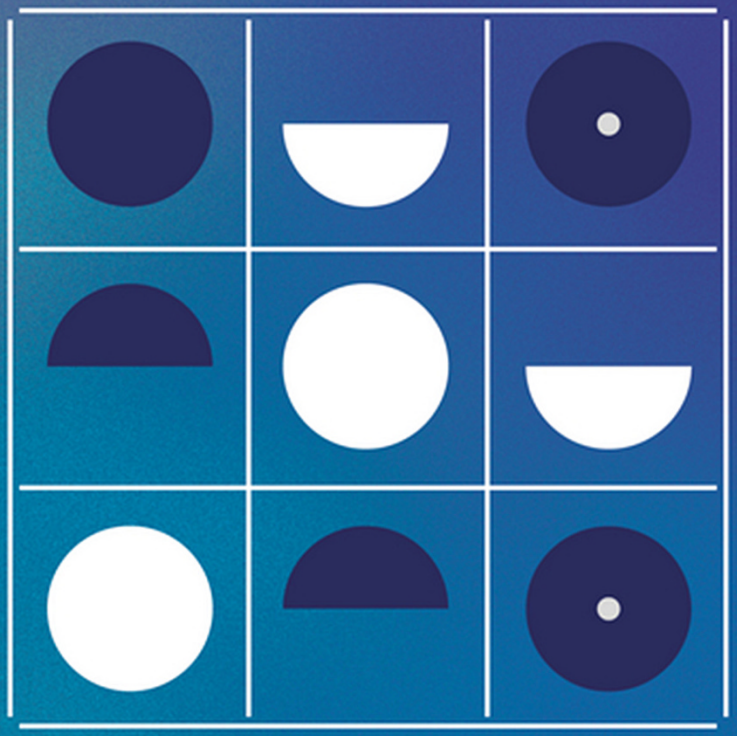
Wiley Series in Survey Methodology

Total Survey Error in Practice



Editors:
Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards,
Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, and Brady T. West

WILEY

Total Survey Error in Practice

WILEY SERIES IN SURVEY METHODOLOGY

Established in Part by Walter A. Shewhart and Samuel S. Wilks

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

Total Survey Error in Practice

Edited by

Paul P. Biemer
RTI International and University of North Carolina

Edith de Leeuw
Utrecht University

Stephanie Eckman
RTI International

Brad Edwards
Westat

Frauke Kreuter
Joint Program in Survey Methodology, University of Mannheim, Institute for Employment Research (Germany)

Lars E. Lyberg
Inizio

N. Clyde Tucker
American Institutes for Research

Brady T. West
University of Michigan and Joint Program in Survey Methodology

WILEY

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

Library of Congress Cataloging-in-Publication Data

Names: Biemer, Paul P., editor. | Leeuw, Edith de, editor. | Eckman, Stephanie, editor. | Edwards, Brad, editor. | Kreuter, Frauke, editor. | Lyberg, Lars E., editor. | Tucker, N. Clyde, editor. | West, Brady T., editor.
Title: Total survey error in practice / edited by Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, Brady T. West.
Description: Hoboken, New Jersey : John Wiley & Sons, 2017. | Includes index.
Identifiers: LCCN 2016031564 | ISBN 9781119041672 (cloth) | ISBN 9781119041696 (epub)
Subjects: LCSH: Error analysis (Mathematics) | Surveys.
Classification: LCC QA275 .T685 2016 | DDC 001.4/33–dc23
LC record available at https://lccn.loc.gov/2016031564

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Notes on Contributors

Manfred Antoni
Research Data Centre (FDZ)
Institute for Employment Research (IAB)
Nuremberg
Germany

Christopher Antoun
Center for Survey Measurement
U.S. Census Bureau
Suitland, MD
USA

Reg Baker
Marketing Research Institute International
Ann Arbor, MI
USA

Stefan Bender
Research Data and Service Centre
Deutsche Bundesbank
Frankfurt am Main
Germany

Grant Benson
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Heather Bergdahl
Process Department
Statistics Sweden
Stockholm
Sweden

Marcus E. Berzofsky
Division for Statistics and Data Science
RTI International
Research Triangle Park, NC
USA

