to-face, telephone and web surveys and on the role of respondents' motivation in survey data quality. She has taught sampling and survey methods around the world.

Michelle L. Edwards is Assistant Professor of Sociology, Sociology and Anthropology Department, Texas Christian University. Her main research interests lie in research methodology, environmental risk, and public perceptions of science. She has previously co-authored an article with Don A. Dillman and Jolene D. Smyth in *Public Opinion Quarterly* on the effects of survey sponsorship on mixed-mode survey response.

Michèle Ernst Stähli is Head of group International Surveys at FORS (Swiss Centre of Expertise in the Social Sciences), running in Switzerland the European Social Survey (ESS), the International Social Survey Programme (ISSP), the European Values Study (EVS) and the Survey of Health, Ageing and Retirement in Europe (SHARE). Holding a PhD in sociology of work, since 2010 she has focused her research on topics related to survey methodology such as translation, nonresponse and mixed mode.

Jessica Fortin-Rittberger is Professor of Comparative Politics at the University of Salzburg and a former member of the Secretariat of the Comparative Study of Electoral Systems (CSES). Her main areas of research interest include political institutions and their measurement, with particular focus on electoral rules. Her work has appeared in *Comparative Political Studies*, *European Journal of Political Research*, *European Union Politics*, and *West European Politics*.

Kathleen A. Frankovic retired in 2009 as CBS News Director of Surveys and Producer, where she managed the CBS News Polls and (after 2000) CBS News election night projections. Since then, she has consulted with CBS News, YouGov, Harvard University and the Open Society Foundations, among others. A former AAPOR and WAPOR President, Frankovic has published many articles on the linkages between journalism and polling.

Siegfried Gabler is the leader of the statistics team at GESIS – Leibniz Institute for the Social Sciences and Privatdozent at University of Mannheim. He is a member of the Sampling Expert Panel of the European Social Survey. His research area covers sampling designs, especially for telephone surveys and for cross-cultural surveys, weighting for nonresponse, design effects, and decision theoretic justification of sampling strategies. He is involved in several projects in the context of telephone surveys and was jointly responsible for the Census 2011 project for Germany. He has published on a wide field of statistical topics.

Peter Granda is Associate Director of the Inter-university Consortium for Political and Social Research (ICPSR). Most recently he has participated in a number of collaborative projects with colleagues at the University of Michigan including acting as Director of Data Processing for the National Survey of Family Growth and as Co-Principal Investigator of the Integrated Fertility Survey Series. He has interests in the creation and use of comparative and harmonized data collections and has had a long association with the cultures of South Asia, where he spent several years of study in the southern part of the Indian subcontinent.

Sabine Häder is Senior Statistician at GESIS – Leibniz Institute for the Social Sciences, Mannheim. She is also a member of the Sampling Expert Panels of the European Social Survey. Sabine Häder holds a Doctorate in Economics. Current research areas are: sampling designs, especially for telephone surveys and for cross-cultural surveys. She has published widely on sampling topics.

Thomas Hinz is Professor of Sociology at the University of Konstanz. His research interests cover social inequality, labor market sociology, economic sociology, and survey research methodology, particularly survey experiments. Together with Katrin Auspurg, he authored *Factorial Survey Experiments* (SAGE Series Quantitative Applications in the Social Sciences, Vol. 175, 2015).

Lynette Hoelter is an Assistant Research scientist and Director of Instructional Resources at ICPSR and a research Affiliate of the Population Studies Center at the University of Michigan. At ICPSR, she is involved in projects focusing on assisting social science faculty with using data in the classroom, including the Online Learning Center and TeachingWithData.org, and generally oversees efforts focused on undergraduate education. Lynette is also a Co-Principal Investigator of the Integrated Fertility Survey Series, an effort to create a dataset of harmonized variables drawn from national surveys of fertility spanning 1955–2002. Her research interests include the relationship between social change and marital quality, gender in families, and the study of family and relationship processes and dynamics more broadly. She has also taught in the Department of Sociology and the Survey Methodology Program at the University of Michigan.

David Howell is Associate Director of the Center for Political Studies at the University of Michigan, and Director of Studies for the Comparative Study of Electoral Systems (CSES). His interests include public opinion, cross-national research, survey methodology, and developing local research capacity in international settings.

Ben Jann is Professor of Sociology at the University of Bern. His research interests include social-science methodology, statistics, social stratification, and labor market sociology. Recent publications include the edited book *Improving Survey Methods: Lessons from Recent Research*

(Routledge 2015) and various methodological papers in journals such as *Sociological Methodology*, *Sociological Methods & Research*, the *Stata Journal*, the *Journal of Survey Statistics and Methodology*, or *Public Opinion Quarterly*.

Kathy Joe is Director, International Standards and Public Affairs at ESOMAR, the World Association of Social, Opinion and Market Research. She works with international experts in the development of strategies relating to data privacy legislation, and global professional standards including the ICC/ESOMAR International Code on Market and Social Research. Recent areas of activity also include guidelines on fast-changing areas such as social media research, online research as well as mobile research. Kathy has worked at various publications including *The Economist* and *Euromoney* and she is also co-editor of *Research World*.

Timothy P. Johnson is Director of the Survey Research Laboratory and Professor of Public Administration at the University of Illinois at Chicago. His main research interests include measurement error in survey statistics, cross-cultural survey methodology, and social epidemiology. He has edited one book (*Handbook of Health Survey Methods*), and co-edited two others (*Survey Methods in Multinational, Multiregional and Multicultural Contexts*, and *Hard-to-Survey Populations*).

Jennifer Kelley is a Research Area Specialist in Survey Methodology at the Survey Research Center, Institute for Social Research, University of Michigan. Her main research interests are measurement issues, specifically questionnaire design and interviewer effects. Her operational interests include surveys conducted in international settings, particularly those in developing or transitional countries.

François Laflamme is Chief of data collection research and innovation section at Statistics Canada. His main interests are related operational research on various aspects of survey operations to improve the way data collection is conducted and managed in order to lead to more cost-effective collection or data quality improvements. He is author of many paradata research and responsive design papers.

Pierre Lavallée is Assistant Director at the International Cooperation and Corporate Statistical Methods Division at Statistics Canada. His fields of interest are: indirect sampling, sampling methods for hard-to-reach populations, longitudinal surveys, business survey methods, and non-probabilistic sample designs. Pierre is the author of the book: *Le Sondage Indirect, ou la Méthode Généralisée du Partage des Poids* (Éditions Ellipses) in French and *Indirect Sampling* (Springer) in English. He also contributed in many monographs and papers on survey methods.

Edith de Leeuw is MOA-Professor of Survey Methodology, at the Department of Methodology and Statistics, Utrecht University. Her main research interests lie in online and mixed-mode surveys, new technology, total survey error, and surveying special populations, such as children. Edith has over 140 scholarly publications and is co-editor of three internationally renowned books on survey methodology: *The International Handbook of Survey Methodology*, *Advances in Telephone Methodology*, and *Survey Measurement and Process Quality*.

Geert Loosveldt is Professor at the Center for Sociological Research of the Catholic University of Leuven (KU Leuven) where he teaches Social Statistics and Survey Research Methodology.

His research focuses on evaluation of survey data quality with special interest in the evaluation of interviewer effects and the causes and impact of non-response error. He is a member of the core scientific team of the European Social Survey.

Lars E. Lyberg , PhD, is senior adviser at Inizio, Inc., a research company, and CEO at Lyberg Survey Quality Management, Inc. His research interests lie in comparative surveys, survey quality and general quality management. He has edited and co-edited a number of monographs covering various aspects of survey methodology and is the co-author of the book *Introduction to Survey Quality* (Wiley, 2003). He is the founder of the *Journal of Official Statistics* and served as its Chief Editor between 1985 and 2010.

Jared Lyle is Associate Archivist at the Inter-university Consortium for Political and Social Research (ICPSR), Institute for Social Research, University of Michigan. His main research interests are in data sharing and digital preservation. He is author or co-author of several publications related to managing and curating data.

Alina Matei is senior lecturer at the Institute of Statistics, University of Neuchâtel, Switzerland. Her main research interests and publications concern different features of survey sampling, like sample coordination, estimation in the presence of nonresponse, variance estimation, etc., as well as computational aspects of sample surveys.

Kristen Miller is the Director of the Collaborating Center for Question Design and Evaluation Research at the National Center for Health Statistics. Her writings have focused on question comparability, including question design and equivalence for lower SES respondents, and the improvement of evaluation methods for cross-cultural and cross-national testing studies. She is a co-editor of two survey methodology books: *Cognitive Interviewing Methodology* (2014) and *Question Evaluation Methods* (2011).

Zeina N. Mneimneh is an Assistant Research Scientist at the Survey Methodology Program, Survey Research Center, University of Michigan. Her main research interests include interview privacy, social desirability biases, and interviewer effects on sensitive attitudinal questions. Her main operations interests include monitoring surveys using paradata, reducing survey error in conflict-affected settings, and international survey capacity building. She has published more than 25 peer-reviewed articles and book chapters.

Beth-Ellen Pennell is the Director of International Survey Operations at the Institute for Social Research's Survey Research Center at the University of Michigan. Pennell also serves as the Director of the Data Collection Coordinating Centre for the World Mental Health Survey Initiative, a joint project of the World Health Organization, Harvard University and the University of Michigan. Her research interests focus on cross-cultural survey methods and their application in resource poor settings. Pennell is an elected member of the International Statistical Institute, led the development of the Cross-cultural Survey Guidelines (http://ccsg.isr.umich.edu) and was one of the co-editors of *Survey Methods in Multinational, Multiregional and Multicultural Contexts*, edited by J. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler, B-E. Pennell, and T.W. Smith.

Adam Phillips is a research consultant and Managing Director of Real Research. He has been Managing Director of AGB Nielsen, Euroquest and Mass-Observation and CEO of

Winona Research. He chairs ESOMAR's Legal Affairs Committee, and through his knowledge of the public affairs arena and experience in liaising with UK and EU regulatory bodies, has broad experience in compliance and self-regulation. He chaired ESOMAR's Professional Standards Committee for 15 years, worked with the Committee, to set up an international disciplinary process that binds ESOMAR members to uphold the ICC/ESOMAR International Code.

Amy Pienta is a senior researcher at ICPSR in the Institute for Social Research. She is a faculty affiliate in the Population Studies Center, the Michigan Center for Demography of Aging, and the Michigan Center for Urban and African American Aging Research at the University of Michigan. Her training is in sociology (PhD from SUNY Buffalo) and demography (NIA postdoctoral fellowship at the Population Research Institute at the Pennsylvania State University). Her research centers on secondary analysis of the Health and Retirement Study (and other key datasets such as SIPP and NLS) exploring how marriage and/or family relationships affect a range of later life outcomes including: retirement, chronic disease, disability, and mortality. Her current research seeks to understand the underpinnings of a culture of data sharing in order to incentivize and strengthen this ethos across a broad range of scientific disciplines. Dr Pienta directs the National Addiction and HIV Data Archive Program (NAHDAP), funded by NIDA, and the National Archive of Data on Arts and Culture (NADAC), funded by the National Endowment for the Arts.

Dr. Stephen Quinlan is Senior Researcher at the GESIS Leibniz Institute, Mannheim and Project Manager at the Comparative Study of Electoral Systems project. His research focuses on comparative electoral behavior, public opinion, and the impact of social media on politics. His research has been published in the journals *Electoral Studies* and *Irish Political Studies*.

Finn Raben is Director General of ESOMAR, the World Association of Social, Opinion and Market Research, and has spent most of his working career in market research. Prior to joining ESOMAR, he had worked at Millward Brown IMS in Dublin, AC Nielsen, TNS and at Synovate. He is an ex Officio Director of MRII, the online educational institute partnered with the University of Georgia (USA); he serves as an external examiner at the International School of Management in Avans University (Breda, NL) and has joined the advisory board for the Masters of Marketing Research programme at Southern Illinois University Edwardsville.

Melanie Revilla is a researcher at the Research and Expertise Centre for Survey Methodology at the University Pompeu Fabra, Barcelona, Spain. Her main research interests lie in survey methodology, quality of questions, questionnaire design, web and mobile web surveys. She is author of a series of papers about quality of questions in different modes of data collection, and for several years she has been teaching courses on survey design, measurement errors, etc.

Caroline Roberts is Assistant Professor in Survey Methodology in the Institute of Social Sciences at the University of Lausanne, Switzerland. Her research interests relate to the measurement and reduction of different types of survey error, particularly in mixed mode surveys. She teaches courses on survey research methods and questionnaire design for the MA in Public Opinion and Survey Methodology, and is a member of the Scientific Committee of the European Survey Research Association.

Willem Saris is Emeritus Professor of the University of Amsterdam and momentarily visiting professor at the Universitat Pompeu Fabra in Barcelona. His research interests have been Structural Equation models and its application in improvement of survey methods, especially the correction for measurement errors. In that context he developed together with others the program SQP that makes it possible to predict the quality of questions and the improvement of them. Besides that he has been involved with Irmtraud Gallhofer in the study of argumentation of politicians. In all three fields he has made many publications.

Peter Schmidt is Professor Emeritus of social science methodology at the University of Giessen. His research concentrates on foundations and applications of generalized latent variable models, especially structural equation models. Applications include cross-country, repeated cross-sections and panel data. The substantive topics deal with values, attitudes toward minorities, national identity and innovation. He is, together with Anthony Heath, Eva Green, Eldad Davidov, Robert Ford, and Alice Ramos, a member of the Question Design Team for the immigration module of the European Social Survey 2014.

Silke L. Schneider is senior researcher and consultant at GESIS – Leibniz Institute for the Social Sciences, Mannheim. Her research interests cover comparative social stratification research, attitudes towards migrants, and survey methodology, especially the (comparative) measurement of socio-demographic variables. She has served as an expert with respect to education measurement and the International Standard Classification of Education for several cross-national surveys (e.g. ESS, SHARE), international organizations (e.g. UNESCO, OECD) and individual research projects.

Rainer Schnell is the Director of the Centre for Comparative Surveys at City University London and holds the chair for Research Methodology in the Social Sciences at the University of Duisburg-Essen, Germany. His research focuses on nonsampling errors, applied sampling, census operations, and privacy preserving record linkage. Rainer Schnell founded the German Record Linkage Center and was the founding editor of the journal Survey Research Methods. He is the author of books on *Statistical Graphics* (1994), *Nonresponse* (1997), *Survey Methodology* (2012), and *Research Methodology* (10th ed. 2013).

Barry Schouten is Senior Methodologist at Statistics Netherlands. His research interests are nonresponse reduction, nonresponse adjustment, mixed-mode survey design and adaptive survey design. He has written several papers in these areas and was project coordinator for EU FP7 project RISQ (Representativity Indicators for Survey Quality).

Kuniaki Shishido is an Associate Professor of the Faculty of Business Administration, Osaka University of Commerce. His areas of specialty are social gerontology, social survey and quantitative analysis of survey data. He takes charge of designing questionnaires of the Japanese General Social Surveys (JGSS). He also participates in cross-cultural survey projects such as the East Asian Social Survey (EASS).

Jolene D. Smyth is an Associate Professor in the Department of Sociology and Director of the Bureau of Sociological Research at the University of Nebraska-Lincoln. Her research focuses on challenges with questionnaire design, visual design, and survey response/nonresponse. She is co-author of the book *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design*

Method (Wiley, 2014) and has published many journal articles focusing on issues of questionnaire design and nonresponse.

Martin Spiess is Professor of Psychological Methods and Statistics, Institute of Psychology, University of Hamburg. His main interests include survey and psychological research methodology, techniques to compensate for missing data, robust and semi-/non-parametric statistical methods as well as causal inference. From 1998 until 2008 he was responsible for compensating unequal selection and response probabilities as a researcher at the German Socio-Economic Panel study.

Stephanie Steinmetz is an Assistant Professor of Sociology at the University of Amsterdam and senior researcher at the Amsterdam Institute for Advanced Labour Studies (AIAS), Netherlands. Her main interests are quantitative research methods, web survey methodology, social stratification, and gender inequalities.

Ineke A. L. Stoop is senior survey methodologist at The Netherlands Institute for Social Research/SCP. She is also Deputy Director Methodological of the European Social Survey, and Chair of the European Statistical Advisory Committee. Her main research interests lie in nonresponse and cross-national surveys. She is author and co-author of several books and book chapters on these topics.

Victor Thiessen† was a Professor in the Department of Sociology and Social Anthropology at Dalhousie University, Halifax, Nova Scotia Canada. During his carrier he served as Chair of his department, Dean of the Faculty of Arts and Social Sciences, and Academic Director of the Atlantic Research Data Centre, which made Statistics Canada surveys available to academic researchers. His substantive work focused on various transitions in young people's lives. Victor loved to play with statistics and to teach others how to do the same, something he continued to do as Professor Emeritus. He passed away suddenly and unexpectedly at the age of 74 on the evening of February 6th, 2016, in the company of his wife Barbara and very close friends. He is survived by his wife, his sister, two daughters and their partners, and six grandchildren.

Yves Tillé is professor at the Institute of Statistics, University of Neuchâtel. His main research interests are the theory of sampling and estimation from finite population, more specifically sampling algorithms, resampling, estimation in presence of nonresponse, estimation of indices of inequality and poverty.

Bojan Todosijević is Senior Research Fellow at the Center for Political Studies and Public Opinion Research, Institute of Social Sciences, Belgrade. His research interests include political psychology, political attitudes and behavior, and quantitative research methods. He has been affiliated with the Comparative Study of Electoral Systems (CSES) for a decade, mostly dealing with the integration of micro- and macro-level cross-national data. His work has been published in *Political Psychology*, *International Political Science Review*, and *European Journal of Political Research*.

Vera Toepoel is an Assistant Professor at the Department of Methods and Statistics at Utrecht University, the Netherlands. Her research interests are on the entire survey process, with a particular focus on web and mobile surveys. She is the chairwoman of the Dutch and Belgian Platform for Survey research and author of the book *Doing Surveys Online*.

Mary Vardigan, now retired, was an Assistant Director and Archivist at the Inter-university Consortium for Political and Social Research (ICPSR), a large archive of social and behavioral science data headquartered at the University of Michigan. At ICPSR, Vardigan provided oversight for the areas of metadata, website development, membership and marketing, and user support. She also served as Executive Director of the Data Documentation Initiative (DDI), an effort to establish a structured metadata standard for the social and behavioral sciences and as Chair of the Data Seal of Approval initiative.

Vasja Vehovar is a Professor of Statistics, University of Ljubljana, Slovenia. His interests are in survey methodology, particularly web surveys. He co-authored the book *Web Survey Methodology* and is also responsible for the corresponding website (http://WebSM.org).

James Wagner is a Research Associate Professor in the University of Michigan's Survey Research Center. His research interests include survey nonresponse, responsive or adaptive survey design, and methods for assessing the risk of nonresponse bias. He has authored articles on these topics in journals such as *Public Opinion Quarterly, the Journal of Survey Statistics and Methodology, Survey Research Methods, and the Journal of Official Statistics*.

Herbert F. Weisberg is Professor Emeritus of Political Science at The Ohio State University, Columbus, Ohio. His main research interests include American politics, voting behavior, and political methodology. He is author of *The Total Survey Error Approach: A Guide to the New Science of Survey Research*.

