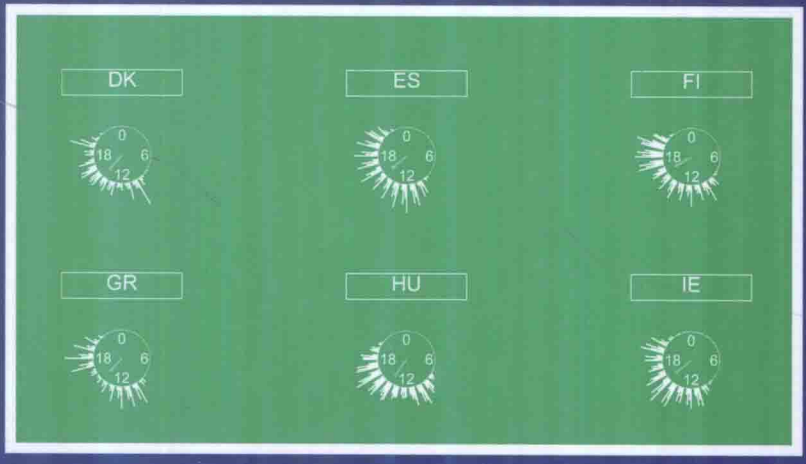
Wiley Series in Survey Methodology

Improving Surveys with Paradata

Analytic Uses of Process Information



Edited by

Frauke Kreuter

WILEY

IMPROVING SURVEYS WITH PARADATA

Analytic Uses of Process Information

Edited by

FRAUKE KREUTER
Joint Program in Survey Methodology, University of Maryland
Institute for Employment Research, Nuremberg
Ludwig Maximilian University, Munich

WILEY

Cover Design: John Wiley & Sons, Inc.
Cover Illustration: Courtesy of Frauke Kreuter

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

Library of Congress Cataloging-in-Publication Data:

Improving surveys with paradata: analytic uses of process information / [edited by] Frauke Kreuter, University of Maryland, College Park, Maryland, Institute for Employment Research, Nuremberg, Ludwig-Maximilians-University, Munich.
pages cm
Includes bibliographical references and index.
ISBN 978-0-470-90541-8 (cloth)
1. Surveys–Statistical methods. 2. Social surveys–Statistical methods. 3. Social sciences–Research–Statistical methods. I. Kreuter, Frauke.
HA31.2.147 2013
001.4′33–dc23

2013000328

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

IMPROVING SURVEYS WITH PARADATA

WILEY SERIES IN SURVEY METHODOLOGY

Established in Part by WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *Mick P. Couper, Graham Kalton, J. N. K. Rao, Norbert Schwarz, Christopher Skinner*
Editor Emeritus: *Robert M. Groves*

A complete list of the titles in this series appears at the end of this volume.

PREFACE

Newspapers and blogs are now filled with discussions about "big data," massive amounts of largely unstructured data generated by behavior that is electronically recorded. "Big data" was the central theme at the 2012 meeting of the World Economic Forum and the U.S. Government issued a Big Data Research and Development Initiative the same year. The American Statistical Association has also made the topic a theme for the 2012 and 2013 Joint Statistical Meetings.

Paradata are a key feature of the "big data" revolution for survey researchers and survey methodologists. The survey world is peppered with process data, such as electronic records of contact attempts and automatically captured mouse movements that respondents produce when answering web surveys. While not all of these data sets are massive in the usual sense of "big data," they are often highly unstructured, and it is not always clear to those collecting the data which pieces are relevant, and how they should be analyzed. In many instances it is not even obvious which data are generated.

Recently Axel Yorder, the CEO of the company Webtrends, pointed out that just as "Gold requires mining and processing before it finds its way into our jewelry, electronics, and even the Fort Knox vault [...] data requires collection, mining and, finally, analysis before we can realize its true value for businesses, governments, and individuals alike."[1] The same can be said for paradata. Paradata are data generated in the process of conducting a survey. As such, they have the potential to shed light on the survey process itself, and with proper "mining" they can point to errors and breakdowns in the process of data collection. If captured and analyzed immediately paradata can assist

[1] http://news.cnet.com/8301-1001_3-57434736-92/big-data-is-worth-nothing-without-big-science/

with efficiency during data collection field period. After data collection ends, paradata that capture measurement errors can be modeled alongside the substantive data to increase the precision of resulting estimates. Paradata collected for respondents and nonrespondents alike can be useful for nonresponse adjustment. As discussed in several chapters in this volume, paradata can lead to efficiency gains and cost savings in survey data production. This has been demonstrated in the U.S. National Survey of Family Growth conducted by the University of Michigan and the National Center for Health Statistics.