Paul P. Biemer
Social, Statistical, and Environmental Sciences
RTI International
Research Triangle Park, NC
Odum Institute for Research in Social Science
University of North Carolina
Chapel Hill, NC
USA

Paul Burton
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Christine Bycroft
Statistics New Zealand
Wellington
New Zealand

Jennifer Hunter Childs
Research and Methodology Directorate
U.S. Census Bureau
Washington, DC
USA

Sue Connor
Westat
Rockville, MD
USA

Frederick G. Conrad
Survey Research Center
University of Michigan
Ann Arbor, MI
Joint Program in Survey Methodology
University of Maryland
College Park, MD
USA

Mick P. Couper
Survey Research Center
University of Michigan
Ann Arbor, MI
Joint Program in Survey Methodology
University of Maryland
College Park, MD
USA

Edith de Leeuw
Department of Methodology and Statistics
Utrecht University
Utrecht
The Netherlands

Stephanie Eckman
Survey Research Division
RTI International
Washington, DC
USA

Brad Edwards
Westat
Rockville, MD
USA

Barbara Felderer
Collaborative Research Center SBF 884
"Political Economy of Reforms"
University of Mannheim
Mannheim
Germany

Jamie Griffin
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Heidi Guyer
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Kristen Cibelli Hibben
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Daniela Hochfellner
Center for Urban Science
and Progress
New York University
New York, NY
USA

Anders Holmberg
Statistics Norway
Oslo
Norway

Joop Hox
Department of Methodology
and Statistics
Utrecht University
Utrecht
The Netherlands

Yuli Patrick Hsieh
Survey Research Division
RTI International
Chicago, IL
USA

Frost Hubbard
Survey Solutions Division
IMPAQ International
Columbia, MD
USA

Andrew L. Hupp
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Joost Kappelhof
Department of Education, Minorities, and Methodology
Institute for Social Research/SCP
The Hague
The Netherlands

Alan F. Karr
Center of Excellence for Complex Data Analysis
RTI International
Research Triangle Park, NC
USA

Jennifer Kelley
Institute for Social and Economic Research
University of Essex
Colchester
UK

Meena Khare
National Center for Health Statistics
Centers for Disease Control and Prevention
Hyattsville, MD
USA

Yumi Kim
Department of Research Methods
Market Strategies International
Livonia, MI
USA

Antje Kirchner
Department of Sociology
University of Nebraska-Lincoln
Lincoln, NE
Survey Research Division
RTI International
Research Triangle Park, NC
USA

Thomas Klausch
Department for Epidemiology and Biostatistics
VU University Medical Center
Amsterdam
The Netherlands

Frauke Kreuter
Joint Program in Survey Methodology
University of Maryland
College Park, MD
USA
Department of Sociology
University of Mannheim
Mannheim
Statistical Methods Group
Institute for Employment Research (IAB)
Nuremberg
Germany

John LaFrance
Market Strategies International
Livonia, MI
USA

Thomas Laitila
Department of Research and Development
Statistics Sweden
Department of Statistics
Örebro University School of Business
Örebro
Sweden

JiaoJiao Li
Market Strategies International
Livonia, MI
USA

Karin Lindgren
Process Department
Statistics Sweden
Stockholm
Sweden

Peter J. Lugtig
Institute for Social and Economic Research
University of Essex
Colchester
UK
Department of Methodology and Statistics
Utrecht University
Utrecht
The Netherlands

Lars E. Lyberg
Inizio
Stockholm
Sweden

Peter Lynn
Institute for Social and Economic Research
University of Essex
Colchester
UK

Aaron Maitland
Westat
Rockville, MD
USA

Aigul Mavletova
Department of Sociology
National Research University Higher School of Economics
Moscow
Russia

Peter Ph. Mohler
University of Mannheim
Mannheim
Germany

William D. Mosher
Bloomberg School of Public Health
Johns Hopkins University
Baltimore, MD
USA

Mary H. Mulry
Research and Methodology Directorate
U.S. Census Bureau
Washington, DC
USA