Brady T. West is a Research Assistant Professor in the Survey Methodology Program, located within the Survey Research Center of the Institute for Social Research on the University of Michigan-Ann Arbor campus. His main research interests lie in interviewer effects, survey paradata, the analysis of complex sample survey data, and regression models for longitudinal and clustered data. He is the first author of the book *Linear Mixed Models: A Practical Guide using Statistical Software* (Second Edition; Chapman and Hall, 2014), and also a co-author of the book *Applied Survey Data Analysis* (Chapman and Hall, 2010).

Gordon B. Willis is Cognitive Psychologist and Survey Methodologist at the National Cancer Institute, National Institutes of Health, Bethesda, MD. His main research interests are questionnaire design, development, pretesting, and evaluation; especially in the cross-cultural area. He has written two books on the use of Cognitive Interviewing in questionnaire design.

Heike Wirth is senior researcher at the Leibniz Institute for the Social Sciences, GESIS, Mannheim, and also a member of the German Data Forum. She works in the areas of social stratification, sociology of the family, data confidentiality, and research methodology. She is author or co-author of several articles or chapters on the measurement of social class.

Diana Zavala-Rojas is a survey methodologist and a researcher at the Universitat Pompeu Fabra, Barcelona. Her main research interests lie in questionnaire design, survey translation, linguistic equivalence in multilingual surveys, structural equation modeling and measurement error. She is a member of the Core Scientific Team of the European Social Survey and the Synergies for Europe's Research Infrastructures in the Social Sciences project.

Preface

The story of this Handbook covers five continents and five years! In the summer 2011, during the ESRA conference in Lausanne, SAGE contacted one of us in order to develop the idea of a Handbook of survey methodology and a team of an American, a German, a Swiss quickly joined by a Taiwanese began to elaborate the concept for the volume. Taking advantage of scientific meetings in the USA, Croatia, and Australia the editors developed a detailed proposal for the Handbook which then was reviewed by colleagues in the field contacted by SAGE (thanks to them). On the basis of these reviews the table of content was finalized and approved. The contract for the volume between SAGE and us was signed when the four of us met in Santiago de Chile for the annual ISSP meeting in 2013.

This marked the kick-off of the second stage of producing this Handbook by reaching out to a group of internationally acknowledged experts and inviting them to contribute a chapter. We started out hoping to recruit scholars from across the world, but were only partially successful: the 73 authors contributing to this Handbook reside in Asia, Europe, and North America.

While the chapters were solicited, written, and reviewed, we used the opportunity of a meeting in summer 2014 in Yokohama to coordinate the content and make last adjustments. Again one year later we met at the annual ISSP meeting, this time in Cape Town, and later in the summer in Reykjavik in order to finalize the last chapters, do a last adjustment to the Table of Contents and organize the writing of the introduction. A final meeting of the editors took place in Zurich in January 2016 bringing us back to Switzerland where it all started in 2011. The story of the development of this Handbook signifies its international character and reflects the importance and value we put on cross-national and cross-cultural perspectives while at the same time striving for a fair and balanced synthesis of current knowledge. Hopefully this Handbook will stimulate more survey research and the population of survey scholars will grow to the critical mass in even more regions.

Putting together this Handbook would not have been feasible without the support of our close collaborators, colleagues, and families whom we thank for their encouragement and the freedom to pursue this work. We are also grateful for the continuous support and encouragement we have received from SAGE.

April 2016 CW, DJ, TWS, YF

PART I

Basic Principles

1

Survey Methodology: Challenges and Principles

Dominique Joye, Christof Wolf,
Tom W. Smith and Yang-chih Fu

INTRODUCTION

There are a lot of reasons to publish a new handbook of survey methodology. Above all, the field of survey methodology is changing quickly in the era of the Internet and globalization. Furthermore, survey methodology is becoming an academic discipline with its own university programs, conferences, and journals. However, survey methodology could also be seen as a bridge between disciplines, resting on the shared methodological preoccupations between specialists of very different fields. These are some of the challenges we are addressing here.

Discussing the actual practices in many contexts is an invitation to think in a global perspective, along two directions. On the one hand, surveys are realized today all around the world in very different settings, which we call the globalization of surveys. But on the other hand, the 'Total Survey Errors' paradigm considers the complete survey life cycle and the interrelation of the different elements involved in the data collection process. That means that it would not be wise to pay too much attention to a single element at the risk of losing sight of the complete picture. This is of course valid for survey designers but also for estimating the quality and potential for use of existing surveys. A global perspective also requires a comparative frame. We even argue that integrating a comparative perspective from the beginning can enlighten many different aspects of survey design, even in a single national context.

These points will be developed throughout this handbook beginning in this introduction, with the idea of simultaneously providing a 'state of the art' and a perspective on the upcoming challenges. One important point in this respect is not to consider surveying as a technique but to consider it an integrated part of the scientific landscape and socio-political context. But first, we should explicate what we mean by a 'survey'.

WHAT IS A SURVEY?

The *Oxford Dictionary of Statistical Terms* begins with a broad definition of surveys: 'An examination of an aggregate of units, usually human beings or economic or social institutions' (Dodge 2010 [2003]: 398). Although many authors, such as Ballou (2008), agree on the polysemy of the concept, a more precise definition is given by Groves et al. (2004): 'A survey is a systematic method for gathering information from (a sample of) entities for the purpose of constructing quantitative descriptors of the attributes of the large population of which the entities are members' (p. 2). In this sense, the French word for survey, 'enquête', the same term used for a criminal inquiry, denotes well this systematic quest for information.

Sometimes 'surveying' is defined as obtaining information through asking questions, in line with the German word for survey: 'Befragung'. Dalenius (1985) recalls that observations are to be done according to a measurement process (measurement method and a prescription on its use) (Biemer and Lyberg 2003, see also Dodge 2010 [2003]: 399). That means that surveys defined in this sense share a lot of commonalities with other forms of data collection.

The idea of a systematic method for gathering observation includes for example exhaustive censuses as well as the use of a sample. In fact, some specialists explicitly limit surveys to data collection exercises conducted on samples (de Leeuw et al. 2008: 2). This handbook includes many chapters (Chapters 21, 22, 23) on the question of sampling, and the sample survey will be the first target, even though we see no reason to exclude by definition censuses which share a lot methodologically with surveys and are of great importance in the history of the quantitative observation of society.

'Quantitative descriptors' implies not only 'numbers' but also their interpretation, which in turn is placed in a broader interpretative frame. There is a process of operationalization that progresses from theory to measurement (Chapters 9, 14 and 34). In this sense, 'descriptors', i.e., point estimates, can only be understood by taking into account the structure and functioning of a given society. In other words, how to build the measure of 'items' is also one of the main topics to address, and a full part of the handbook is dedicated to survey-based measurement (Chapters 14–20). One strength of this handbook is the attention it gives to measurement and survey quality.

The definition of survey nevertheless excludes a lot of approaches useful for social research that are outside the scope of this handbook (but see Bickman and Rog 2009). Generally, qualitative methods are not considered, as we focus on quantitative descriptors. However, in certain parts of the survey life cycle, qualitative methods are well established and important to consider, such as in pretesting (Chapter 24). Along the same lines, 'big data', e.g., administrative data or data from the Internet, are not considered here because they are not organized a priori as a 'systematic model for gathering information'. to rephrase Dalenius. Nevertheless, such data are becoming vital to understanding social life, and must be taken as complementary with surveys.[1] In the last part of the handbook we will take into account the growing integration of surveys into a set of different sources of information (predominately Chapter 42, but also Chapters 40 and 41 in some aspects).

There are multiple ways of collecting information through surveys, and some distinctions between them are useful. Although a complete typology is outside the scope of this introduction, Stoop and Harrison (2012), for example, classify surveys based on the interrogatives *who*, *what*, *by whom, how*, *when*, *where* and *why*. Without mimicking their excellent chapter, we can discuss some lines.

In the 'by whom' category, different types of actors that sponsor activities can be mentioned:

- The scientific community tries to develop theory and analytical models in order to explain behavior, attitudes or values as well as the distribution of health, wealth and goods in given societies.
- The public administration; quantitative information is needed for governance – it is no coincidence that the words 'statistic' and 'state' have the same root – and an important task of a state is to assess the number of inhabitants or households it contains.
- Commercial enterprises need knowledge about their clients and their clients' reactions to their products in order to be as efficient as possible in their markets.
- Mass media are a special actor in the survey field and were mentioned as a particular category already in the 1980s (Rossi et al. 1985). Sometimes, they use results of polls more as a spectacular result to gain an audience rather than as a piece of systematic knowledge about society. That is part of the debate about the concept of public opinion (Chapter 5).

This type of distinction is also of importance when thinking for example about ethical aspects of surveys (Chapter 7). Of course, the boundaries between these actors are not always clear and depend on the national context, at least for the relation between administrative organizations and academia, and this could be important for the definition of measurement tools (Chapters 5 and 20). Nevertheless, we can expect that these actors have different expectations of surveys, their quality and their precision. In fact, most of the examples used in this book are taken from the academic context, implying that we focus more on the link between theory and measurement (Chapters 9 and 14) than other indicators of quality used, for example, in official statistics.[2] This example also reminds us that no absolute criterion for quality exists independently of the goals. This is clearly stated in the definition of quality given by

> Biemer and Lyberg, 'Quality can be defined simply as 'fitness for use.' In the context of a survey, this translates to a requirement for survey data to be as accurate as necessary to achieve their intended purposes, be available at the time it is needed (timely) and be accessible to those for whom the survey was conducted' (2003: 9). This handbook has

many chapters assessing data quality and aspects that can jeopardize quality (Chapters 34 to 39), an important aspect that must also be taken into account in the design state of a survey (Chapter 16 for example). This is of prime importance when developing the total survey error frame (Chapter 3).

We can further distinguish between the different ways to acquire information, the 'how' mentioned by Stoop and Harrison (2012). The first distinction is between modes of data collection, even though the boundaries between them are blurring, and multi or mixed modes are more and more often utilized (Chapter 11). We will come back to this later in the introduction when considering the development of surveys during the last century.

Who (or what) are the units of analysis of the survey, is another question. In most of the chapters in this handbook it would be individuals or households, but this is clearly a choice: a fairly big proportion of the surveys conducted by statistical offices are on establishments, even if it is individuals who give the information. In other cases we can aggregate information at some meso or macro level, such as occupations or regions. We include one chapter examining how survey data could be augmented by macro indicators (Chapter 42). Along the same lines, complex structures of data, such as members of a network or connections or interactions among these network members, as a basic unit is left to the specialized literature, such as the *Sage Handbook of Social Network Analysis* (Scott and Carrington 2011).

Stoop and Harrison also mention the time dimension as an important classificatory factor, the 'when', mostly to distinguish longitudinal and cross-sectional approaches. We cover the discussion on cross-sectional vs longitudinal survey designs in Chapter 9, and further details can be found in dedicated publications (e.g., Lynn 2009). But 'when' may also refer to the historical context (Chapter 8), a topic we turn to next.

SURVEYS IN HISTORICAL PERSPECTIVE

Though the idea of probability originated many centuries ago, and the art of counting people through censuses has been known for several millennia (Hecht 1987 [1977]), the modern survey, organized on the basis of a random sample and statistical inference, is more or less only one century old (Seng 1951; see also Chapter 21). Of course, some precursory works can be mentioned, such as for studying poverty or social mobility,[3] as well as the so-called straw polls for prediction of political results. However, the use of a small random sample, for example to predict elections, instead of a large, but fallacious, non-random selection of cases, was proposed only in the interwar years (Chapter 8). More or less at the same time official statistics began using samples.

Even if the advent of modern surveying was relatively recent, it has experienced major changes during the last decades. It is interesting to quickly discuss these changes, as they have structured the way to realize surveys as well as the way to use them or even think about them. One aspect of this change can be seen in the predominant mode of data collection, which – at least for Western countries – shows the following sequence:

- Just after the Second World War, most surveys were conducted face to face or by mail if enough people were considered to be literate.
- One generation later, in the 1980s, the telephone survey was seen as a new and efficient technology, at least when the coverage was sufficient. In some countries it was obligatory to be listed in the telephone directory, which therefore was seen to constitute an excellent sampling frame.
- Another generation later, telephones seem more and more difficult to use, as mobile phones tend to replace landlines, and centralized directories are no longer available. With the spread of the Internet, the web survey is seen as THE new methodology to adopt, in particular when considering the price of interviewing.
- Nowadays, in a context of declining response rates (Chapter 27), in many cases the idea is that combining modes (Chapters 11 and 18) is the way to follow, either in order to contact the greatest possible part of the sample or to have the best cost/benefit ratio, such as by using adaptive sampling (Chapter 26). This also explains the choice in this handbook to discuss interviewer interactions and interviewer effects less in depth, as would have been the case in earlier works (e.g., Biemer et al. 1991), but to put more emphasis on quality in general.

All these points pose challenges for data collection, which the survey industry has had to overcome. This has mostly resulted in adapting the field techniques to the changing circumstances. In particular, they must have been more transparent and more systematic but also more flexible.

Another story of change can be told in terms of the growing complexity of survey designs (Chapter 16), linked – once again – to the development of technology:

- At the beginning of the period considered here, after the Second World War, most surveys were single cross-sectional surveys. Their analysis was promising for the disciplines involved, and many important books were based on this type of information. The practical work of analysis was complicated enough with a single survey, in particular considering the (lack of) availability and ease of use of the computers needed for the analysis: software like SPSS was only developed in the seventies, and terminals with video displays became available more or less at the same time.
- In the next period, the time dimension gained importance, but under different modalities:
  - Repeated cross-sectional surveys were put in place in many countries, like the General Social Survey (GSS) in the United States, the British Social Attitudes (BSA) in Great Britain and the Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) in Germany. This was also an important change because it was no more single researchers conducting scientific projects but a tool that had to serve an entire scientific community. In other words, that was the beginning of the implementation of infrastructures in this field.

- Panel surveys, with multiple waves of data collection for the same respondents, have become more frequent. The Panel Study of Income Dynamics (PSID), running in the USA since 1968, is the longest-running longitudinal household survey in the world.[4] Similar initiatives have been launched in other countries, like the German Socio-Economic Panel (G-SOEP), which has run since 1984; the British Household Panel Survey (BHPS), which began in 1991; or the Swiss Household Panel (SHP), running since 1999. Of course there were precursors to these big initiatives; for example, the NORC's College Graduates Study was begun in 1961. In another disciplinary field, we can also mention cohorts like the National Child Development Study in the UK, based on a 1958 cohort, and the National Longitudinal Survey of Youth in the US (beginning in 1979). Most of these studies are further complicated by being household surveys following not only individuals but entire households, which of course change over the years, meaning they involve a very complex data structure.

- The next step was to introduce the comparative dimension, in addition to the time axis. Here we also have to mention three situations:
  - Comparative repeated cross-sectional projects. The European and World Values Surveys (EVS and WVS respectively) and the International Social Survey Program (ISSP) were put in place in the eighties, with precisely the idea to have a tool that allowed putting countries in perspective while evaluating change. Additionally 'barometers' have evolved in Europe and outside of Europe such as the Latino Barometer, Asian Barometer,[5] Afro Barometer, Arab Barometer and Eurasia Barometer. Both the East Asian Social Survey (EASS) and European Social Survey (ESS) were also built from the beginning with the same idea to measure social change while keeping strict comparability and high quality.
  - Harmonization of national panel studies to allow comparability. This was the challenge of the Cross-National Equivalent File (CNEF),[6] for example, pulling together, among others, the British (BHPS), German (G-SOEP), Swiss (SHP) and US (PSID) panels.
  - Comparative panel surveys designed from the beginning in a comparative perspective. There are not many examples, but we can nevertheless mention the case of the Survey of Health, Ageing and Retirement in Europe (SHARE) and the European Statistics on Income and Living Conditions (EU-SILC).

- The tendency nowadays is also to combine different sources of data to exploit the growing availability of information, as well as to pursue a movement initiated some years ago by social scientists such as Stein Rokkan (Dogan and Rokkan 1969). Multilevel models are more and more often used in comparative projects, making use of data at the country level. Some other projects use other types of information at the contextual level, which could be not only geographic but linked, to mention some examples, to social networks or to occupation. Other examples include the ESS which tracked the main events arising during the fieldwork period or the CSES which integrates not only geographical but also institutional information. SHARE is integrating not only answers to questions but also biomarkers which will probably gain even more importance in the development of health-related surveys these next years. This is described in Chapter 42, among other projects.

- A further sign of growing complexity comes from the fieldwork which in recent years has given more attention to paradata giving supplementary information on respondents and the contexts in which they live. These data are a potential basis to identify and correct biases (Kreuter 2013) but also provide a means to improve fieldwork monitoring and adapt to the best design (Chapter 26). Control of the production process is an important aspect of survey quality (Chapter 25; see also Stoop et al. 2010).

- In recent years, another type of survey based on the Internet has gained visibility – on-line or access panels:[7]
  - These panels typically are opt-in surveys meaning that potential respondents sign-up for them. This allows for conducting very cheap surveys on a large number of respondents that can be in the range of 10,000 to 100,000 panelists. Being a self-selected group, coverage of the population usually is poor and though companies try to improve representativeness by weighting according a few socio-demographic variables external validity typically is a problem. This therefore calls into question the quality of such surveys.

- In reaction to such a model, academic-driven surveys began to use an offline random recruitment process, called probabilistic panels. This strategy tries to combine the advantages of Internet surveys, which are cheap and easy to set up, and those of true random sampling. One of the most famous is probably the LISS panel, which has run in the Netherlands since 2007 (http://www.lissdata.nl/lissdata/About_the_Panel), but similar experiments are running in France with the ELLIPS initiative (http://www.sciencespo.fr/dime-shs) and in Germany with the German Internet Panel (http://reforms.uni-mannheim.de/internet_panel/home/) and the GESIS Panel (http://www.gesis.org/unser-angebot/daten-erheben/gesis-panel/). The rise of these Internet surveys does not mean that some other aspects of survey design, such as sampling (Chapter 22) as well as attrition and selection of respondents, are totally solved.

Looking at the history of surveys, there is something a little bit paradoxical: the recent proliferation of surveys, mainly Internet surveys, without random sampling or a clear description of the inference possibilities, put forward data by emphasizing the number of respondents more than the quality of data. This in a sense harkens back to the situation of the thirties and the discussion about the Literary Digest poll, although there is perhaps a difference: we now have a better knowledge of non-probability sampling (Chapter 22) and conditions of use.[8]

In this history, we can mention a last important distinction, between data produced by design, when a survey is designed according to a specific goal, and data produced by a process like administrative data or even the 'big data' mentioned earlier. At the time of writing, it does not seem useful to us to claim the superiority of one type of data over the other in an historical, or prospective, perspective, but it seems to us more important to think about the articulation between the research question and the way to answer it, in function of the available sources of information, but also considering the limitations that each kind of data may have.

A DISCIPLINE WITHIN DISCIPLINES?

The first survey practitioners were first of all substantive researchers who were learning methodology by conducting surveys and accumulating experience. In this sense, in the middle of the last century, surveying was something like a craft or an art. For example, a famous book of Stanley Payne (1951) was precisely entitled *The Art of Asking Questions*. Survey methodology began to cumulate as a discipline later on; for example, the first handbook was published in the eighties (Rossi et al. 1985). In order to move from an 'art' or a 'craft' to a 'science', there was therefore a need for a unifying paradigm. Total survey error was a perfect candidate for this (Chapters 3 and 10).

Survey methodology tends more and more to be seen as a discipline of its own: it has its own journals, such as *Public Opinion Quarterly* (POQ), *Survey Research Methods* (SRM), and the *Journal of Official Statistics* (JOS); its own conferences and associations, including the American Association for Public Opinion Research (AAPOR), European Survey Research Association (ESRA) and World Association for Public Opinion Research (WAPOR); as well as handbooks, this one being one of them. These associations have contributed to establishment of standards for the discipline (Chapter 2). This is also the case for market-oriented associations (Chapter 7).