However, just as for big data in general, many questions remain about how to turn paradata into gold. Different survey modes allow for the collection of different types of paradata, and depending on the production environment, paradata may be instantaneously available. Fast-changing data collection technology will likely open doors to real-time capture and analysis of even more paradata in ways we cannot currently imagine. Nevertheless some general principles regarding the logic, design, and use of paradata will not change, and this book discusses these principles. Much work in this area is done within survey research agencies and often does not find its way into print, thus this book also serves as a vehicle to share current developments in paradata research and use.

This book came to life during a conference sponsored by the Institute for Employment Research in Germany, November of 2011 when most of the chapter authors participated in a discussion about it. The goal was to write a book that goes into more detail than published papers on the topic. Because this research area is relatively new we saw the need to collect information that is otherwise not easily accessible and to give practitioners a good starting point for their own work with paradata. The team of authors decided to use a common framework and standardized notation as much as possible. We tried to minimize overlap across the chapters without hampering the possibility for each chapter to be read on its own. We hope the result will satisfy the needs of researchers starting to use paradata as well as those who are already experienced. We also hope it will inspire readers to expand the use of paradata to improve survey data quality and survey processes. As we strive to update our knowledge on behalf of all authors, I ask you to tell us about your successes and failures in dealing with paradata.

We dedicate this volume to Mick Couper and Robert Groves. Mick Couper coined the term "paradata" in a presentation at the 1998 Joint Statistical Meeting in Dallas where he discussed the potential of paradata to reduce measurement error. For his vision regarding paradata he was awarded the American Association for Public Opinion Research's Warren J. Mitofsky Innovators Award in 2008. As the director of the University of Michigan Survey Research Center and later as Director of the U.S. Census Bureau, Robert Groves implemented new ideas on the use of paradata to address nonresponse, showing the breadth of applications paradata have to survey errors and operational challenges. After a research seminar in the Joint Program in Survey Methodology on this topic, I remember him saying: "You should write a book on paradata!" Both Mick and Bob have been fantastic teachers and mentors for most of the chapter authors and outstanding colleagues to all. Their perspectives on Survey

Methodology and the Total Survey Error Framework are guiding principles visible in each of the chapters.

I personally also want to thank Rainer Schnell for exposing me to paradata before they were named as such. As part of the German DEFECT project that he led, we walked through numerous villages and cities in Germany to collect addresses. In this process we took pictures of street segments and recorded, on the first generation of handheld devices, observations and judgments about the selected housing units. Elizabeth Coutts, my dear friend and colleague in this project, died on August 5, 2009, but her ingenious contributions to the process of collecting these paradata will never be forgotten.

We are very grateful to Paul Biemer, Lars Lyberg and Fritz Scheuren for actively pushing the paradata research agenda forward and for making important contributions by putting paradata into the context of statistical process control and the larger metadata initiatives. This book benefitted from discussions at the International Workshop on Household Survey Nonresponse and the International Total Survey Error Workshop and we are in debt to all of the researchers who shared their work and ideas at these venues over the years. In particular, we thank Nancy Bates, James Dahlhamer, Mirta Galesic, Barbara O'Hare, Rachel Horwitz, François Laflamme, Lars Lyberg, Andrew Mercer Peter Miller and Stanley Presser for comments on parts of this book. Our thanks also goes to Ulrich Kohler for creating the cover page graph.

The material presented here provided the basis for several short courses taught during the Joint Statistical Meeting of the American Statistical Association, continuing education efforts of the U.S. Census Bureau, the Royal Statistical Society, and the European Social Survey. The feedback I received from course participants helped to improve this book, but remaining errors are entirely ours.