Joe Murphy
Survey Research Division
RTI International
Chicago, IL
USA

Elizabeth M. Nichols
Research and Methodology Directorate
U.S. Census Bureau
Washington, DC
USA

Anders Norberg
Process Department
Statistics Sweden
Stockholm
Sweden

Daniel L. Oberski
Department of Methodology and Statistics
Utrecht University
Utrecht
The Netherlands

Beth-Ellen Pennell
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Gregg Peterson
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Vicki J. Pineau
NORC at the University of Chicago
Chicago, IL
USA

Joseph W. Sakshaug
Cathie Marsh Institute for
Social Research
University of Manchester
Manchester
UK
Department of Statistical Methods
Institute for Employment Research (IAB)
Nuremberg
Germany

Michael F. Schober
Department of Psychology
New School for Social Research
New York, NY
USA

James A. Singleton
National Center for Immunization and Respiratory Diseases
Centers for Disease Control and Prevention
Atlanta, GA
USA

Benjamin Skalland
NORC at the University of Chicago
Chicago, IL
USA

Philip J. Smith
National Center for Immunization and Respiratory Diseases
Centers for Disease Control and Prevention
Atlanta, GA
USA

Diana Maria Stukel
FHI 360
Washington, DC
USA

Can Tongur
Process Department
Statistics Sweden
Stockholm
Sweden

Roger Tourangeau
Westat
Rockville, MD
USA

Dennis Trewin
Former Australian Statistician
Australian Bureau of Statistics
Canberra
Australia

James Wagner
Survey Research Center
University of Michigan
Ann Arbor, MI
Joint Program in Survey Methodology
University of Maryland
College Park, MD
USA

Brady T. West
Survey Research Center
University of Michigan
Ann Arbor, MI
Joint Program in Survey Methodology
University of Maryland
College Park, MD
USA

Kirk M. Wolter
NORC at the University of Chicago
Chicago, IL
USA

Gelaye Worku
Department of Statistics
Stockholm University
Stockholm
Sweden

Yingfu Xie
Process Department
Statistics Sweden
Stockholm
Sweden

H. Yanna Yan
Survey Research Center
University of Michigan
Ann Arbor, MI
USA

Ting Yan
Methodology Unit
Westat
Rockville, MD
USA

David Yankey
National Center for Immunization and Respiratory Diseases
Centers for Disease Control and Prevention
Atlanta, GA
USA

Wei Zeng
NORC at the University of Chicago
Chicago, IL
USA

Zhen Zhao
National Center for Immunization and Respiratory Diseases
Centers for Disease Control and Prevention
Atlanta, GA
USA

Preface

Total survey error (TSE) refers to the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data. In this context, a survey error can be defined as any error contributing to the deviation of an estimate from its true parameter value. Survey errors arise from misspecification of concepts, sample frame deficiencies, sampling, questionnaire design, mode of administration, interviewers, respondents, data capture, missing data, coding, and editing. Each of these error sources can diminish the accuracy of inferences derived from the survey data. A survey estimate will be more accurate when bias and variance are minimized, which occurs only if the influence of TSE on the estimate is also minimized. In addition, if major error sources are not taken into account, various measures of margins of error are understated, which is a major problem for the survey industry and the users of survey data.

Because survey data underlie many public policy and business decisions, a thorough understanding of the effects of TSE on data quality is needed. The TSE framework, the focus of this book, is a valuable tool for understanding and improving survey data quality. The TSE approach summarizes the ways in which a survey estimate may deviate from the corresponding parameter value. Sampling error, measurement error, and nonresponse error are the most recognized sources of survey error, but the TSE framework also encourages researchers not to lose sight of the less commonly studied error sources, such as coverage error, processing error, and specification error. It also highlights the relationships between errors and the ways in which efforts to reduce one type of error can increase another, resulting in an estimate with more total error. For example, efforts to reduce nonresponse error may unintentionally lead to measurement errors, or efforts to increase frame coverage may lead to greater nonresponse.