But survey research is also a bridge between disciplines. For example, medical cohorts and sociological panels use the same methodology and they could begin to speak to each other just because they share so much in terms of data collection and analysis.

It is nevertheless important to recall in this context that methodology is also deeply embedded into disciplines. In this context, it is probably vain to develop a methodology for methodology, independent of the substantive goal of the research. There is always the risk of development inside an ivory tower without taking into account the most

important social and scientific challenges. For this reason we argue that a comparative perspective in methodology is a way to reflect about the limits and conditions for the validity of survey results and, as such, a very important safeguard. In other words, methodology must be open to the preoccupations of the other disciplines and keep a broad perspective. Methodological excellence will also be better accepted by substantive researchers if it shares, at least in part, the same disciplinary vocabulary and preoccupations. Survey research is therefore positioned in an area of tension between methodology on the one hand and substantive research on the other. But the reverse is true also: from a substantive researcher's point of view, it is important to be aware of data quality and, more generally, of the question of how much methodology affects empirical results.

We should not only consider if survey methodology is a discipline by itself or a specific field of expertise within several disciplines but also to take into account that some disciplines are sources of support for solving problems specific to survey methodology:

- The first discipline to mention is statistics, which can sometimes be seen as part of mathematics but also has its own position in the scientific landscape. From our point of view, statistics is important not only for sampling or data analysis but also for proposing tools for measurement (Chapter 17) and estimation of quality (Chapter 34).
- Psychology and in particular cognitive psychology as well as social psychology can be seen as a key when looking at the modeling of the answer process and the interaction between interviewers and interviewees. These disciplines are also important when looking at models of answers, like in Chapter 15.
- Psychometry is important to consider in the discussion about measurement and measurement models (Chapter 17 or 34).
- Sociology and political science can be considered when trying to understand the differences in survey participation, for example, by social position. It could be inspired by theory, like the one of Bourdieu, but may also refer to the idea of social exchange models as posited by Dillman et. al (2014). Along the same lines, the study of housing and living conditions of respondents can benefit from the work of urban sociology in order to conceptualize lifestyle and social conditions (Smith 2011). But sociology and political science are also important when embedding the construction of indicators into social and political constraints (Chapters 5 and 20).
- Linguistics and translation studies are also important to consider for questions of formulation as well as tools when considering translations (Chapter 19).

Interdisciplinarity is important in this frame. It is one of the conditions needed to fruitfully integrate methods, statistics and disciplinary perspectives. However, for survey methodologists, it is also important to find a common paradigm in order to address the questions of survey research. That is precisely the goal of the Total Survey Error perspective already mentioned, but some remarks can be added here.

The discussion of survey quality, including how to develop reliable estimators and efficient tools, is probably as old as the history of surveys. A text of Deming (1944) is one of the oldest milestones along these lines; interestingly, it was published in the *American Sociological Review*. From this perspective, it is important to consider all the possible sources of errors, the way that they interact and can endanger, or not, the results as well as the conclusions that can be drawn from the data (Smith 2005). This is why Chapter 3 is dedicated specifically to considering the possible sources of errors, their consequences and their interrelations. This does not prevent us from dedicating a full part of this work to the different facets of data quality (Chapters 34 to 39).

This attention to data quality and its consequences is one of the characteristics of this handbook and is related to the other transversal theme: comparative design, to which the last chapter of each part is dedicated as well as a general chapter on challenges of comparative survey research (Chapter 4). This attention to comparison and comparability of

course has a lot of consequences for the way we consider survey methodology.

COMPARATIVE FRAME AND THE NEED FOR MORE RESEARCH AND EXCHANGE BETWEEN CONTINENTS

For many reasons, to begin with the size of the scientific community and the need of information from the administration of a big country, survey methodology developed quickly in the United States after the Second World War. Part of this knowledge was 'exported' to Europe with the creation of important firms such as the IFOP with Jean Stoezel in France (Meynaud and Duclos (2007 [1985]) and the Allensbach Institute with Elisabeth Noelle-Neumann, both of which exchanged before with George Gallup (Zetterberg 2008; but see also Chapter 43). This movement of exportation and dissemination was of course primarily the case for 'opinion studies' rather than for official statistics, in which the various countries invested more energy in autonomous development.[9]

The origin and development of survey methodology inside the United States has had many consequences. A lot of studies have been established in a US context, but their validity in other contexts has not always been tested. This is the case, for example, for topics like the use of incentives (Chapter 28), for which we can expect that level of income and lifestyle are determinant. Along the same lines, a lot of studies are based on meta-analysis of published results and have never taken into account the cultural origin of the studies on which they were based. These are of course strong arguments to also consider and promote studies conducted in different contexts.

By the way, as mentioned in Chapter 43, surveys have been developed on all the continents. For example, the Global Barometer Surveys mentioned above that were inspired by the Eurobarometer now include Latino Barometer, Asian Barometer, Afro Barometer, Arab Barometer and Eurasia Barometer (http://www.globalbarometer.net/). The ISSP and the world extension of the value surveys also cover six continents since many years. This represents not only the dissemination of a technique all over the globe, which is interesting, but also the possibility of very interesting scientific developments (Haller et al. 2009). For example, to what extent can we compare different systems (Chapter 12)? What is the importance and impact of translation (Chapter 19)? What is the link between the general conditions of a country and the way of conducting surveys, either in drawing a sample or choosing the most adequate mode (Chapter 23)? The comparative perspective is clearly a central point here. Comparison is considered for the planning, measurement, use and quality of the survey process and the resulting data.

Another point can be mentioned here. Even though survey methodology is first of all a discipline founded on an empirical basis, there are still a lot of practical elements that are unknown and need to be explored, especially from a comparative perspective. For example, what is the relation between interest in a topic or in the questionnaire on the one hand and quality of response on the other? If such a question seems rather simple, there are still difficulties in measuring 'interest' in an appropriate manner. Likewise, even if the words in a questionnaire are chosen very carefully and discussed between experts after extensive cognitive testing or pretesting, there is still room to discuss the choice of a particular wording. This kind of information has to be documented in depth and published in order to allow scientific validation. In this context, it is a little bit astonishing that experiments in survey methodology are less often archived and re-analyzed than substantive surveys, even though the survey methodologists are probably the people most trained to do secondary analysis in an appropriate way (Chapter 40; see also Mutz 2011).

Once again, the comparative dimension adds a level of difficulty here: what is true in a given context is not necessarily true in another one (Chapters 12 and 13). That means that a

lot of experiments and empirical analysis have to be multiplied in different countries before solutions can be adopted or adapted. As mentioned, the incentives presented in Chapter 28 have been until recently only discussed in a US context, without knowing how appropriate it would be to implement them in the same way in Africa, Asia, Europe or South America. Another case could be item validity, which may vary by behavior patterns or rules of social exchange deeply embedded in cultures. For example, the prevailing norm of 'saving face' during social interactions may lead Chinese respondents to give socially desirable responses during face-to-face interviews to an excessive extent.[10] In some comparative surveys, such as the ISSP, a significantly large number of East Asian respondents also choose the mid-point response category (e.g., neither agree nor disagree) on attitudinal items, consistently refraining from revealing definite opinions. It also remains unknown how such cultural differences could be taken into account for comparative survey studies. Similarly, linguistic properties of questions and wordings are far better known in an Anglo-Saxon context that in other languages. In other words, we still have to make progress in order to find functional equivalence between countries when designing surveys (Chapters 19 and 33).

But insisting on the difference of conditions between countries is also an invitation to examine the importance of differences between different social groups inside a country, in terms of shared validity and reliability, as well as functional equivalence. In other words, if we follow such an idea, every survey is comparative by nature! That means that considering the challenges posed by multicultural surveys also make us aware of the heterogeneity of conditions and of respondents in each national context.

USE AND USEFULNESS OF SURVEYS IN A CHANGING WORLD

We have already mentioned the usefulness of surveys, a minimal proof being the development of the discipline. Let us discuss some points about this in more detail:

- In any country the statistical office and other governmental agencies are an important source of information needed for governing and for making informed decisions. For example, the European Commission conducts the Eurobarometer in order to gain regular information on the attitudes of Europeans. More generally, the use of statistical data is part of a movement wherein decisions are based on information and facts. All of the social reporting movement and evaluation studies are based on this line of reasoning. 'Evidence-based policy' is similarly in line with the push for 'evidence-based medicine' and the recent developments around genomics could be a further incentive in this sense.
- A lot of scientific work has been developed on the basis of the Comparative Study of Electoral Systems (CSES), ESS, E(W)VS and ISSP, to mention four important international comparative projects. Combining these sources, there are probably more than one thousand comparative journal articles published each year, showing the integration of survey methodology in the scientific activity of social sciences today, especially in a context of comparison. This is not only a benefit in terms of knowledge of the social system of the countries concerned, but also considering that of sharing methodological excellence and innovations between researchers. That means that the use of these important surveys increases the level of knowledge and competences inside the scientific community of the participating countries. The encouragement of data infrastructures in Europe, including data production with projects like ESS or SHARE and through the creation of ERICs,[11] is probably one more sign of the vitality of surveys as a tool for knowledge production in the academic field, at least in a European context.
- The relation to the media is sometimes more ambiguous, also because of the question of the accuracy of electoral predictions, sometimes based on surveys lacking the necessary transparency (Chapters 2 and 5). On the other hand, the question of feedback to citizens and participants in surveys is clearly an important point and it is even sometimes seen as a crucial element of a democratic system (Henry 1996). This is also part of the idea of a 'survey climate' developed in Chapter 6, at least if we think that discussion about surveys

and feedback to the people participating are part of a democratic culture (Chapter 5).

- We have already mentioned the close relation between the words 'statistic' and 'state', which clearly puts on the agenda some concepts related to political science. Four aspects at least are of interest here.
  - Surveys are part of the democratic system, in which everyone must have the right to express his or her own opinion. In this sense, every act against freedom of expression is problematic, and surveys are clearly relevant in the context of establishing democracy (Chapter 5).
  - There is also another link between surveys and democracy, as in many cases surveys mimic democracy by following the model 'one (wo)man, one vote', which we find when each adult inhabitant of a country has the same probability of being invited to participate in a survey.
  - Along the same lines, potential respondents have some rights: the right not to answer and the right to receive feedback about the results of what has been done with the information given. More generally, for the respondents, taking part in surveys entails a cost that must be acknowledged by carefully considering asking only meaningful questions.
  - But surveys are also ways of forming opinions and are not a purely neutral tool of observation. As mentioned in Chapter 20, there is a performative effect in the definition and use of categories and subjects to be asked about. More generally, there is a stream of research that questions the relation between surveys and public opinion or even the pertinence of the latter as a scientific concept.[12]

In fact, presented in this way, survey methodology can be not only seen as an interdisciplinary field but even considered in terms of transdisciplinarity (Hirsch Hadorn et al. 2008), which means taking into account the social conditions in which interdisciplinarity operates. We hope to have demonstrated in this handbook the interest in discussing surveying from a methodological point of view as well as from a far more general perspective, including social and political challenges. The condition for this is the practical possibility of using the information contained in surveys in the most pertinent way. Above all, that means documentation of the data and their conditions of production (Chapter 29) as well as good practices of analysis (Chapters 30, 31 and 32).

ORGANIZATION OF THE HANDBOOK

The handbook begins by introducing basic principles in Part I. It also introduces readers to the two main organizing principles: the idea of total survey error and the comparative perspective.

Part II underlines that surveys are not just a technical tool described by a simple methodology but that they are anchored in societies and historical contexts. They are useful for many actors, such as the state, and also have an impact because they have developed in a historical context. That is one reason that allows us to speak about 'survey climate' (Chapter 6) and one more reason to take into account ethical issues (Chapter 7).

The remainder of the handbook follows a simple flow model of conducting a survey: planning a survey, deciding about measurement, choosing a sample method, thinking about specific features of data collection, preparing the data for use, and finally assessing and, where possible, improving data quality. The questions that are posed and that have to be answered in each of these steps are even more challenging if a survey is to be carried out as part of cross-cultural, comparative research. Therefore, this particular aspect is, as mentioned earlier, dealt with in specific chapters discussing the particular challenges of comparative survey research in the different phases of the survey process.

The next chapters are dedicated to planning a survey (Part III). In this part, the research question that drives the choice of a suitable survey design has to be made explicit. A specific survey mode or modes is also to be determined. It also covers a discussion of the

total survey error paradigm in practice as well as a discussion of surveying in multicultural contexts, with an emphasis on doing surveys in difficult situations.

The chapters in the remaining parts discuss more specific aspects of the survey process through the question of measurement on the one hand (Part IV) and of sampling on the other (Part V), which are the practical tools that surveys require. Finally, specific issues in the data collection phase are discussed (Part VI). From the total survey error perspective, all the different steps are important when it comes to assessing the final quality of the outcome. In other words, a survey is always a combination of interlinked steps and its finally achieved quality depends on the weakest one.

Data are of no value if they are not used. That is why a lot of people put a lot of effort into making data available for secondary research. This of course implies properly documenting the data, organizing access to them and respecting the characteristics of the samples through weighting as well as ensuring comparability. Preparing data for use is covered in Part VII.

Quality can be threatened by a number of factors. Detecting, and hopefully correcting for, the potential biases is central in this respect. Part VIII addresses this in terms of measurement questions, non-response and missing values, as well as comparability challenges.

Part IX is dedicated to further issues. As mentioned, they can be divided into three components: better use of resources, beginning with secondary analysis, putting together different sources, comprising the different ways to link data and framing all that in a process of globalization of science and surveying.

For us as editors an important contribution of this handbook is not only to give tools to solve problems but also to offer elements to frame surveys in a more general context, allowing methodology and scientific practices to be linked in the most fruitful ways. One of the most challenging developments in the near future will be to combine data from different sources including surveys. However, we firmly believe that the survey model based on a random sample of a population will continue to play an important role in the advancement of the social sciences and knowledge society in general.

NOTES

1 A report recently published by the AAPOR writes, 'The term Big Data is used for a variety of data as explained in the report, many of them characterized not just by their large volume, but also by their variety and velocity, the organic way in which they are created and the new types of processes needed to analyze them and make inference from them. The change in the nature of the new types of data, their availability, the way in which they are collected, and disseminated are fundamental'. And, as a recommendation: 'Surveys and Big Data are complementary data sources not competing data sources' (AAPOR report on big data, 12.2.2015, accessed 29.2.2016 from https://www.aapor.org/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf).

2 See for example *Quality Assurance Framework for the European Statistical System*, version 1.2, edited by the European Statistical System, http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646, accessed 28.11.2015.

3 For poverty see for example Bowley (1915; discussed in Kruskal and Mosteller 1980). For social mobility, some Scandinavian studies of the nineteenth century are mentioned by Merllié (1994). We can also think of the works of Galton and Pearson on the transmission of quality between generations, as reported for example by Desrosieres (2002). For general histories of survey research see Oberschall (1972) and Converse (2009).

4 Cf. https://psidonline.isr.umich.edu/, accessed 3.12.2015.

5 Asian Barometer (http://www.asianbarometer.org/), a partner in the Global Barometer network, is not to be confused with Asia Barometer, an independent regional comparative survey network jointly sponsored by governmental agencies and business firms in Japan (https://www.asiabarometer.org/).

6 See https://cnef.ehe.osu.edu/, accessed 3.12.2015.
7 See also the ISO norm, http://www.iso.org/iso/catalogue_detail?csnumber=43521, accessed 14.1.2016.
8 See the AAPOR report on the use of non-probability sampling, http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf, accessed 29.2.2016.
9 This question of different development between the academic world, the private survey organizations and the national statistical institutes is still of relevance and was one of the reasons for the launch of the Data without Boundaries project (DwB) in the context of the 7th Framework Program of the European Union.
10 For theoretical arguments about such social norms, see Hwang (1987).
11 For these institutions see for example https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric, accessed on 26.12.2015.
12 In the French tradition, such a critical perspective exists in the stream initiated by Bourdieu in 1972 in the famous paper 'L'opinion publique n'existe pas' (reproduced at http://www.hommemoderne.org/societe/socio/bourdieu/questions/opinionpub.html, accessed on 26.12.2015). More recently see also the work of Blondiaux (1998). In English we can mention Bishop (2004).

REFERENCES

Ballou J. (2008) 'Survey', in the *Encyclopedia of Research Methods, Vol. 2*, Paul Lavrakas (ed.), Beverly Hills: Sage, p. 860.

Bickman L. and Rog D.J. (2009) *Sage Handbook of Social Research Methods*, Beverly Hills: Sage.

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

Biemer P.P., Groves R.M., Lyberg L.E., Mathiowetz N.A. and Sudman S. (1991) *Measurement Errors in Surveys*, Hoboken: Wiley.

Bishop G.F. (2004) *The Illusion of Public Opinion: Fact and Artifact in American Public Opinion Polls*, United States of America: Rowman & Littlefield Publishers.

Blondiaux L. (1998) *La fabrique de l'opinion*, Paris: Seuil.

Converse J. (2009) *Survey Research in the United States: Roots and Emergence 1890–1960*, New Brunswick, NJ, Transaction publisher.

Dalenius T. (1985). *Elements of Survey Sampling.* Swedish Agency for Research Cooperation with Developing Countries. Stockholm, Sweden.

Deming, W.E. (1944) 'On Errors in Surveys', *American Sociological Review* 9(4), pp. 359–369.

Desrosieres A. (2002) *The Politics of Large Numbers: A History of Statistical Reasoning*, Cambridge: Harvard University Press.

Dillman D.A., Smyth J.D. and Leah M.C. (2014) *Internet, Phone, Mail and Mixed Mode Surveys: The Tailored Design Method*, Hoboken: Wiley.

Dodge Y. (2010 [2003]) *The Oxford Dictionary of Statistical Terms*, Oxford: Oxford University Press.

Dogan M. and Rokkan S. (eds) (1969) *Quantitative Ecological Analysis in the Social Sciences*, Cambridge: MIT Press.

Groves R.M., Fowler F.J., Couper M.P., Lepkowski J.M., Singer E. and Tourangeau R. (2004) *Survey Methodology*, Hoboken: Wiley.

Haller M., Jowell R. and Smith T.W. (2009) *The International Social Survey Programme 1984-2009: Charting the Globe*, London: Routledge.

Hecht J. (1987 [1977]) 'L'idée de dénombrement jusqu'à la révolution', *Pour une histoire de la statistique, Vol. 1*, pp. 21–81, Paris: Economica/INSEE.

Henry G. (1996) 'Does the Public Have a Role in Evaluation? Surveys and Democratic Discourse', in Marc T. Braverman and Jana Kay Slater (eds) *Advances in Survey Research, New direction for evaluation*, No. 70, pp. 3–15.

Hirsch Hadorn G., Hoffmann-Riem H., Biber-Klemm S., Grossenbacher-Mansuy W., Joye D., Pohl C., Wiesmann U. and Zemp E. (2008) *Handbook of Transdisciplinary Research*, New York: Springer.

Hwang Kwang-kuo (1987) 'Face and Favor: The Chinese Power Game', *American Journal of Sociology* 92(4), 944–974.

Kreuter F. (2013) *Improving Surveys by Paradata*, Hoboken: Wiley.

Kruskal J. and Mosteller F. (1980) 'The Representative Sampling IV: The History of the Concept in Statistics', *International Statistical Review* 48(2), 169–195.