On the practical side, this book would not have found its way into print without our LaTeX wizard Alexandra Birg, the constant pushing of everybody involved at Wiley, and the support from the Joint Program in Survey Methodology in Maryland, the Institute for Employment Research in Nuremberg, and the Department of Statistics at the Ludwig Maximilian University in Munich. We thank you all.

FRAUKE KREUTER

Washington D.C.
September, 2012

CONTRIBUTORS

Melania Calinescu, VU University Amsterdam, NL

Mario Callegaro, Google London, UK

Julia D'Arrigo, Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, Southampton, UK

Gabriele B. Durrant, Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, Southampton, UK

Stephanie Eckman, Institute for Employment Research (IAB), Nuremberg, Germany

Matt Jans, University of California Los Angeles, Los Angeles, California, USA

Nicole G. Kirgis, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA

Frauke Kreuter, Institute for Employment Research (IAB), Nuremberg, Germany; University of Maryland, College Park, Maryland, USA; Ludwig Maximilian University, Munich, Germany

James M. Lepkowski, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA

David Morgan, U.S. Census Bureau, Washington, DC, USA

Gerrit Müller, Institute for Employment Research (IAB), Nuremberg, Germany

Kristen Olson, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

BRYAN PARKHURST, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

JOSEPH W. SAKSHAUG, Institute for Employment Research (IAB), Nuremberg, Germany

JOSEPH L. SCHAFER, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, USA

BARRY SCHOUTEN, Statistics Netherlands, Den Haag and University of Utrecht, NL

JENNIFER SINIBALDI, Institute for Employment Research (IAB), Nuremberg, Germany

ROBYN SIRKIS, U.S. Census Bureau, Washington DC, USA

JAMES WAGNER, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan

BRADY T. WEST, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA

TING YAN, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA

ACRONYMS

AAPOR	American Association for Public Opinion Research
ACASI	Audio Computer-Assisted Self-Interview
ACS	The American Community Survey
AHEAD	Assets and Health Dynamics Among the Oldest Old
ANES	American National Election Studies
BCS	British Crime Survey
CAI	Computer-Assisted Interviewing
CAPI	Computer-Assisted Personal Interviews
CARI	Computer-Assisted Recording of Interviews
CASRO	Council of American Survey Research Organizations
CATI	Computer-Assisted Telephone Interviews
CE	Consumer Expenditure Interview Survey
CHI	Contact History Instrument
CHUM	Check for Housing Unit Missed
CPS	Current Population Survey
CSP	Client-side Paradata
ESOMAR	European Society for Opinion and Market Research
ESS	European Social Survey
FRS	Family Resources Survey
GSS	General Social Survey
HINTS	Health Information National Trends Study
HRS	Health and Retirement Study
IAB	Institute for Employment Research
IVR	Interactive Voice Response System
KPI	Key Performance Indicators

LAFANS	Los Angeles Family and Neighborhood Study
LCL	Lower Control Limits
LFS	Labour Force Survey
LISS	Dutch Longitudinal Internet Studies for the Social Sciences
LMU	Ludwig Maximilian University Munich
NCHS	National Center for Health Statistics
NHANES	National Health and Nutrition Examination Survey
NHEFS	The NHANES Epidemiologic Follow-up Study
NHIS	National Health Interview Survey
NSDUH	National Survey of Drug Use and Health
NSFG	National Survey of Family Growth
NSHAP	National Social Life, Health, and Aging Project
NSR	Non-self Representing
OMB	Office of Management and Budget
PASS	Panel Study of Labour Market and Social Security
PDA	Personal Digital Assistant
PSU	Primary Sampling Units
RDD	Random Digit Dial
RECS	Residential Energy Consumption Survey
RMSE	Root Mean Squared Error
RO	Regional Office
SCA	Survey of Consumer Attitudes
SCF	Survey of Consumer Finances
SHS	Survey of Household Spending
SPC	Statistical Process Control
SQC	Statistical Quality Control
SR	Self-Representing Areas
UCL	Upper Control Limits
UCSP	Universal Client Side Paradata

CONTENTS

PART I PARADATA AND SURVEY ERRORS