Leeuw Edith D. de, Hox Joop J. and Dillman Don A. (2008) *International Handbook of Survey Methodology*, New York: Lawrence Erlbaum.

Lynn P. (2009) *Methodology of Longitudinal Surveys*, Hoboken: Wiley.

Merllié D. (1994) *Les enquêtes de mobilité sociale*, Paris: PUF.

Meynaud H.Y. and Duclos D. (2007 [1985]), *Les sondages d'opinion*, Paris: La Découverte.

Mutz D. (2011) *Population-Based Survey Experiments*, Princeton: Princeton University Press.

Oberschall A. (ed.) (1972) *The Establishment of Empirical Sociology*, New York: Harper & Row.

Payne S. (1951) *The Art of Asking Questions*, Princeton: Princeton University Press.

Rossi P.H., Wright J.D. and Anderson A.B. (eds) (1985) *Handbook of Survey Research*, Orlando: Academic Press.

Scott J.G. and Carrington P.J. (eds) (2011) *The Sage Handbook of Social Network Analysis*, Beverly Hills: Sage.

Seng Y.P. (1951) 'Historical Survey of the Development of Sampling Theories and Practice', *Journal of the Royal Statistical Society*, Series A (General) 114(2), 214–231.

Smith T.W. (2005) 'Total Survey Error,' in Kempf-Leonard, K. (ed.) *Encyclopedia of Social Measurement*, New York: Academic Press, pp. 857–862.

Smith T.W. (2011) 'The Report on the International Workshop on Using Multi-level Data from Sample Frames, Auxiliary Databases, Paradata, and Related Sources to Detect and Adjust for Nonresponse Bias in Surveys', *International Journal of Public Opinion Research*, 23, 389–402.

Stoop I. and Harrison E. (2012) 'Classification of Surveys', in Gideon L. (2012) *Handbook of Survey Methodology in the Social Sciences*, Heidelberg: Springer, pp. 7–21.

Stoop I., Billiet J., Koch A. and Fitzgerald R. (2010) *Improving Survey Response: Lessons Learned from the European Social Survey*, Hoboken: Wiley.

Zetterberg H.L. (2008) 'The Start of Modern Public Opinion Research', *Sage Handbook of Public Opinion Research*, London: Sage, pp. 104–112.

2

Survey Standards

Tom W. Smith

DIFFERENT TYPES OF STANDARDS

First, there are informal common or customary practices. For example, in the field of survey research (as well as in many other disciplines), the general norm is to accept probabilities of 0.05 or smaller as 'statistically significant' and thus scientifically creditable. As far as I know, this rule is not codified in any general, formal standards, but it is widely taught in university courses and applied by peer reviewers, editors, and others at journals, publishers, funding agencies, etc. (Cowles and Davis, 1982). Other examples are the use of null hypotheses and including literature reviews in research articles (Smith, 2005).

Second, standards are adopted by professional and trade associations.[1] These may apply only to members (often with agreement to follow the organizational code as a condition of membership) or may be deemed applicable to all those in a profession or industry regardless of associational membership. Enforcement is greater for members (who could be sanctioned or expelled for violating standards), but can also be applied to non-members (Abbott, 1988; Freidson, 1984, 1994; Wilensky, 1964).

Third, standards are developed by standards organizations. These organizations differ from particular professional and trade associations in that they do not represent a specific group and they are not designed to promote and represent individual professions or industries, but to establish standards across many fields. The main international example is the International Organization for Standardization (ISO) and the many national standards organizations affiliated with the ISO (e.g. in the US the American National Standards Institute or in Togo the Superior Council of Normalization). Standards organizations typically both promulgate rules and certify that organizations are compliant with those rules (Smith, 2005).

Fourth, standards are written into specific contracts to conduct surveys. Contracts of

course can stipulate any mutually agreeable, legal provisions. But in many cases they incorporate specific requirements based on the codes/standards of professional and trade associations and/or standard organizations.

Finally, there are legally-binding standards established by governments. These can be local, national, or international. They may be set directly by legislation and/or established by regulatory agencies. Examples are the restrictions that many countries impose on pre-election polls (Chang, 2012; Smith, 2004). Enforcement can be through civil sanctions or criminal prosecutions. Government agencies sometimes work together with private organizations (usually trade, professional, or standards groups) to formulate and even enforce rules. In addition, governments also set standards by establishing rules for data collected by their own agencies (e.g. the US Bureau of the Census) or by those working for the government (OMB, 1999, 2006; Smith, 2002a; Subcommittee, 2001).

TYPES OF CODES OF PROFESSIONAL AND TRADE ASSOCIATIONS

One key component of professionalization is the adoption of a code of standards which members promise to follow and which the professional association in turn enforces (Abbott, 1988; Freidson, 1984, 1994; Wilensky, 1964). Codes for survey-based research can have several different components.

First, there are ethical standards that stipulate certain general and specific moral rules. These would include such matters as honesty, avoiding conflicts of interest, and maintaining confidentiality (American Statistical Association, 1999; Crespi, 1998). Even when applied to a specific industry/profession like survey research, they usually reflect general principles applicable across many fields.

Second, there are disclosure standards that stipulate certain information, typically methodological, that must be shared with others about one's professional work (Guide of Standards for Marketing and Social Research, n.d.; Hollander, 1992; Kasprzyk and Kalton, 1998; Maynard and Timms-Ferrara, 2011; Smith, 2002a). A prime example of this approach is the Transparency Initiative launched by the American Association for Public Opinion Research (AAPOR) and endorsed by such other organizations as the World Association for Public Opinion Research (WAPOR) and the American Statistical Association (AmStat).

Third, there are technical and definitional standards. Essentially, these are detailed elaborations on what is meant by other standards. For example, AAPOR and WAPOR both require that the response rate of surveys be disclosed and both endorse *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (www.aapor.org/Standard_Definitions2.htm) as the way in which those and other outcome rates must be calculated and reported (see also Lynn et al., 2001; Kaase, 1999).

Fourth, there are procedural standards. These indicate specific steps or actions that need to be executed when a professional activity is carried out. For example, checking cases through monitoring centralized telephone calls or recontacts might be stipulated procedures for interview validation.

Finally, there are outcome or performance standards. These specify acceptable levels that are expected to be reached before work is considered as satisfactory. This includes such things as having dual-entry coding show a disagreement rate below a certain level (e.g. less than 2 in 1000) or obtaining a response rate above some minimum (e.g. 70%).

Codes can encompass few, many, or all of these types of standards. The different types are not independent of one another, but interact with each other in various, complex ways. For example, procedural standards would have to be consistent with ethical standards and disclosure and technical/definitional standards are closely inter-related.

Professional and Trade Associations

There are many professional and trade associations in the field of survey research. First, there are the core professional and trade associations of the profession and industry of survey research. These include two, major, international professional associations: ESOMAR (formerly the European Society for Opinion and Marketing Research) and WAPOR; regional associations like the European Survey Research Association, Asian Network for Public Opinion Research, and the Latin American Chapter of WAPOR; many national professional associations such as the AAPOR, the Social Research Associations (SRAs) in Wales, Scotland, and Ireland, and the British Market Research Association (BMRA); and national trade associations such as the Association of the Marketing and Social Research Industry (Canada) (AMSRI), Council of American Survey Research Organizations (CASRO), Council of Canadian Survey Research Organizations (CCSRO), and the National Council of Public Polls (USA) (NCPP).

Second, there are professional and trade associations in closely-related fields: market research, the social sciences, and statistics. Market research associations include ESOMAR, which bridges the fields of survey and market research, and such other groups as the Advertising Research Foundation (ARF), the Alliance of International Market Research Institutes (AIMRI), the American Marketing Association (AMA), the Association of Consumer Research (ACR), the Association of European Market Research Institutes (AEMRI), the European Federation of Associations of Market Research Organizations (EFAMRO), the Market Research Quality Standards Association (MRQSA), the Marketing Research Association (MRA), and more specialized groups within market research such as the Audit Bureau of Circulation (ABC) and the Media Ratings Council (MRC).

The social-science disciplines most closely tied to survey research are sociology, political science, and psychology and they are represented by such cross-national groups as the International Sociological Association (ISA), the International Political Science Association (IPSA), and International Association of Applied Psychology (IAAP).

The main international statistical groups are the International Association of Survey Statisticians (IASS) and the International Statistical Institute (ISI).

Finally, since survey research is often public and widely distributed to the mass media and also sometimes done by or in collaboration with the media, standards relating to the media and journalism are also relevant. First, the survey-research field reaches out to and promotes best practices by journalists in their use of surveys. The NCPP focuses on the media and both AAPOR (Zukin, 2012) and WAPOR (Smith, 2002b) have guides for journalists. Second, numerous media organizations have standards about reporting their own surveys and the surveys of others, such as the Canadian Broadcasting Corporation (n.d.) and Reuters (2009). Lastly, various media professional and trade associations have standards relating to surveys such as the Canadian Association of Journalists (2012) and the German Press Council (2006). The organizational and associational codes of the media usually only touch on a few general points about using surveys.

Existing Professional and Trade Codes

Most of the professional and trade associations discussed above have codes of standards that address survey research. But they are quite variable in what is and is not covered. A few examples will illustrate this.

First, for codes of disclosure a comparison was made of nine documents (codes and supporting documents) by five organizations (AAPOR, CASRO, ESOMAR, NCPP, and

WAPOR) (Smith, 2002b). All organizations agreed on the reporting of the following elements of a survey: who conducted, who sponsored, sample design, sample size, sampling error, mode of data collection, when collected/dates, question wording, question order, sample population, and response rate. Also, mentioned in most of these codes and related documents were weighting/imputing and indicating the purpose of the survey.

Second, standards on response rates were examined. The codes and official statements of 20 professional, trade, and academic organizations were examined (Smith, 2002a). Four have neither codes nor any relevant official statements. Another three organizations have only brief general statements about doing good, honest research. Yet another three have general pronouncements about being open about methods and sharing technical information with others, but no details on what should be documented. Then, there are 10 that have some requirement regarding nonresponse.

Of these referring to nonresponse in their codes and statements, all require that response rates (or some related outcome rate) be reported. Only a subset of the 10 mentioning nonresponse require anything beyond a reporting requirement. Six organizations provide at least some definition of response and/or related outcome rates, but only the AAPOR/WAPOR, CASRO, and ABC definitions are detailed.

Three organizations deal with the issues of nonresponse bias in their codes. The WAPOR code, right after requiring the reporting of the nonresponse rate, calls for information on the 'comparison of the size and characteristics of the actual and anticipated samples' and the ESOMAR and MRQSA codes require in client reports 'discussion of any possible bias due to non-response'. Three organizations mention nonresponse bias in official documents. AAPOR in its 'Best Practices', but not its code, urges that nonresponse bias be reported. AmStat addresses the matter in its 'What is a Survey?' series. The AMA in its publication, the *Journal of Market Research*, requires authors to 'not ignore the nonrespondents. They might have different characteristics than the respondents'.

Of the organizations that have an official journal, nine have definite standards about reporting and calculating response rates, two have some general pronouncements mentioning nonresponse bias or the response rate, one has a marginally relevant standard on data sharing, and two have no applicable statement.

In brief, only the professional, trade, and academic organizations at the core of survey research and in the sub-area of media-ratings research take up nonresponse in their codes, official statements, and organizational journals. General market research and statistical organizations do not explicitly deal with nonresponse issues in their codes and standards and only marginally address these in the guidelines of their official journals. Even among the organizations that do address the matter of nonresponse, the proclaimed standards are mostly minimal. Some, but not automatic, reporting is required by all of the core organizations. However, definitions are provided by only six of the 10. Other aspects, such as nonresponse bias and performance standards, are only lightly covered. Thus, even among those organizations that consider nonresponse, reporting standards are incomplete, technical standards are often lacking and/or regulated to less official status, and performance standards are nearly non-existent.

Finally, professional, trade, and academic organizations have advanced the cause of standards by their general promotion and dissemination of research methods at their conferences and official journals (e.g. AAPOR's *Public Opinion Quarterly*, ESRA's *Survey Research Methods*, WAPOR's *International Journal of Public Opinion Research*). As Hollander (1992: 83) has observed, 'the annual AAPOR conference was recognized early on, together with *POQ*, which is older still, as a means of advancing standards'.

STANDARDS ORGANIZATIONS

Recently, the ISO initiated a major effort to develop standards for survey research. In 2003, Technical Committee 225 (TC225) was established to develop standards for 'market, opinion, and social research'. ISO and its national members are bodies specializing in the development of standards per se and lack detailed knowledge of most specific fields and industries. As such, TC225 was composed of survey-research practitioners and relied on direction from technical advisory groups made up of survey researchers in the participating countries and from two international, survey-research associations which are liaison members (ESOMAR and WAPOR) to develop the relevant definitions and rules. ISO 20252 on Market, Opinion, and Social Research – Vocabulary and Service Requirements were issued in 2006 and then updated in 2012 (www.iso.org/iso/catalogue_detail.htm?csnumber=53439). In addition, in 2009 ISO 26363 on Access Panels in Market, Opinion, and Social Research was adopted.

The ISO standards are largely consistent with the existing codes of professional and trade associations. For example, its disclosure list of information to be included in research reports closely follows the existing minimum disclosure requirements of the major professional and trade associations.

But the ISO standards go beyond most existing codes in two main regards. First, they specify the mutual obligations that exist between clients and research service providers (i.e. data collectors or survey firms). This includes stipulating elements that need to be in agreements between them including such matters as confidentiality of research, documentation requirements, fieldworker training, sub-contracting/outsourcing, effectiveness of quality management system, project schedule, cooperation with client, developing questionnaires and discussion guides, managing sampling, data collection, analysis, monitoring data collection, handling research documents and materials, reporting research results, and maintaining research records.

Second, they have a number of procedural and performance standards. These include the following: (1) methods for doing translations and level of language competency for the translators; (2) type and hours of training for fieldworkers; (3) validation levels for verifying data collected by fieldworkers; (4) use of IDs by fieldworkers; (5) the notification that potential respondents must receive; (6) documenting the use of respondent incentives; (7) guarantees of respondent confidentiality; and (8) what records should be kept and for how long they should be retained.

INTERNATIONAL COLLABORATIONS

Most major cross-national collaborations have standards for their participating members (Lynn, 2001). These include the European Social Survey, the International Social Survey Programme (ISSP), the OECD Programme for International Student Assessment, the Survey of Health and Retirement in Europe, the World Health Survey (WHS), and the World Values Survey. A fuller listing of these programs appears in Chapter 43 in this volume. For example, the ISSP Working Principles (www.issp.org/page.php?pageId=170) contain various standards for data collection and documentation, including requirements about mode, sample design, the calculation of response rates, and methods disclosure. For the WHS rules, see Ustun et al. (2005).

OTHER ASSOCIATIONS AND ORGANIZATIONS

Standards are promoted and developed by other groups besides the national and international professional and trade associations. These include several conference/workshop

series such as the Household Nonresponse Workshop (since 1990), the International Field Directors and Technologies Conference (since 1993), the International Total Survey Error Workshop (since 2005), the International Workshop on Comparative Survey Design and Implementation (CSDI) (since 2002), and the loosely-associated series of survey methodology conferences starting with the International Conference on Telephone Survey Methodology in 1987 through the International Conference on Methods for Surveying and Enumerating Hard-to-Reach Populations in 2012. CSDI for example has issued the *Guidelines for Best Practice in Cross-Cultural Surveys* (http://ccsg.isr.umich.edu/).

Also, survey archives around the world have created standards for the documentation of survey research data (Maynard and Timms-Ferrara, 2011). The metadata initiatives in particular have increased the documentation required for deposited surveys and also made that information more accessible to users. Of particular importance is the Data Documentation Initiative (www.ddialliance.org/).

IMPLEMENTING AND ENFORCING CODES

If the proof of the pudding is in the tasting, then the proof of standards is in their enforcement. Codes matter only if they are followed and here the experience of survey research is mixed. Three examples illustrate the present situation and its limitations.

First, most codes indicate specific methodological information about survey methodology that must be reported. Numerous studies over the years in various countries and covering both television and newspapers have repeatedly found that the basic methodological components required by disclosure standards are often not reported (Bastien and Petry, 2009; Sonck and Loosveldt, 2008; Szwed, 2011; Weaver and Kim, 2002; Welch, 2002). For example, the share of news stories reporting sample size ranged from 21% in the USA in 2000, to 37% in the USA in 1996–98, 49% in Canadian newspapers in 2008, and 65% in Poland in 1991–2007. For question wording, reporting was even lower, ranging between 6% and 25% in studies.

Similarly, reporting in academic journals also falls short of the disclosure standards. Presser (1984) examined what methodological information was reported in articles in the top journals in economics, political science, social psychology, sociology, and survey research. He found that in articles using surveys reporting ranged as follows: (1) sampling method from 4% in economics to 63% in survey research; (2) question wording from 3% in economics to 55% in survey research; (3) mode of data collection from 18% in economics to 78% in social psychology; (4) response rate from 4% in economics to 63% in survey research; (5) year of survey from 20% in social psychology to 82% in political science; and (6) interviewer characteristics from 0% in economics to 60% in social psychology.

Likewise, when looking at response rates, Smith (2002a) also found that reporting levels were low in top academic journals – 34% in survey research articles, 29% in sociology, and 20% in political science. In follow-up work, Smith (2002c) found in 1998–2001 that response-rate reporting remained low in political science and sociology, but was improving in survey research. However, even in survey research in 2001 only 53% of articles reported a response rate and just 33% provided any definition (see also Hardmeier, 1999; Turner and Martin, 1984).

Perhaps even more telling are the shortfalls in the methodology reports released by the survey-research, data collectors themselves. A study in Canada (Bastien and Petry, 2009) during the 2008 election found that sample size was reported 100% of the time, question wording 97%, weighting factors 62%, and interview mode 55%. A US study of 2012 pre-election polls found reporting for sample

size 98%, question wording 73%, weighting 37%, and interview mode 86% (Charter et al., 2013).

Second, the professional associations are not well-equipped to handle specific instances of alleged code violation which are commonly called standards cases. For example, WAPOR has no mechanism for or tradition of handling standard cases. AAPOR does have procedures and does conduct such reviews, but it has found that formal standards cases involve considerable effort, take a long time to decide, and, under some outcomes (e.g. exoneration or private censure), do not result in educating the profession. AAPOR procedures are by necessity complex and legalistic in order to protect the rights of the accused. Also, since the handling of standards cases is done by volunteers who must find time to participate, this creates a burden and takes considerable time to adjudicate. AAPOR believes that standards in the field can better be advanced by methods other than formal standards cases, such as by task-force reports and the Transparency Initiative.

Finally, many professions in part enforce their codes through the certification of members. But this practice is rare in the field of survey research. Globally, neither WAPOR nor ESOMAR has certification, nor does AAPOR or CASRO in the United States. However, MRA started a Professional Researcher Certification Program in 2005 (see www.mra-net.org/certification/overview.cfm). Its certification includes adherence to MRA's Code of Marketing Research Standards. Also, as is true of ISO standards in general, ISO 20252 provides for the certification of survey-research organizations as compliant with its standards.

THE ROAD TO PROFESSIONALIZATION AND THE ROLE OF CODES

Wilensky (1964) proposes five sequential steps that occupations go through to obtain professionalization: (1) the emergence of the profession; (2) establishing training schools and ultimately university programs; (3) local and then national associations; (4) governmental licensing; and (5) formal codes of ethics. Survey research has only partly achieved the second, for although there are some excellent training programs and university programs, most practitioners are formally trained in other fields (statistics, marketing, psychology, sociology, etc.).[2] Survey research has resisted certification and governmental licensing, although recent support for the proscription of fraudulent practices disguised as surveys (e.g. push polls and sugging – selling under the guise of a survey) and the ISO standards have moved the field more in that direction. On the development of the survey-research field, see Converse (1987).

Studies of professionalization indicate that one of the 'necessary elements' of professionalization is the adoption of 'formal codes of ethics…rules to eliminate the unqualified and unscrupulous, rules to reduce internal competition, and rules to protect clients, and emphasize the service ideal …' (Wilensky, 1964: 145) and 'codes of ethics may be created both to display concern for the issue [good character] and to provide members with guides to proper performance at work' (Freidson, 1994: 174).

Survey research has begun to follow the path of professionalization, but has not completed the journey.

In the judgment of Donsbach (1997), survey research is 'semi-professional'. Among other things, it has been the failure of survey researchers 'to define, maintain, and reinforce standards in their area' (Donsbach, 1997: 23) that has deterred full professionalization. As Crespi (1998: 77) has noted, 'In accordance with precedents set by law and medicine, developing a code of standards has long been central to the professionalization of any occupation'. He also adds that 'One hallmark of professionals is that they can, and do, meet performance standards'. In Donsbach's analysis (1997: 26), the problem is that standards

have neither been sufficiently internalized nor adequately enforced:

> We have developed *codes of standards*, but we still miss a high degree of *internalization* in the process of work socialization. We also lack clear and powerful systems of sanctions against those who do not adhere to these standards. It is the professional organizations' task to implement these systems and to enforce the rules.

The limited adoption and enforcement of standards and the incomplete professionalization has several causes. First, the survey-research profession is divided between commercial and non-commercial sectors. Coordinating the quite different goals and needs of these sectors is challenging. There has often been disagreement between these sectors on standards and related matters (Smith, 2002a, 2002d). Moreover, trade associations typically only include for-profit firms and exclude survey-research institutes at universities, government agencies, and not-for-profit organizations. But various steps have been taken to bridge this divide. AAPOR, for example, has certain elected offices rotate between commercial and non-commercial members and more informally WAPOR and other associations try to balance committee appointments between the various sectors. Also, CASRO has opened membership to not-for-profits and universities.

Second, for quite different reasons both sectors have not vigorously pursued professionalization. The academics have been the most open to professionalization in general and standards in particular since most are already members of two types of well-organized professions (university teachers) and their particular discipline (e.g. statistician, psychologist, sociologist, etc.). But while this socialization has made them open to professionalization and standards, it has also hampered the professionalization of survey research since the academics already are usually twice-fold professionals and many have only a secondary interest in survey research as a field/profession.

The commercial practitioners have seen themselves more as businesspersons and less as professionals and many see standards as externally-imposed constraints (akin to government regulations) that intrude on their businesses. Of course it is not inevitable that businesses oppose standards and people in business fields necessarily resist professionalization. For example, the Society of Automobile Engineers was successful from early on in establishing industry-wide standards and recommended practices (Thompson, 1954). However, this has not transpired within the survey-research industry. Suggested reasons for the limited development of cooperation within the survey field include a high level of competition (Bradburn, 1992) and that fewer benefits from collaboration and coordination may exist.[3]

Third, survey research in general and public-opinion research in particular are information fields with strong formative roots in both journalism and politics (Converse, 1987). Some have seen any attempted regulation (especially by government, but even via self-regulation), as an infringement on their freedom of speech and as undemocratic. They lean more towards an unregulated, marketplace-of-ideas approach related to the freedom-of-the-press model.

In sum, the incomplete professionalization of survey research has hindered the development and enforcement of professional standards. Incomplete professionalization in turn occurs due to inter-sector and inter-disciplinary division in survey research and from the high value placed by practitioners on the ideal of independence and proposition that the marketplace can exercise sufficient discipline. Both economic and intellectual laissez-faireism undermine the adoption and enforcement of standards.

CONCLUSION

Standards codes exist for the key professional and trade associations in the field of survey

research and there is a high degree of agreement on many of their provisions. But largely because professionalization has been incomplete, actual practice has often lagged behind the standards and enforcement has been limited. However, this situation has begun to change in recent years. For example, AAPOR and WAPOR have both adopted *Standard Definitions* for the calculation and reporting of response and other outcome rates and the ISO has worked with professional and trade associations in the field of survey research to establish international standards. Thus, the future prospects are for the spread of and greater enforcement of standards and the continued professionalization of survey research.

NOTES

1 Trade or industry associations are those in which organizations rather than individuals belong. Professional and academic associations have individuals in a particular occupation or scholarly discipline as members. There are also hybrid associations that include both individuals and organizations as members.

2 University survey-research programs include the Survey Research and Methodology Program at the University of Nebraska and the Joint Program in Survey Methodology at the Universities of Maryland and Michigan and summer institutes such as the ICPSR Summer Program for Quantitative Methods of Social Research, the Essex Summer School in Social Science Data Analysis, and the GESIS Summer School in Survey Methodology.

3 The setting of a standard gauge for railroads is an example in which several industries benefited. Builders of railroad equipment needed to produce only one size of wheels and axles, shippers gained as transfer costs were reduced, and railroads won increased traffic as unnecessary costs were eliminated.

RECOMMENDED READINGS

Interested readers may begin by studying professional standards, e.g. the AAPOR *Standard Definitions* (see http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions2015_8theditionwithchanges_April2015_logo.pdf) or the ICC/ESOMAR *International Code on Market and Social Research* (see www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf). For comparative surveys the Cross-Cultural Survey Guidelines are of particular importance (see http://ccsg.isr.umich.edu/archive/index.html).

REFERENCES

Abbott, Andrew (1988). *The System of Professions: An Essay on the Division of Expert Labor*. Chicago: University of Chicago Press.

American Statistical Association (1999). ASA Issues Ethical Guidelines. *Amstat News*, 269 (November), 10–15.

Bastien, Frederick and Petry, Francois (2009). The Quality of Public Opinion Poll Reports during the 2008 Canadian Election. Paper presented to the Canadian Political Science Association, Ottawa, May, 2009.

Bradburn, Norman M. (1992). A Response to the Nonresponse Problem. *Public Opinion Quarterly*, 56(Fall): 391–7.

Canadian Association of Journalists (2012). *Canadian Association of Journalists Statement of Principles*, at www.rjionline.org/MAS-Codes-Canada-CAJ-Principles [accessed on 6 June 2016].

Canadian Broadcasting Corporation (n.d.). *Canada Code: Canadian Broadcasting Corporation Journalistic Standards and Practices*, at www.rjionline.org/MAS-Codes-Canada-CBC [accessed on 6 June 2016].

Chang, Robert (2012). *The Freedom to Publish Public Opinion Poll Results: A Worldwide Update of 2012*. World Association for Public Opinion Research.

Charter, Daniela et al. (2013). Transparency in the 2012 Pre-Election Polls. Paper presented to the American Association for Public Opinion Research, Boston.

Converse, Jean M. (1987). *Survey Research in the United States: Roots and Emergence, 1890*. Berkeley: University of California Press.

Cowles, Michael and Davis, Caroline (1982). On the Origins of the .05 Level of Statistical

Significance. *American Psychologist*, 37: 553–8.

Crespi, Irving, (1998). Ethical Considerations When Establishing Survey Standards. *International Journal of Public Opinion Research*, 10(Spring): 75–82.

Donsbach, Wolfgang (1997). Survey Research at the End of the Twentieth Century: Theses and Antitheses. *International Journal for Public Opinion Research*, 9: 17–28.

Freidson, Eliot (1984). The Changing Nature of Professional Control. *Annual Review of Sociology*, 10: 1–20.

Freidson, Eliot (1994). *Professionalism Reborn: Theory, Prophecy, and Policy*. Chicago: University of Chicago Press.

German Press Council (2006). *German Press Code*, at http://ethicnet.uta.fi/germany/german_press_code [accessed on 13 June 2016].

Guide of Standards for Marketing and Social Research (n.d.). L'Association de l'Industrie de la Recherche Marketing et Sociale [Canada].

Hardmeier, Sibylle (1999). Political Poll Reporting in Swiss Print Media: Analysis and Suggestions for Quality Improvement. *International Journal of Public Opinion Research*, 11(Fall): 257–74.

Hollander, Sidney (1992). Survey Standards. In Paul B. Sheatsley and Warren J. Mitofsky (eds), *Meeting Place: The History of the American Association for Public Opinion Research*. American Association for Public Opinion Research: pp 65–103.

International Organization for Standardization/Technical Committee 225 (2005). *Market, Opinion and Social Research Draft International Standard*. Madrid: AENOR.

Kaase, M. (1999). *Quality Criteria for Survey Research*. Berlin: Akademie Verlag.

Kasprzyk, D. and Kalton, G. (1998). Measuring and Reporting the Quality of Survey Data. *Proceedings of Statistics Canada Symposium 97, New Directions in Surveys and Censuses*. Ottawa: Statistics Canada.

Lynn, Peter (2001). *Developing Quality Standards for Cross-National Survey Research: Five Approaches*. ISER Working Paper, 2001–21.

Lynn, Peter J. et al. (2001). *Recommended Standard Final Outcome Categories and Standard Definitions of Response Rate for Social Surveys*. ISER Working Paper 2001–23. Essex University, Institute for Social and Economic Research.

Maynard, Marc and Timms-Ferrara, Lois (2011). Methodological Disclosure Issues and Opinion Data. *Journal of Economic and Social Measurement*, 36: 19–32.

Office of Management and Budget (OMB) (1999). *Implementing Guidance for OMB Review of Agency Information Collection*. Draft, June 2, 1999. Washington, DC: OMB.

Office of Management and Budget (OMB) (2006). *Standards and Guidelines for Statistical Surveys*. Washington, DC: OMB.

Presser, Stanley (1984). The Use of Survey Data in Basic Research in the Social Sciences. In Charles F. Turner and Elizabeth Martin (eds), *Surveying Subjective Phenomena*, Vol. 2. New York: Russell Sage Foundation, pp. 93–114.

Reuters (2009). *Handbook of Journalism*, at http://handbook.reuters.com/index.php?title=Main_Page [accessed on 13 June 2016].

Smith, Tom W. (2002a). Developing Nonresponse Standards. In Robert M. Groves, Donald A. Dillman, John L. Eltinge, and Roderick J.A. Little (eds), *Survey Nonresponse*. New York: John Wiley & Sons, pp. 27–40.

Smith, Tom W. (2002b). *A Media Guide to Survey Research*. WAPOR.

Smith, Tom W. (2002c). Reporting Survey Nonresponse in Academic Journals. *International Journal of Public Opinion Research*, 14(Winter): 469–74.

Smith, Tom W. (2002d). Professionalization and Survey-Research Standards. *WAPOR Newsletter*, 3rd quarter: 3–4.

Smith, Tom W. (2004). Freedom to Conduct Public Opinion Polls Around the World. *International Journal of Public Opinion Research*, 16(Summer): 215–23.

Smith, Tom W. (2005). The ISO Standards for Market, Opinion, and Social Research: A Preview. Paper presented to the American Association for Public Opinion Research, Miami Beach.

Sonck, Nathalie and Loosveldt, Geert (2008). Making News Based on Public Opinion Polls: The Flemish Case. *European Journal of Communication*, 23: 490–500.

Subcommittee on Measuring and Reporting the Quality of Survey Data (2001).

Measuring and Reporting Sources of Error in Surveys. Statistical Working Paper No. 31. Washington, DC: OMB.

Szwed, Robert, (2011). Printmedia Poll Reporting in Poland: Poll as News in Polish Parliamentary Campaigns, 1991–2007. *Communist and Post-Communist Studies*, 44: 63–72.

Thompson, George V. (1954). Intercompany Technical Standardization in the Early American Automobile Industry. *Journal of Economic History*, 14(Winter): 1–20.

Turner, Charles F. and Martin, Elizabeth (1984). *Surveying Subjective Phenomena*. New York: Russell Sage Foundation.

Ustun, T. Bedirhan et al. (2005). Quality Assurance in Surveys: Standards, Guidelines, and Procedures. In *Household Sample Surveys in Developing and Transition Countries*, edited by the United Nations. New York: United Nations Publications, pp. 199–230.

Weaver, David and Kim, Sung Tae (2002). Quality in Public Opinion Poll Reports: Issue Salience, Knowledge, and Conformity to AAPOR/WAPOR Standards. *International Journal of Public Opinion Research*, 14 (Summer): 202–12.

Welch, Reed L. (2002). Polls, Polls, and More Polls: An Evaluation of How Public Opinion Polls are Reported in Newspapers. *Harvard International Journal of Press/Politics*, 7: 102–13.

Wilensky, Harold L. (1964). The Professionalization of Everyone. *American Journal of Sociology*, 70(September): 137–58.

Zukin, Cliff, (2012). *A Journalist's Guide to Survey Research and Election Polls*. AAPOR.

3

Total Survey Error: A Paradigm for Survey Methodology

Lars E. Lyberg and Herbert F. Weisberg

INTRODUCTION

Survey research began as a very practical enterprise, gathering facts and opinions from a large number of respondents, but with little underlying theory. As will be shown in this chapter, the 'Total Survey Error' approach (TSE) has been devised as an inclusive paradigm for survey research to guide the design of surveys, critiques of survey results, and instruction about the survey process. This chapter will review the TSE approach and place it within the broader concern for achieving a 'Total Survey Quality' (TSQ) product.

TSE emphasizes the trade-offs that are required in conducting surveys. Whenever a researcher conducts any type of study, there are constraints: costs, ethics, and time. There are also possible errors in any type of research study. In the survey field, those possible errors, among others, include the error that results from using a sample to represent a larger population ('sampling error'), the error from incomplete response to the survey and its questions ('nonresponse error'), and the error that occurs in survey responses ('measurement error').

The TSE approach emphasizes the trade-offs that must be made in trying to minimize those possible errors within the constraint structure of the available resources. Minimizing all of these error sources at once would require an unlimited budget as well as a very long time schedule, a situation that never prevails in the real world. Instead, the researcher has to decide which errors to minimize, realizing that expending more resources to minimize one type of error means fewer resources are available to minimize another type of error. Decreasing sampling error by greatly increasing the sample size, for example, would take away from the amount of money left to give interviewers extensive training, call back to locate respondents who were not found in the first attempt, pretest the questionnaire more extensively, and so on. The TSE approach suggests that clients commissioning surveys

should weigh these trade-offs, deciding how they want to spend their limited resources to minimize the potential survey errors that they consider most serious.

The goal of survey organizations is to achieve a quality product. Total Survey Quality (TSQ) includes the need for minimizing total survey error ('accuracy'), but it also includes, among other factors, producing results that fit the needs of the survey users ('relevance') and providing results that users will have confidence in ('credibility'). The well-run survey organization pays attention to these quality criteria while trying to minimize survey error within cost and other constraints.

The TSQ argument emphasizes the usability of the survey results. Maximizing TSE does not necessarily produce a useable set of findings. Consider a well-designed simple random sample of 100 likely voters that shows that one candidate has a 60%–40% advantage over the other candidate. While that seems like a small sample, such a result would be statistically significant: there would be less than a 5% chance of obtaining such a result if the other candidate were ahead instead. Yet, regardless of the statistical significance of its findings, such a small survey would not be considered credible. Few newspapers would take it seriously enough to publish an article based on that small survey, and campaign consultants would not find enough useful results to help them shape their campaign. Governments generally have strict quality standards they impose on surveys they contract for. The TSQ approach points to the need to take the likely usefulness of the survey results into consideration when designing and conducting the survey.

This chapter introduces the interrelated TSE and TSQ perspectives that underlie later chapters in this Handbook. We begin by relating the evolution of the TSE approach, and then describe how it is merged with the TSQ goals. We conclude by considering some recent developments that have the potential of upsetting usual practices in the survey research field.

THE DEVELOPMENT OF THE TOTAL SURVEY ERROR APPROACH

While surveys are a very common research procedure today, they were very rare a century ago with the exception of censuses. The basic notion that one could represent a larger population with a probability sample was not yet understood. The first major breakthrough involved statistical sampling theory, particularly Neyman's (1934) landmark article that provided a scientific basis for sampling. Neyman provided mathematical proof that sampling error could be measured by calculating the variance of the estimator. That was followed by Hansen's experiments for the US Census Bureau that showed that small random samples are more accurate than nonrandom judgment samples in which individuals are chosen to represent different groups in the population (Hansen and Hauser 1945). Thus, by the middle of the twentieth century it was well understood that the inevitable 'sampling error' that results from surveying a random sample of a larger population of interest could be estimated mathematically.

There also was an early realization that there are more possible errors in surveys than just sampling error (e.g., Deming 1944). Typical was Kish's (1965) description of survey error as having two components: sampling error and non-sampling error, with the latter including measurement bias. Sampling error was emphasized because it could be computed for probability samples with mathematical precision. Measurement bias could not be formally estimated, but there were many practical efforts in the 1950s through 1980s to decrease measurement bias in surveys by improving interviewing procedures as well by improving question wording.

There are, of course, potential errors in every social science research technique. The idea of systematizing and classifying potential errors began in Campbell and Stanley's (1963) careful classification of potential problems of 'internal validity' and 'external validity' in experimental research. There has

been less attention to detailing the full possible set of errors in content analysis and observation research.

Robert Groves's (1989) book on *Survey Errors and Survey Costs* provided a major theoretical development in the survey research field. It developed the concept of Total Survey Error (TSE), making it a paradigm for the survey field. Groves systematically listed the types of error in surveys and explained the cost calculation in minimizing each type of error. He showed that there are cost-benefit trade-offs involved, since minimizing one type of error within a fixed survey budget leads to less emphasis on controlling other types of error. Groves et al. (2004, 2009) updated Groves (1989) with research that had been done in the interim on the different stages of the survey process.

In addition to survey errors and costs, Weisberg (2005) emphasized another important consideration by explicitly adding survey effects to the trade-offs between survey errors and survey constraints. Rather than involving errors that can be minimized, these effects involve choices that must be made in survey design for which there are no correct decision. For example, asking question A before question B may affect the answers to question B, but asking question B before question A may affect the responses to question A. Thus, it may be impossible to remove question order effects in a survey regardless of how many resources are spent on them. As another example, a male respondent might give different answers on some questions to a male interviewer than he would to a female interviewer. Again, there is no easy way to remove gender-of-interviewer effects in surveys that use interviewers. Instead, survey researchers can seek to estimate the magnitude of such survey effects.

At this point, it is useful to introduce the different types of survey error that will be considered in more detail in subsequent sections of this Handbook. Figure 3.1 provides a depiction that emphasizes that sampling error is just the 'tip of the iceberg', one of the several possible sources of error in surveys.

The first set of errors arises from the respondent selection process. As already mentioned, *sampling error* is the error that occurs when interviewing just a sample of the population. If the sample is selected by probability sampling, then it is possible to compute mathematically the 'margin of error'

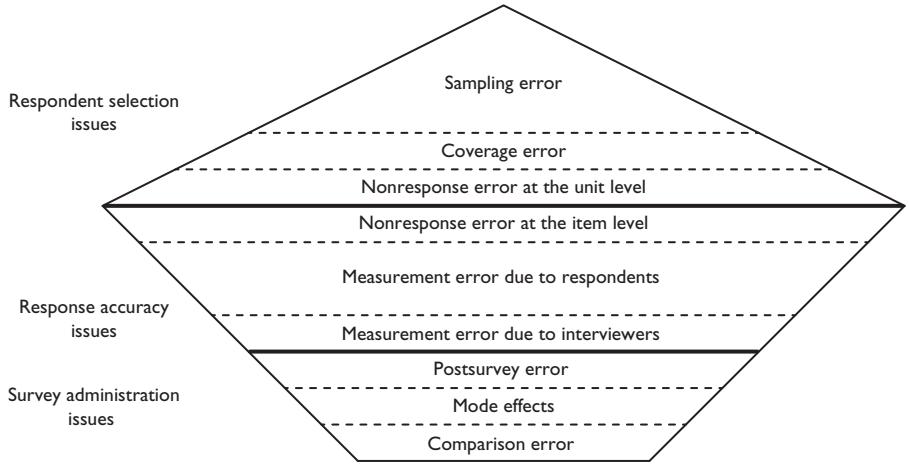

Figure 3.1 The different types of survey error source.

Source: Adapted from Weisberg (2005, p. 19).

corresponding to a 95% confidence interval for the survey results. Sampling issues can be very technical. While it sounds good to give each individual an equal chance of being selected by using simple random sampling, that procedure is often neither possible nor cost-effective. Stratifying the sample can reduce the sampling error, so that the right proportion of individuals are chosen within known subcategories of the population. Clustering the sample within known clusters (such as city blocks or banks of telephone numbers) can reduce costs, though that would increase the sampling error. A more serious problem is that probability sampling is often not feasible, as when conducting an Internet-based sample without having a list of the email addresses of the population of interest. Strictly speaking, sampling errors cannot be computed when the sampling is not probability-based, though many survey reports state what the sample would be for the number of individuals that were interviewed, had simple random sampling been used.

Respondent selection issues involve more than sampling error. 'Coverage error' or 'frame error' occurs when the list from which a sample is taken does not correspond to the population of interest. A telephone sample based exclusively on landline phones would entail coverage error since it would be biased against young people who only have cell phones. Frame errors also occur when a sample includes people who are ineligible, such as a voter survey that includes non-citizens. When a sampled unit does not participate in the survey, there is 'unit nonresponse'. Unit nonresponse occurs both when sampled respondents cannot be contacted and when they refuse to be interviewed. The response rate has fallen considerably in most surveys of the mass public, making it essential to consider this matter when designing a survey. Unit nonresponse becomes especially serious when it is correlated with variables of interest, such as if Republicans were less willing than Democrats to participate in US exit polls.

Another set of survey errors involves response accuracy issues. The usual emphasis is on the measurement error that arises when respondents do not give accurate responses. That may be due to the respondents themselves, as when they are not sufficiently motivated to provide accurate answers. Alternatively, it could be due to the question wording, including unclear questions, biased question wording, or demanding more detailed information than respondents can be expected to know or remember (such as asking people to recall what month they last saw their doctor). Measurement error due to respondents can also be related to questionnaire construction, such as question-order effects or fatigue effects from overly long questionnaires.

Interviewers can also introduce measurement error, which emphasizes the importance of interviewer training. One approach to minimize interviewer error is 'standardized interviewing', when interviewers are trained to ask the identical question in the identical nonjudgmental manner to all respondents. Some researchers instead prefer 'conversational' (or 'flexible') interviewing, with interviewers trained to help respondents understand the questions in the identical manner.

Additionally, the aggregate set of responses on a survey question can be inaccurate when some respondents do not answer all the questions, known as 'item nonresponse' or 'missing data'. This can be a matter of people skipping questions accidentally, intentionally refusing to answer questions, or not having an opinion on attitude questions. One can try to write survey questions in such a way as to minimize the likelihood of missing data, or one can try to deal with missing data problems at the data analysis stage. Missing data may not be a problem if it is truly missing at random, but the results would be biased if the occurrence of missing data were correlated with the variables of interest.

Finally, there is a set of possible errors related to survey administration. 'Postsurvey error' can occur in processing and analyzing

the data, including errors made in coding the survey responses. While converting survey responses to code categories may seem routine, it often involves difficult and subjective judgments that can introduce considerable error. There can also be 'mode effects' related to how the survey is conducted (e.g., telephone versus web surveys). Mode effects could be related to whether or not there is a human interviewer, as well as whether the respondent hears or reads the questions. Mode differences can be particularly important on so-called 'sensitive topics'. For example, people might be less willing to admit drug or alcohol use when an interviewer is asking the questions than when filling out an anonymous written questionnaire. 'Social desirability effects' can also be more common when there is a human interviewer, as when answering questions relating to racial prejudice. 'Comparison error' (Smith 2011) deals with the issue of non-equivalence of estimates on the same survey topic for different populations, which is relevant in cross-national and cross-cultural surveys. Comparison error also can occur when comparing answers on the same topic from surveys taken in different years if, as frequently is the case, the surveys word the questions differently.

Each of these different errors can either be random or systematic. Random errors are ones that vary from case to case but are expected to cancel out. For example, human interviewers are likely to skip a word occasionally when reading questions to respondents, but there should not be a pattern to such slips. Systematic errors are ones that bias the results, distorting the mean value on variables. For example, interest groups that sponsor surveys often word their questions so as to make it more likely that respondents will support those interest groups' positions, biasing the results to make it look like their position has more support than it would have with more neutral question wording.

Another issue is whether the errors are uncorrelated or correlated. Ideally, errors would be uncorrelated, as when an interviewer mistakenly records a respondent's 'yes' answer as a 'no'. What is more serious is when the errors for different respondents are correlated, which occurs when interviewers take multiple interviews and when cluster sampling is used. Having interviewers take multiple interviews (which is the only feasible way of taking interviews) and using cluster sampling help contain the cost of a survey, but correlated errors increase the variance of estimates due to an effective sample size that is smaller than the intended one and thereby make it more difficult to achieve statistically significant results. Correlated variances occur whenever multiple coders, editors, interviewers, supervisors and/or crew leaders are given assignments and affect their assignments in systematic but different ways. Thus, it is important to balance the cost savings from having to train just a small number of interviewers and other survey staff and from using cluster sampling versus the greater margin of error that results from those design decisions.

The survey design goal is to minimize the 'mean squared error' (MSE), which is the sum of (1) the variance components (including sampling error) and (2) the squared bias from measurement and other sources. MSE has become the metric for measuring Total Survey Error. While MSE cannot usually be calculated directly, it is useful conceptually to consider how large the different components can be and how much they add to the total survey error.

This discussion has spoken of 'potential' errors, since there is often no objective way to determine the 'truth' being measured in surveys. For example, ideally each person in a sample would answer the survey, but it is common to have some people refuse to cooperate. It may well be that the people who do not respond would have answered the same way as those who did respond, in which case their refusals did not create any error. However, their refusals create the possibility for error, since their answers might have been very different from those who responded to

the survey – and we generally have no way of knowing whether that is the case, so we must recognize the potential for error when some people in the sample are not interviewed.

In dealing with error that is not related to sampling, the survey design decisions are whether to ignore such error, whether to try to control it, or whether to try to measure it. While the ideal might be to minimize every type of error, that is impossible under fixed monetary and time constraints, so many researchers instead try to measure those error sources they cannot control. For example, some resources might be used to gain side information about individuals who were selected for the sample but would not participate, so one can estimate how much bias was introduced by their nonparticipation.

As the field has developed since the Groves (1989) book, there has been further research on each of the different error sources. That research will be presented in later sections of this Handbook, but it is important to stress a few of the most important developments and relevant controversies. One of the most important developments has been greater focus on the role of cognition in survey research, particularly as regards how respondents process survey questions. While early work viewed interviewing as a conversation, more recent theorizing has focused on how an interview is a cognitive task for the respondent. Of particular importance is the Cognitive Aspects of Survey Methodology (CASM) movement (Jabine et al. 1984), which began the process of using insights from the cognitive revolution in psychology to reduce measurement error in surveys.

The CASM movement led to the idea of 'think-aloud protocols', in which respondents tell the interviewer what they are thinking as they formulate their answers. That is particularly useful in testing question wording when developing a questionnaire. The emphasis on cognitive processes also led to Krosnick and Alwin's (1987) theory that there are two different levels of effort that respondents can exert in answering survey questions. While researchers hope that respondents will follow the 'high road' that requires real thinking, respondents will instead frequently use the 'low road' of giving an answer that sounds plausible so as to get through the task quickly. Such 'satisficing' behavior increases measurement error. Tourangeau et al. (2000) provided a further breakthrough with their separation of the high-road response process into four cognitive components: comprehending the question, retrieving relevant information from memory, judgment of the appropriate answer, and, finally, selecting and reporting the response. Improving survey question wording requires understanding possible errors in each of these four steps as well as trying to minimize the likelihood of satisficing.

Another important development in the TSE approach has been a greater focus on 'selection bias', which occurs when the actual respondents differ systematically from the intended population on the attitudes or behavior being measured. Non-probability samples introduce the possibility that the sample selection criterion is related to the attitudes or behavior of interest, thus biasing the survey results. The response rate in telephone surveys has fallen drastically over the years, leading to increased reliance on recruiting respondents who are willing to participate in web surveys. However, selection bias is a serious potential problem for web surveys because people who opt-in for a web survey might be very different from those who do not. Even weighting the sample on the basis of known population characteristics may not handle this problem because people who respond to such a web survey may differ from those with the same demographics who do not respond. Some pollsters try to resolve this problem by conducting a small supplementary telephone sample to use to weight the web sample. However, the basic argument of web survey proponents is that the response rate on telephone surveys today is so low that conventional random phone surveys also suffer from selection bias: the people who are willing to respond

to a telephone survey might be very different from those who do not. In any case, selection bias has become a key concern for surveys of the mass public.

DESIGNING SURVEYS FROM A TSE PERSPECTIVE

From a TSE perspective, the goal in survey planning is to have the smallest possible Mean Squared Error within the constraints of a fixed monetary and time budget. Stating that is, however, easier than achieving it. For one thing, it is difficult to estimate the magnitude of many of the possible errors, and, for another, costs of minimizing particular types of error are hard to estimate in advance. The task becomes even more difficult in the usual situation where several variables are being measured and they have different likely error structures. Working to reduce the measurement error on one set of survey questions could increase the error for a different set of questions in the same survey. Cross-national surveys pose even greater difficulty, especially when different survey organizations conduct the fieldwork in the different countries. This is not to discount the importance of thinking about MSE, but it points out that one cannot expect precision in estimating it at the survey planning stage.

The broader point is that there is not an overall formal survey planning theory. The TSE approach provides a framework for thinking about the several elements involved in planning a survey, but, as Lyberg (2012: 110) emphasized, there is no planning manual for surveys and 'no design formula is in sight'. Later in this chapter we review a recent attempt by Biemer et al. (2014) to provide more of an overall assessment, but we do not expect a formal survey planning theory to be developed in the near future.

The TSE perspective is certainly very useful as an outline for instruction about survey research. It is important that researchers contemplating a survey understand the trade-offs required, and the TSE approach helps make those trade-offs clear. Still, it would be hard in practice for an investigator to make the choices required, particularly because of the difficulty in stating trade-offs effects with precision.

The most common trade-offs situation is between sampling error and unit nonresponse. One way to reduce sampling error is to increase the sample size, but that increases interviewer costs considerably. Alternatively, one could expend more money on trying to obtain interviews with more people in the original sample, such as through more callbacks to people who could not be contacted originally, through conversion attempts to get interviews from people who refused on first contact, through offering alternative ways of answering the survey (such as web completion instead of a telephone interview) and/or through offering monetary inducements to respondents. In this day of big data, one might even be able to buy information about non-cooperating designated respondents, possibly including their consumer behavior, their house value, their frequency of voting in recent elections, and which political candidates they have contributed to – assuming that such inquiries about people without their consent can pass the ethical muster of an Institutional Review Board.

Another aspect of the TSE approach is to include in the survey some measurement of survey effects that cannot be minimized. For example, when there is not a perfect way to word a survey question, random half-samples of the respondents can be asked different versions of the same question to measure how robust answers on a topic are to how the question is worded. Similarly, when asking closed-ended questions, the order of response options can be varied for different random half-samples to measure how robust results are to the ordering of the response options. Such survey experiments increase the cost of programming a survey and require extra effort to ensure that the results are analyzed

correctly, but these extra costs are minimal compared to the considerable usefulness of the information they can provide.

It is similarly possible to design a survey so that some interviewer effects can be estimated. For example, it can be useful to keep track of the gender of interviewers in order to test for interactive effects between the gender of the respondent and the interviewer. Similarly, the race of interviewers can be included in the data so that race-of-interviewer effects can be analyzed. Though it is not always feasible, respondents would ideally be assigned to interviewers randomly (a procedure known as 'interpenetration'), so interviewer effects can be estimated.

The TSE approach has its roots in cautioning against the common practice to focus on just the sampling error. A focus just on sampling error results in underestimating the real error, sometimes considerably so. Ideally the service provider should, together with the client or the main users, identify error sources that are major contributors to the MSE, control them during the implementation stage, and potentially modify the survey design based on the analysis of paradata collected from relevant processes (Couper 1998, Groves and Heeringa 2006, Kreuter 2013).[1] In practice, however, errors and error structures are difficult to discuss with interested parties, since their complexity does not invite user scrutiny. Concepts such as correlated interviewer variance, design effects, and cognitive phenomena such as context effects and telescoping can be very difficult to discuss with a client or a user.

Instead, the average client thinks that good accuracy is the responsibility of the service provider, and the service providers are selected based on their perceived credibility. Thus, service providers or producers of statistics usually place high priority on being trustworthy and accurate regarding data quality, while users place high priority on aspects that they can easily assess. Examples of aspects or dimensions of quality that users appreciate include relevance (data satisfy user needs), timeliness (data are delivered on time), accessibility (access to data is user-friendly), interpretability (documentation is clear and comprehensible), and cost (data give good value for money). These dimensions are the components of so-called quality frameworks and there are a number of slightly different ones used by statistical organizations (Lyberg 2012). Groves et al. (2009) call them non-statistical dimensions but none-the-less they are important to consider at the design stage, since resources have to be set aside to satisfy user needs regarding these dimensions. Typically, users are not only interested in data quality (the total survey error as measured by the MSE) but also in some of these other dimensions. This means that we have a trade-off situation not only when it comes to the TSE components but also between TSE components and other dimensions. For instance, if a user wants data really fast, there is less time for nonresponse follow-up and accuracy might decrease compared to a situation where there is ample time for this activity.

The TSE framework is a typology of error sources with a prescription of how to control, measure, and evaluate their impact on survey estimates. It is a great conceptual foundation but very difficult to practice. In practice, survey organizations do not produce estimates of TSE on a regular basis because of costs, complexity, and/or lack of methodology. Also the number of error sources increases as new technology is introduced, and some error sources might even defy expression.

Many questions associated with the TSE framework have remained unanswered over the years. For instance, why are some error sources such as coding understudied when their consequences can be considerable depending on the use of coded data? One example is when movements on the labor market are studied and errors in repeated occupation coding result in an exaggerated picture of such movements. Other questions concern the allocation of resources. How should resources be allocated between measurement

of error sizes and actual improvement of the processes involved (Spencer 1985), and how should resources be allocated between prevention of errors, quality control, and evaluation, i.e., before, during, and after the survey? Also, most surveys are multipurpose, which is problematic from a design optimization point of view. Usually this problem is solved by identifying the most important variables and then working out a compromise design that best estimates those variables.

It seems very unrealistic to expect statisticians to develop expanded confidence intervals or margins of error that take all major error sources into account. A much more realistic scenario is to work on continuous improvement of various survey processes so that biases and unwanted variations are gradually reduced to the extent that estimates of variances become approximations of the mean squared error. One way of accomplishing that would be to vigorously apply proven design principles and survey standards together with ideas from the world of quality management, most notably the notion of continuous quality improvement. This calls for a new way of thinking, where TSE is extended to total survey quality (TSQ), where survey quality is more than a margin of error. In the next section we will describe how a gradual merging between TSE and TSQ has evolved.

THE MERGING OF TOTAL SURVEY ERROR WITH TOTAL SURVEY QUALITY

In the late 1980s and early 1990s many statistical organizations became interested in aspects of survey quality beyond traditional measures of accuracy. The quality frameworks represent one such development, where users were informed about several dimensions of survey quality. Just measuring and describing quality dimensions was, however, not sufficient. The quality management movement became part of the work in these organizations, with ideas about continuous quality improvement, two-way communication with users, handling competition from other providers of surveys, streamlining survey processes by observing metrics so that unnecessary variation is reduced, trying to eliminate waste, and minimizing costs of poor quality (Lyberg et al. 1998). At the core is the idea that measuring quality must be combined with systematic improvement activities.

Thus survey quality is more than a specified TSE. To be able to improve the TSE, we need to move to a state that we might call Total Survey Quality, TSQ, where the ingredients listed above are present. TSQ is illustrated by Table 3.1 that is adapted from Lyberg and Biemer (2008). The table shows that it is possible to view TSQ as a three-level

Table 3.1 Survey quality on three levels

Quality level	*Main stakeholders*	*Control instrument (examples)*	*Measures and indicators (examples)*
Product	Users, clients	Product requirements, evaluation studies, quality frameworks, minimum standards	Estimates of MSE components, degree of compliance to requirements, results of user surveys
Process	Survey designer	Metrics and paradata, control charts, verification and other quality control measures, process standards and checklists	Paradata analysis, analysis of common and special cause process variation, results of evaluation studies
Organization	Service provider, government, society	Excellence models, audits, self-assessments	Assessment scores, identification of strong and weak points, staff surveys

Source: Adapted from Lyberg and Biemer (2008)

concept: product quality, process quality, and organization quality.

The product quality is the extent to which agreed-upon product requirements have been met. Such requirements might include a certain response rate or that the translation of survey materials has been checked according to specifications.

A good product quality rests on good underlying processes. A good process is one that is stable in the sense that it always delivers what is expected. For instance, the process of interviewing implies that a number of elements must be in place for this to happen. Examples of such elements are a proper training program, a compensation system that encourages interviewers to strive for good quality, and supervision and feedback activities that help interviewers improve. Quality is built into the process by such quality assurance measures. Then quality control measures are used to check if these quality assurance elements are carried out and function as intended. This is done by using paradata or other metrics, i.e., data about the process. When paradata are plotted it is usually possible to distinguish between process variation that has common causes and variation due to special causes by using statistical process control theory and methods, especially the control chart (Montgomery 2005). There are also simpler ways of checking parts of the interview process. For instance, checking with respondents that interviews have actually taken place can discover interviewer falsification and so can simple response pattern analyses. Simple Pareto diagrams of monitoring outcomes can identify questions that were especially problematic for the interviewers.

Finally, a good process quality cannot be achieved without good organizational quality. For instance, a survey organization must have leadership that makes sure that the organization has staff with the right competence, that processes are continuously improved, that suggestions and opinions from users and staff are taken care of, and that good processes are promoted within the organization and evaluated regularly.

While the idea that survey quality should be measured by the mean squared error encompassing all variance and squared bias terms associated with an estimate seems reasonable, this becomes complicated in practice. It would be very expensive to estimate the sizes of various bias terms since that would entail comparisons between the regular survey result and the result of a preferred survey procedure (which for some reason, probably budget constraints, could not be used in the regular survey). It would also be very complicated and expensive to estimate the correlated response variance due to interviewers, coders, editors, and supervisors, since that would require interpenetration experiments comparing the outcomes of clustered assignments with random ones. Furthermore, models decomposing the MSE of an estimate generally do not include all major error sources. For example, the US Census Bureau survey model (Hansen et al. 1964) does not include nonresponse and noncoverage errors. It also takes time to conduct such studies. On the other hand it is quite disturbing that so many margins of error in surveys are understated. As mentioned one way out of this dilemma is to gradually improve survey processes so that they approach ideal ones associated with small errors. The quality management literature has given us philosophies, methods, and tools to do that (Breyfogle 2003).

The ASPIRE system (A System for Product Improvement, Review, and Evaluation) developed at Statistics Sweden (Biemer et al. 2014) is a recent attempt at handling survey quality assessment emphasizing TSE while using certain quality management tools. ASPIRE is a general system for evaluating TSE for the most important products at the agency. It uses six components for quality monitoring and process improvement, namely:

- MSE decomposed into sampling error, frame error, nonresponse error, measurement error, data processing error, and model/estimation error.
- Five quality criteria, namely the product's knowledge of risks of each of the MSE components on

the accuracy of the product, communication with data providers and users regarding these risks, compliance with best practices in the survey field regarding mitigation of errors, available expertise for monitoring and controlling errors, and the product's achievements toward risk mitigation and improvement plans.

- Quality rating guidelines for each of the quality criteria with descriptions of what the assessments poor, fair, good, very good, and excellent mean to ensure consistent ratings.
- Rating and scoring rules that help summarize progress in quality.
- Risk assessment including the intrinsic risk of doing nothing, the residual risk remaining after mitigation measures have been applied, and where the quality criteria are weighted by intrinsic risks, low, medium, and high. This allows for individual error source scores as well as a weighted overall product score.
- An evaluation process including pre-activities such as reviewing existing quality declarations and any program self-assessments, a quality interview with program staff, and post activities such as reviewing comments from product owners on product ratings possibly resulting in scoring adjustments.

ASPIRE is a comprehensive approach that is easily understood by management. The use of quality management principles and tools such as self-assessment, risk assessment, staff capacity building, and identification of areas most important to improve makes it possible to gradually mitigate the TSE. ASPIRE has so far been used during five rounds of quality assessment of the most important products at Statistics Sweden. Quality improved over the five rounds for most of these products.

ASPIRE does not really reflect all parts of the TSE, but it does a better job than other existing approaches toward mitigating the MSE. Admittedly the scoring process can be somewhat subjective and is, of course, highly dependent on the knowledge and skills of the evaluators. Furthermore, it is important that the evaluators are external, since most internal assessments have a tendency to underreport problems (Lyberg 2012).

It is possible to go beyond Total Survey Quality to consider the quality of the total research study. Total Research Quality (TRQ) would include not only the survey itself, but also the information needs that led to choosing to conduct a survey rather than a different research strategy. It would also include the analysis stage, checking whether the data analysis approach is appropriate for the research needs.

Accordingly, Kenett and Shmueli (2014) recently launched a new concept called Information Quality, InfoQ, which takes survey quality one step beyond Total Survey Quality. InfoQ attempts at assessing the utility of a particular data set for achieving a given analysis goal by employing statistical analysis or data mining. Obviously it is possible to increase InfoQ at the survey design stage by investigating the various known information needs. A formal definition of the concept is

$$\text{InfoQ}\ (f, X, g) = U[f(X|g)]$$

where g is the quality of goal definition, X is data quality, f is quality of analysis, and U is quality of the utility measure. InfoQ is in some sense a measure of the Total Research Quality (TRQ). Given a stated goal, InfoQ can be assessed at the design stage, at the data release stage, or before embarking on any secondary analyses. It can discover a faulty translation from statistics to domain, and it should be potentially useful when integrating data sets. InfoQ is clearly a few steps away from the one-size-fits-all frameworks we have mentioned above but it is still to be tested in practice.

Thus, in our review of survey quality we have not only moved from sampling error to total survey error, but we have moved beyond both to total survey quality and ultimately to total research quality.

THE FUTURE OF TOTAL SURVEY ERROR AND TOTAL SURVEY QUALITY

It is appropriate to conclude with some discussion of issues that are arising in contemporary

survey research that impinge on the TSE, TSQ, and even TRQ approaches. This involves recognition both of new data sources and of renewed debates as to how to conduct scientific surveys.

First, there has been an explosion of new data sources in recent years, principally due to the Internet. Much more 'administrative data' is collected nowadays and posted in a manner that is accessible to researchers. The term 'big data' is often used to include the large amount of commerce and government data that is becoming available to researchers. Add to this the data from social media, such as from Google searches or Facebook posts. This explosion of data sources is leading to a field of 'Data Science' that is apart from but not unrelated to survey research.

One result of this explosion is that the number of cases for data analysis is often exponentially greater than the number of respondents in most surveys. While sample size has traditionally varied between different types of surveys, non-governmental surveys were usually in the area of 800–2,000 respondents. Web surveys now can yield ten times those numbers of respondents, while 'big data' can involve thousands of times that number of cases.

On the one hand, the greater number of respondents makes it easier to detect small effects. A 1% difference is usually not statistically significant in a survey of 800–2,000 respondents, but it might be if there were 80,000–200,000 cases being analyzed. However, a 1% difference is still a small difference, often too small to matter substantively.

On the other hand, the greater number of respondents is rarely achieved through probability-based sampling. Strictly speaking, that makes statistical significance calculations inappropriate. Some researchers simply claim that conclusions based on a very large number of cases should be accepted, regardless of how the respondents were recruited. Other researchers use complex weighting procedures to weight the respondents on the basis of known population characteristics. 'Propensity weighting' has become common in web surveys as a means to compensate for people's differential willingness to participate in surveys, such as by weighting a large web survey by some aspects of a much smaller telephone survey.

A further complication is the debate between two statistical schools. On one side are the 'frequentists' who have provided the basis for how survey results have traditionally been interpreted. On the other side are the 'Bayesians' who weight survey results with their prior knowledge, often based on previous surveys. A good example of the Bayesian perspective is the US election forecasting models by Nate Silver and others who build models based on a state's voting in previous election and then update those models with new polls as they come out, giving more credence to polls from nonpartisan sources and human interviewers than to polls from survey organizations with ideological biases or to automated polls.

The election forecasting example is a good illustration of what is becoming known as data-driven journalism. Whereas journalists 50 years ago would develop forecasts for an election by talking to insiders who were considered specialists in the politics of a state, nowadays they use polls, polls of polls, and complex statistical models to develop their election forecasts. Such a development is occurring also in many decision-making areas, such as companies using a combination of focus groups and consumer surveys in deciding on which new products to bring to market and how to prepare effective marketing campaigns for those products.

Time and cost considerations are creating additional problems for traditional survey research. Many data users have a distinct need for improved timeliness, needing data on very short notice and not being willing to wait for a traditional survey to deliver that data. Furthermore, in-person national surveys have become very expensive in large industrial countries, and the cost of hiring interviewers has even made telephone surveys more expensive than many research budgets

permit. As a result, many users are not willing to pay for traditional surveys. There is still a lot of room for traditional surveys, but the survey community needs to adjust to these time and cost concerns.

The Total Survey Error approach and the Total Survey Quality emphasis have both been important developments. They have created better understanding of the survey process and more focus on how to improve surveys. Yet the movement away from traditional survey procedures to big data, opt-in panels, and other modes of data collection may signify another paradigm shift that yields even greater changes in how survey research is conducted. The development from TSE and TSQ to Total Research Quality seems very relevant since future data collection will be very different from today's. We can expect a plethora of data sources and a need to combine these. We need new theories and methods that can help us do that in a scientific way.

NOTE

1 Paradata in a survey are data about how the survey data were collected. For example, it would include how many attempts were made to contact a respondent, how long the interview took, and what time of day the interview was taken. Alternative definitions do not restrict paradata to the data collection process but rather all survey processes (Lyberg and Couper 2005).

RECOMMENDED READINGS

The following works are recommended for more information: Groves (1989), Groves et al. (2009), Lyberg (2012), Lyberg and Biemer (2008), Smith (2011), and Weisberg (2005).

REFERENCES

Biemer, P. P., Trewin, D., Bergdahl, H., and Japec, L. (2014). A System for Managing the Quality of Official Statistics. *Journal of Official Statistics*, 30(3), 381–442.

Breyfogle, F. W. (2003). *Implementing Six Sigma: Smarter Solutions using Statistical Methods*. Hoboken, NJ: John Wiley & Sons.

Campbell, D. T., and Stanley, J. (1963). *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand-McNally.

Couper, M. P. (1998). Measuring Survey Quality in a CASIC Environment. Paper presented at the Joint Statistical Meetings, American Statistical Association, Dallas, TX.

Deming, W. E. (1944). On Errors in Surveys. *American Sociological Review*, 9(4), 359–369.

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

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004, 2009). *Survey Methodology* (1st and 2nd edn). New York: Wiley.

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

Hansen, M. H., and Hauser, P. M. (1945). Area Sampling – Some Principles of Sample Design. *Public Opinion Quarterly*, 9(2), 183–193.

Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1964). The Estimation and Interpretation of Gross Differences and Simple Response Variance. In C. Rao (ed.), *Contributions to Statistics* (pp. 111–136). Oxford: Pergamon Press.

Jabine, T. B., Straf, M. L., Tanur, J. M., and Tourangeau, R. (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*. Washington DC: National Academy Press.

Kenett, R. S., and Shmueli, G. (2014). On Information Quality. *Journal of the Royal Statistical Society, A,* 177(1), 3–27.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

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

Krosnick, J. A., and Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2), 201–219.

Lyberg, L. E. (2012). Survey Quality. *Survey Methodology*, 38(2), 107–130.

Lyberg, L. E., and Biemer, P. P. (2008). Quality Assurance and Quality Control in Surveys. In E. de Leeuw, J. Hox, and D. Dillman (eds), *International Handbook of Survey Methodology*, Chapter 22 (pp. 421–441). New York: Lawrence Erlbaum Associates.

Lyberg, L. E., and Couper, M. P. (2005). The Use of Paradata in Survey Research. Invited paper, International Statistical Institute, Sydney, Australia.

Lyberg, L. E., Japec, L., and Biemer, P. P. (1998). Quality Improvement in Surveys – A Process Perspective. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 23–31.

Montgomery, D. C. (2005). *Introduction to Statistical Quality Control* (5th edn). Hoboken, NJ: John Wiley & Sons.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), 558–606.

Smith, T. W. (2011). Refining the Total Survey Error Perspective. *International Journal of Public Opinion Research*, 28(4), 464–484.

Spencer, B. D. (1985). Optimal Data Quality. *Journal of the American Statistical Association,* 80(391), 564–573.

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

Weisberg, H. F. (2005). *The Total Survey Error Approach*. Chicago: University of Chicago Press.

4

Challenges of Comparative Survey Research

Timothy P. Johnson and Michael Braun

INTRODUCTION

The origins of comparative survey research date back to the late 1940s (Smith, 2010). From those earliest experiences, the dangers of uncritically exporting social science research methodologies to new cultures and social environments were quickly recognized (c.f., Buchanan and Cantril, 1953; Duijker and Rokkan, 1954; Wallace and Woodward, 1948–1949; Wilson, 1958). Over the ensuing decades, the research literature began to demonstrate increasing awareness of both the opportunities and challenges of comparative survey research (Bulmer and Warwick, 1993; Casley and Lury, 1981; Cicourel, 1974; Frey, 1970; Przeworski and Teune, 1970; Tessler et al., 1987; van de Vijver and Leung, 1997; van Deth, (2013 [1998]). Today, the importance of comparative survey research is reflected in dramatic increases in the availability of internationally comparative survey data (Smith, 2010; see also Chapter 43 by Smith and Fu, in this Handbook), the volume of substantive analyses of these data, and the continued growth and sophistication of methodological research focused on the problems associated with comparative survey research (c.f., Davidov et al., 2011; Harkness et al., 2010a; van de Vijver et al., 2008). In this chapter, we provide an overview of the challenges posed by comparative survey research and some potential strategies for addressing them.

THE CHALLENGE OF COMPARABILITY

In comparative survey research, much more than the problems common to all monocultural surveys and measures need to be taken into consideration. In addition to depending on the quality of each individual national or cultural survey and measurement component, cross-cultural research is also dependent on their 'comparability'. Within the comparative framework, the same

challenges apply whether the goal is comparison of different ethnic groups within a single country or comparisons across multiple countries. In the first case, multiple issues, such as sampling, accessibility, translations, and interviewer effects, may be relevant. In the second case, larger contextual factors must also be considered. These latter contexts might include considerations such as the social structures and information available for the drawing of samples and conducting fieldwork, as well as socio-economic and political environments. Because cross-national comparative surveys are the more general case, we will focus in this chapter on comparisons of different countries. However, comparing different cultural groups within a country will be addressed as well.

An initial question which emerges early in the conduct of a comparative project is that of which countries are to be compared. The selection of countries or cultural groups is important both for the design of an original comparative survey project and for the analysis of secondary data. Depending on the research question, different strategies have been considered (e.g., Frey, 1970; Küchler, 1998; Przeworski and Teune, 1970; Scheuch, 1968). One approach involves the selection of countries which are contextually as different as possible. An important advantage of this 'most different systems' design (Przeworski and Teune, 1970: 34ff) is its ability to demonstrate the generalizability of relationships between the variables under study. If relationships between constructs are found to be consistent across countries that have little in common, then there is reason to believe that they are indeed generalizable and not dependent on the conjunction of very specific contextual conditions. If, on the other hand, the research aim is to demonstrate that a relationship originally found in one country or cultural group is unique and does not hold everywhere, then using different contexts (e.g., countries which differ in a large number of characteristics) is a disadvantage. This holds because, if the data seemingly are in accordance with the expectation of different relationships, alternative explanations might make an unambiguous interpretation impossible. In this case it is rather helpful to select highly similar countries, which is consistent with Przeworski and Teune's (1970) 'most similar systems' design. This is because, having once found differences in the relationships already with highly similar countries, it becomes more likely that this will also be the case across more different national contexts. In research practice, however, these considerations are increasingly losing relevance, at least as far as the selection of countries is concerned, due to the ongoing development of analytic strategies that are able to address many of these concerns (see below).

Many projects of international comparative survey research such as the International Social Survey Program (ISSP, http://www.issp.org) or the World Values Survey (WVS, http://www.worldvaluessurvey.org) strive for a more comprehensive coverage of all countries. And the survey projects restricted to specific regions of the world such as the European Social Survey (ESS, http://www.europeansocialsurvey.org), the European Values Study (EVS, http://www.europeanvaluesstudy.eu) or the Eurobarometer surveys (http://ec.europa.eu/public_opinion/index_en.htm) also pursue, in their domains, a maximally complete coverage of existing countries (for these projects, see Chapter 43 by Smith and Fu, this Handbook). Modern developments of regression analysis, in particular multilevel models, have provided adequate statistical procedures for the appropriate analysis of data from a large number of countries. These procedures allow one to investigate whether relationships between variables differ between countries and how such differences can be explained by characteristics on the level of countries.

The problem of comparability not only concerns the nations under investigation but also the selection of groups to be compared between countries. In comparative research, attention has to also be paid to the possibility

that what is found is not a difference between countries but a comparison between different groups. Scheuch (1968: 187) for instance, discusses whether farmers in the US are comparable with those in Europe: '… if one compares responses for both groups, much of what is done actually shows that similar labels refer to different groups, rather than demonstrating cross-cultural differences between the responses of otherwise comparable groups'.

PROBLEMS OF COMPARABILITY OF PROJECT COMPONENTS

In cross-cultural comparative research, the adequacy of conclusions depends on the quality and comparability of single national studies. In the presence of errors, similarities and differences between countries might simply be due to methodological artifacts. Indeed, a critical task of comparative survey research is to attempt to prevent, *ex ante*, and to detect and adjust for, *ex post*, these potential artifacts.

Two broad classes of errors can be distinguished: (1) The degree to which samples which were drawn and realized (after fieldwork) represent intended populations in a comparable way (sampling, coverage, and non-response error), and (2) the extent to which the questionnaire in general and specific items in particular after translation are processed the same way by respondents from different national and/or ethnic groups (measurement error). For a discussion of the total survey error framework, see Chapter 10 by Biemer in this Handbook.

Sampling, Coverage, and Non-response Errors

'Sampling error' results from analyzing just a sample instead of the entire population. It can be computed exactly, if the sample has been drawn at random and the other, systematic, error components (see below) can be neglected. 'Coverage errors' exist if not all units of the population under investigation have a chance to be included in the sample. The conditions of the sampling might be responsible for this. That is, the existence, access to and practicability of complete and up-to-date lists of the members of the population which are necessary for an unbiased sampling may not be available in all nations (see Chapter 23 by Gabler and Häder, this Handbook).

'Non-response errors' (see Chapter 27 by Stoop, this Handbook) refer to non-participation of those units of the population that have been selected as part of the sample. It is composed of non-contacts and refusals and those who are not able to take part in the survey for other reasons, such as limited accessibility, poor health, or language problems. Each of these components can differ significantly across countries, depending on the survey climate in a society (that is how Western-originated, quantitative surveys are perceived in general, and the population's general willingness to participate in them), as well as its survey literacy, or experience with survey research practices (Harkness, 2007), fieldwork duration, contact protocols, use of incentives, the physical mobility and accessibility of the population, the acceptability of refusal conversion practices, average household size and the resources of the research organization. Lack of awareness of or appreciation for important religious or public holidays when planning fieldwork can also have serious effects on non-response (Wuelker, 1993).

The consequences of excluding specific parts of the population – due to coverage and/or non-response errors – depend on the topics of the survey and the size of the groups affected but in particular on differences with respect to the variables of interest between these groups and those actually participating in the survey.

The utilization of identical procedures within each participating country is no

guarantee for comparability (Pennell et al., 2010). On the contrary, different strategies which are adapted to the respective contexts might be preferable. It makes obviously little sense to prescribe the conduct of telephone surveys for all countries, irrespective of the local conditions. Otherwise, in countries with a low telephone density a large part of the population would be excluded from the survey to begin with. Instead, in these countries, other survey modes should be employed that can be equally successful in covering and contacting the population. It might also be necessary to accept differences in sampling procedures and data collection mode, if the approaches optimal for each single country should be applied. In Germany, for instance, a sample drawn from municipal registers is regarded as the gold standard for personal interviews. In the US, in contrast, such registers do not exist. There, the standard approach for area probability surveys consists of drawing a random sample of city blocks first, then another sample of households, and finally the sampling of actual respondents using within household selection procedures. In some developing countries, even such a procedure is not feasible, for instance there is less reliable information regarding the distribution of the population in small geographical units, blocks of houses cannot be identified or a part of the population consists of nomads. In these cases less precise procedures must sometimes be used of necessity, although the now widespread availability of global positioning systems (GPS) and associated technologies have done much to alleviate these problems (c.f., Galway et al., 2012; Himelein et al., 2014). It would be hardly acceptable to use such sub-optimal procedures also in those countries in which good random samples could be drawn, only to preserve the formal equality of procedures.

Measurement Errors

Discrepancies between actual and reported values for measures targeted for assessment by survey data collection methods are known as 'measurement errors'. These include sources of error that may be contributed by survey instruments, respondents, interviewers, survey design and procedures, and differences in social systems (Biemer et al., 1991). In comparative research, each of these may have differential influence across nations and individual cultures.

Survey instruments may contribute to measurement error by way of complex designs, the formulation of single questions (and, of course, inadequate translation), the succession of questions and response alternatives. At the most basic level, the construction of individual items needs to be attended to carefully when developing measures for use in cross-cultural contexts. As an example, the President of the US has an entirely different function than the President in Germany (where he or she has a mostly representative role, only). When it comes to measuring attitudes towards the head of government, 'President' thus needs to be relabeled as 'Chancellor' in a German questionnaire. Harkness (2007) provides another example that considers the challenge of measuring religiosity in a comparable manner across nations using items that assess frequency of church attendance.

At the respondent level, asking for both objective and subjective information can be affected by cultural variability in question information processing, including problems of understanding and interpretation, difficulties of recall of the desired information from memory, and variability in judgment formation, response mapping, and response editing processes (Johnson et al., 1997).

In addition, it is known that social interactions and communication patterns are largely mediated by cultural norms, which may influence 'standardized' survey data in numerous ways. In comparative analyses, these differences may be misinterpreted as substantively meaningful respondent differences in attitudes, beliefs, and/or behaviors when they in fact represent variability in how respondents

react and respond to survey questions during social encounters. There are currently several conceptual frameworks for understanding cultural dimensions that may be useful for interpreting respondent behaviors during survey interviews. Some of these include those identified by Hofstede (2001), Inglehart and Oyserman (2004), Schwartz et al. (1992), Triandis (1996), and Trompenaars and Hampden-Turner (1998). Hofstede's (2001) individualistic vs collectivistic orientations, for example, have been linked to national and cultural group differences in propensity to give socially desirable, acquiescent, and extreme responses when answering survey questions (Johnson et al., 2011; Johnson and van de Vijver, 2003). In addition, unique cultural patterns such as the Asian and Middle Eastern Courtesy bias (Ibrahim, 1987; Jones, 1963; Mitchell, 1993), the 'ingratiation bias' (Bulmer, 1993), the 'sucker bias' (Keesing and Keesing, 1956), Simpatía and other dimensions of Hispanic culture (Davis et al., 2011; Triandis et al., 1984), and the 'honor' culture found in the southern United States (Nisbett and Cohen, 1996), may also differentially influence respondent behaviors.

Interviewers may be an additional source of variability in measurement error in comparative survey research. They may misread question text or register the answers of the respondents in a biased way, thereby contributing to error. They also might drift – consciously or unconsciously – from their role as a friendly but neutral observer, by cuing respondents to answer in a conforming or socially desirable manner. Recent research suggests, for example that in some countries, religious clothing may influence respondent answers (Benstead, in-press; Blaydes and Gillum, 2013; Koker, 2009). Countries differ in their survey traditions and, therefore, also in the standards of conduct, for instance the training of interviewers, and the acceptability of interactions between interviewers and respondents of different age or gender (Bulmer, 1993; Kearl, 1976; Newby et al., 1998; Pennell et al., 2010). The presence of others during survey interviews is also likely to have differential effects on the quality of self-reported information across countries (Mitchell, 1993; Mneimneh, 2012).

Survey design and protocols may also differentially add to measurement errors cross-nationally. Tendencies to give socially desirable answers, for example, are known to vary by survey mode in that they are most pronounced with personal interviews and least pronounced with self-administered surveys (Kreuter et al., 2008). As comparative projects of necessity must sometimes rely on varying modes of data collection across nations for various reasons, such as coverage issues (see above), efforts to minimize one potential source of cross-national error may unfortunately contribute to other sources of error. Differences in operational protocols, such as procedures for supervising and monitoring field interviewers, may have similar effects.

Structural differences across social systems, if not properly accounted for, may also lead to serious measurement errors. For example, potentially different definitions in different countries concern nearly all socio-demographic variables (Braun und Mohler, 2003; Hoffmeyer-Zlotnik and Wolf, 2003; Scheuch, 1968). Measures of education (Braun and Müller, 1997) are particularly problematic due to significant differences in the underlying education systems. Measuring income cross-nationally is equally challenging.

COMPARABILITY OF CONSTRUCTS AND ITEMS

Researchers addressing cross-cultural measurement error usually express concerns with the 'equivalence' of their measures and instruments. However, as Frey (1970: 240) has observed: 'equivalence is never total'. Accordingly, Mohler and Johnson (2010) conceive of equivalence as an 'ideal concept'

or 'ideal type' that is unattainable in practice, albeit a useful concept when considering the practical similarity of constructs, indicators, and measures across cultural groups. The notion of equivalence, however, is employed as a common heuristic in the literature concerned with cross-cultural methods and we a use it here for the same purpose.

It is of course essential to work to establish the comparability of the stimuli to which respondents are asked to respond. If there is no such comparability, the data will not only represent real differences between the countries or cultures under investigation but also measurement artifacts. Separating both can be a difficult challenge. Obviously, researchers want to avoid any lack of comparability through careful construction of measurement instruments *ex ante*. However, such efforts have not been successful in many studies, despite significant effort. This means that the users of secondary data must often face the task of demonstrating *ex post* whether or not equivalence can be demonstrated for the measures being analyzed.

Securing Equivalence of Measurement Instruments Ex Ante

The comparability of data can be compromised both by an inadequate rendering of measures in different languages and in the social reality of different national contexts. Researchers typically attempt to achieve functional equivalence, *ex ante*, through careful construction of the source questionnaire, making ample use of existing multicultural competence and experience (Harkness et al., 2010b), and via professional translation and adaptation of questionnaires (Harkness et al., 2010c).

To give an example, in the ISSP a so-called drafting group consisting of representatives of approximately six – ideally highly diverse – member countries prepares a preliminary source questionnaire in English. In an iterative procedure, this draft is then discussed with representatives of the other member countries, tested in a smaller number of countries, and finally approved by the ISSP General Assembly after discussion and voting. During this process, country-specific particularities are taken into account, both with regard to the inclusion of single items and their formulation. For instance, in the past East Asian countries have suggested a more abstract formulation of items on religion than would be necessary if all countries had a Christian background. The source questionnaire is then translated into the languages of the participating countries. For the translation, a team approach is recommended, in which several translators first translate the questionnaire independently from each other into the language of the country and then discuss their translations with experts for the specific topics dealt with in the questionnaire and survey experts (see Chapter 19 by Behr and Shishido, this Handbook). Special attention is given to the question whether the translated items are understood the same way as the items in the source questionnaire. Finally, the country-specific versions of the questionnaire are tested in a (cognitive) pretest, with special attention to problems of comprehensibility of the translation and comparability of the stimuli with the source questionnaire.

Often there is no clear distinction made between translation and adaptation, though this might be useful (see Chapter 19 by Behr and Shishido, this Handbook). Translation (in a narrow sense) refers to linguistic aspects. However, as a properly translated question might lead to different stimuli in different cultures, an adaptation is necessary which takes into consideration non-linguistic cultural information. Paradoxically, errors may be more frequent where cultural differences are seemingly smaller and more likely to be overlooked, such as when a common language is shared between countries that have somewhat unique cultures. An example are questions regarding whether children suffer if their mother works, where eastern German respondents have a tendency to think

of younger children and a higher amount of labor-force participation than do western German respondents. Thus, the answers of respondents are not directly comparable. Nevertheless, the revealed attitudes are still less traditional in eastern compared to western Germany.

Both qualitative procedures and quantitative pretests are helpful to achieve equivalence *ex ante*. Among qualitative procedures, cognitive interviews are particularly helpful to investigate problems in the response process (Beatty and Willis, 2007; Willis, 2015; see also Chapter 15 by Miller and Willis in this Handbook). *International* comparative cognitive studies, however, are infrequent compared to *intercultural* comparative cognitive studies in one country, due to the high coordination effort necessary (as exceptions: Fitzgerald et al., 2011; Miller et al., 2011). In quantitative pretests the main study questionnaire is tested under actual conditions. When pretests are conducted in different countries and high numbers of cases are available, statistical procedures for testing equivalence can also be employed.

It is, however, rare in practice that large quantitative pretests or qualitative studies are conducted in all participating countries. Even for the European Social Survey (ESS), qualitative studies were the exception and quantitative pretests have been conducted only in two countries, before the English-language source questionnaire was finalized. Moreover these pretests – as in the ISSP – were conducted mainly for the purpose of selecting substantively interesting and methodologically adequate questions and not in order to test the final measurement instrument to be used in the main study. Thus, it is often necessary to test equivalence *ex post* on the basis of the data collected in the main study.

Securing Equivalence of Measurement Instruments Ex Post

There are a large number of quantitative procedures for testing equivalence of measurement instruments *ex post* (Braun and Johnson, 2010; Davidov et al., 2011; van de Vijver and Leung, 1997). One of the most frequently used procedures is confirmatory factor analysis (Brown, 2006; Vandenberg and Lance, 2000). For cross-cultural comparisons, multiple-group confirmatory factor analysis tests whether a measurement instrument which has proved to be adequate for one country can also be applied to other countries. Comparability can exist on different levels which refer to the criteria for equivalence. Braun and Johnson (2010) illustrate the use of several complex and less complex procedures for the same substantive problem, demonstrating that they lead to essentially the same conclusions.

An important drawback of all quantitative procedures – perhaps with the exception of multilevel models (in particular multilevel confirmatory factor analysis) – is that they can only point to a problem but not account for the underlying causes. For these purposes – as already with establishing equivalence *ex ante* (that is in the course of pretesting) – probing techniques can be used. Because the conduct of cognitive interviews – particularly on the international level – is very time consuming, the conduct of additional web-based studies is a possible alternative. In such studies, probing questions can be included in a regular web questionnaire – comparable to Schuman's (1966) suggestion for 'random probes'.

Behr et al. (2012), for example, use both 'comprehension' ('What ideas do you associate with the phrase "civil disobedience"? Please give examples') and 'category-selection probes' ('Please explain why you selected [response alternative]') for an item on civil disobedience from ISSP. They document two reasons for the lower support for civil disobedience in countries such as the US or Canada, compared to countries such as Germany and Spain. On the one hand, in the first group of countries civil disobedience is associated with violence to a higher degree than in the second group. This is a

methodological artifact as the respondents are actually answering different questions. On the other hand, trust in politicians is (even) lower in the second group of countries than in the first one. This is a substantive result. However, in the data methodological artifacts and real differences are actually confounded, and it is difficult to distinguish between the two.

Braun et al. (2012) show that respondents from different countries, when asked to evaluate migrants, do think of largely comparable groups which correspond to the reality of migration in the different countries, though there are distortions. They use 'specific probes' ('Which type of immigrants were you thinking of when you answered the question?'). Conventional statistical methods alone could not have answered the question whether the attitudes of respondents in different countries towards migrants are incomparable because respondents think of different groups even in the case of comparable realities of migration.

ETHICAL AND OTHER CONCERNS IN COMPARATIVE SURVEYS

Conducting comparative survey research also requires consideration of potential variability and reconciliation of ethical standards of research conduct across the nations and/or groups being examined. In this regard, perhaps the most fundamental ethical question confronting comparative survey research is the appropriateness of imposing Western individualistic, quantitative social science methodologies onto cultures that do not share those traditions and values. Arguably, survey research is itself a culture-bound methodology designed to study social structures that are most specific to Western nations. Western assumptions, such as the equal value of all individual opinions, can be contested elsewhere (Rudolph and Rudolph, 1958) and suggests an inherent 'democratic bias' (Lonner and Berry, 1986) embedded within survey methodologies. From such a perspective, applications of survey research in other cultural environments may be interpreted by some as a form of 'scientific imperialism' (Drake, 1973).

One important source of conflict over the past several decades in this regard has centered around questions of the traditional Western requirement of obtaining informed consent from all individual survey respondents. In some traditional, collectivist societies, the concept of the self continues to be de-emphasized in favor of group identities. In these environments, researchers have commented on the importance of obtaining approval from village leaders or family elders in lieu of individual informed consent from potential respondents (Bulmer, 1998; Drake, 1973; Kearl, 1976). It has been argued that requiring individual informed consent where it is not understood or valued is a coercive form of 'ethical imperialism' that may undermine local institutions and social practices (Barry, 1988; Newton, 1990). In contrast, Ijsselmuiden (1992) cautions that it is not always clear who, other than the research participant, would be the most appropriate person to provide consent, and how to be certain that such a gatekeeper is in fact acting in the respondent's best interests. Common sense suggests that comparative researchers be aware of the importance of respecting and showing deference to local cultural traditions and values while also demonstrating respect for the importance of individual autonomy when making decisions regarding research participation.

A related issue is the level of information that is required to meet informed consent requirements in different nations. Concerns have been expressed that some national requirements for complex, multi-page consent documents may be confusing and not always appropriate when exported to other cultural contexts, leaving potential respondents with the impression that researchers are more concerned with their own

legal protections rather than the welfare of respondents (Marshall, 2001).

Some additional ethical concerns that are specific to comparative survey research include cross-national data sharing practices (Freed-Taylor, 1994), the ethics of collecting data in nations that may have 'no intellectual or material interests in the results', also known as 'academic colonialism' (Schooler et al., 1998), and abandonment of the strict standards of methodological rigor commonly seen in national surveys when conducting cross-national research (Jowell, 1998).

In addition, the importance of equal professional and power status for researchers from various participating nations, as opposed to exploitive 'hired hands' research (Clinard and Elder, 1965; Kleymeyer and Bertrand, 1993; Warwick, 1993), cannot be over-emphasized. As O'Barr et al. (1973: 13) have commented 'when decisions about research in one society are made by persons from another society or, worse yet, by foreign governments, the potential for abuse increases considerably, as does the anxiety of those being studied'.

Fortunately, accumulated insights and experience over the past several decades have both brought to light these various ethical concerns and explored approaches to addressing them using strategies that recognize and demonstrate respect for societal values and for personal well-being. In particular, both Warwick (1980) and Hantrais (2009) have presented sets of recommendations for engaging in sound ethical practice while also maintaining high scientific standards in the conduct of comparative survey research.

SUMMARY

The collection of survey data across nations or cultures presents many additional challenges beyond those confronted when conducting research in more homogeneous environments. We have sought to provide an overview of those challenges specifically associated with the design and execution of high quality comparative survey research, along with some of the potential solutions that are being actively pursued by investigators around the world.

RECOMMENDED READINGS

For further reading, we recommend the following recent texts concerned with unique conceptual and technical aspects of comparative survey research: Davidov et al. (2011); Hantrai (2009); Harkness et al. (2003, 2010a); Hoffmeyer-Zlotnik and Wolf (2003); and Porter and Gamoran (2002) In particular, Harkness et al. (2010a) provide a broad overview of cross-national and cross-cultural survey research methodologies.

In addition to these texts, Scheuch (1968) is one of the earliest papers to carefully describe the challenges associated with comparability in cross-cultural survey research.

REFERENCES

Barry, M. (1988). Ethical considerations of human investigation in developing countries: The AIDS dilemma. *New England Journal of Medicine* 319(16): 1083–1086.

Beatty, P.C. and Willis, G.B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly* 71: 287–311.

Behr, D., Braun, M., Kaczmirek, L. and Bandilla, W. (2012). Item comparability in cross-national surveys: Results from asking probing questions in cross-national Web surveys about attitudes towards civil disobedience. *Quality & Quantity*. DOI: 10.1007/s11135-012-9754-8.

Benstead, L.J. (in-press). Effects of interviewer-respondent gender interaction on attitudes toward women and politics: Findings from Morocco. *International Journal of Public Opinion Research*. Advanced access at: http://ijpor.oxfordjournals.org/content/early/2013/09/27/ijpor.edt024.full.pdf+html. Accessed on 6 June 2016.

Biemer, P.P., Groves, R.M., Lyberg, L., Mathiowetz, N.A. and Sudman, S. (eds) (1991). *Measurement Errors in Surveys*. New York: Wiley.

Blaydes, L. and Gillum, R.M. (2013). Religiosity-of-interviewer effects: Assessing the impact of veiled enumerators on survey response in Egypt. *Politics and Religion*. Accessed on April 25, 2015 at: http://web.stanford.edu/~blaydes/Veil3.pdf.

Braun, M., Behr, D. and Kaczmirek, L. (2012). Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in Web Surveys. *International Journal of Public Opinion Research*. DOI:10.1093/ijpor/eds034.

Braun, M. and Johnson, T.P. (2010). An illustrative review of techniques for detecting inequivalences, pp. 375–393 in J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P.P. Mohler, B.-E. Pennell and T. Smith (eds), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, NJ: Wiley.

Braun, M. and Mohler, P.P. (2003). Background variables, in J. Harkness, F.J.R. van de Vijver and P.P. Mohler (eds), *Cross-Cultural Survey Methods*. Hoboken, NJ: Wiley, pp. 101–116.

Braun, M. and Müller, W. (1997). Measurement of education in comparative research, in L. Mjøset et al. (eds), *Methodological Issues in Comparative Social Science*. Comparative Social Research 16, Greenwich, CT: JAI Press, pp. 163–201.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.

Buchanan, W. and Cantril, H. (1953). *How Nations See Each Other: A Study of Public Opinion*. Urbana: University of Illinois Press.

Bulmer, M. (1993). Interviewing and field organization, pp. 205–217 in M. Bulmer and D.P. Warwick (eds), *Social Research in Developing Countries: Surveys and Censuses in the Third World.* London: UCL Press.

Bulmer, M. (1998). Introduction: The problem of exporting social survey research. *American Behavioral Scientist* 42: 153–167.

Bulmer, M. and Warwick, D.P. (1993). *Social Research in Developing Countries: Surveys and Censuses in the Third World*. London: UCL Press.

Casley, D.J. and Lury, D.A. (1981). *Data Collection in Developing Countries*. Oxford: Clarendon Press.

Cicourel, A.V. (1974). *Theory and Method in a Study of Argentine Fertility*. New York: Wiley.

Clinard, M.B. and Elder, J.W. (1965). Sociology in India. *American Sociological Review* 30: 581–587.

Davidov, E., Schmidt, P. and Billiet, J. (eds) (2011). *Cross-cultural Analysis: Methods and Applications.* New York: Routledge.

Davis, R.E., Resnicow, K. and Couper, M.P. (2011). Survey response styles, acculturation, and culture among a sample of Mexican American adults. *Journal of Cross-Cultural Psychology* 42: 1219–1236.

Drake, H.M. (1973). Research method or culture-bound technique? Pitfalls of survey research in Africa, pp. 58–69 in W.M. O'Barr, D.H. Spain and M.A. Tessler (eds), *Survey Research in Africa: Its Applications and Limits.* Evanston, IL: Northwestern University Press.

Duijker, H.C.J. and Rokkan, S. (1954) Organizational aspects of cross-national social research. *Journal of Social Issues* 10: 8–24.

Fitzgerald, R., Widdop, S., Gray, M. and Collins, D. (2011). Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. *Journal of Official Statistics* 27, 569–599.

Freed-Taylor, M. (1994). Ethical considerations in European cross-national research. *International Social Science Journal* 46: 523–532.

Frey, F.W. (1970). Cross-cultural survey research in political science, pp. 173–294 in T.T. Holt and J.E. Turner (eds), *The Methodology of Comparative Research*. New York: Free Press.

Galway, L.P., Bell, N., Sae, A.S., Hagopian, A., Burnham, G., Flaxman, A., Weiss, W.M., Takaro, T.K. (2012). A two-stage cluster sampling method using gridded population data, a GIS, and Google Earth™ imagery in a population-based mortality survey in Iraq. *International Journal of Health Geographics* 11: 12. Accessed on April 25, 2015 at: http://www.ij-healthgeographics.com/content/11/1/12.

Hantrais, L. (2009). *International Comparative Research: Theory, Methods and Practice.* New York: Palgrave Macmillan.

Harkness, J.A. (2007). In pursuit of quality: Issues for cross-national survey research, pp. 35–50 in L. Hantrais and S. Mangen