This book is written to provide a review of the current state of the field in TSE research. It was stimulated by the first international conference on TSE that was held in Baltimore, Maryland, in September 2015 (http://www.TSE15.org). Dubbed TSE15, the conference had as its theme, "Improving Data Quality in the Era of Big Data." About 140 papers were presented at the conference which was attended by approximately 300 persons. The conference itself was the culmination of a series of annual workshops on TSE called the International TSE Workshops (ITSEWs) which began in 2005 and still continue to this day. This book is an edited volume of 25 invited papers presented at the 2015 conference spanning a wide range of topics in TSE research and applications.

TSE15 was sponsored by a consortium of professional organizations interested in statistical surveys—the American Association of Public Opinion Research (AAPOR), three sections of the American Statistical Association (Survey Research Methods, Social Statistics, and Government Statistics), the European Survey Research Association (ESRA), and the World Association of Public Opinion Research (WAPOR). In addition, a number of organizations offered financial support for the conference and this book. There were four levels of contributions. Gallup,

Inc. and AC Nielsen contributed at the highest level. At the next highest level, the contributors were NORC, RTI International, Westat, and the University of Michigan (Survey Research Center). At the third level were Mathematica Policy Research, the National Institute of Statistical Sciences (NISS), and Iowa State University. Finally, the Council of Professional Associations on Federal Statistics (COPAFS) and ESOMAR World Research offered in-kind support. We are deeply appreciative of the sponsorship and support of these organizations which made the conference and this book possible.

Stephanie Eckman (RTI International) and Brad Edwards (Westat) cochaired the conference and the organizing committee, which included Paul P. Biemer (RTI International), Edith de Leeuw (Utrecht University), Frauke Kreuter (University of Maryland), Lars E. Lyberg (Inizio), N. Clyde Tucker (American Institutes for Research), and Brady T. West (University of Michigan). The organizing committee also did double duty as coeditors of this volume. Paul P. Biemer led the editorial committee.

This book is divided into five sections, each edited, primarily, by three members of the editorial team. These teams worked with the authors over the course of about a year and were primarily responsible for the quality and clarity of the chapters. The sections and their editorial teams were the following.

Section 1: The Concept of TSE and the TSE Paradigm (Editors: Biemer, Edwards, and Lyberg). This section, which includes Chapters 1 through 4, provides conceptual frameworks useful for understanding the TSE approach to design, implementation, evaluation, and analysis and how the framework can be extended to encompass new types of data and their inherent quality challenges.

Section 2: Implications for Survey Design (Editors: De Leeuw, Kreuter, and Eckman). This section includes Chapters 5 through 11 and provides methods and practical applications of the TSE framework to multiple-mode survey designs potentially involving modern data collection technologies and multinational and multicultural survey considerations.

Section 3: Data Collection and Data Processing Applications (Editors: Edwards, Eckman, and de Leeuw). This section includes Chapters 12 through 15 and focuses on issues associated with applying the TSE framework to control costs and errors during data collection activities.

Section 4: Evaluation and Improvement (Editors: West, Biemer, and Tucker). This section includes Chapters 16 through 21 and describes a range of statistical methods and other approaches for simultaneously evaluating multiple error sources in survey data and mitigating their effects.

Section 5: Estimation and Analysis (Editors: Kreuter, Tucker, and West). This section includes Chapters 22 through 25 which deal with issues such as the appropriate analysis of survey data subject to sampling and nonsampling errors, potential differential biases associated with data collected by mixed modes and errors in linking records, and reducing these errors in modeling, estimation, and statistical inferences.

The edited volume is written for survey professionals at all levels, from graduate students in survey methodology to experienced survey practitioners wanting to imbue cutting-edge principles and practices of the TSE paradigm in their work. The book highlights use of the TSE framework to understand and address issues of data quality in official statistics and in social, opinion, and market research. The field of statistics is undergoing a revolution as data sets get bigger (and messier), and understanding the potential for data errors and the various means to control and prevent them is more important than ever. At the same time, survey organizations are challenged to collect data more efficiently without sacrificing quality.

Finally, we, the editors, would like to thank the authors of the chapters herein for their diligence and support of the goal of providing this current overview of a dynamic field of research.

We hope that the significant contributions they have made in these chapters will be multiplied many times over by the contributions of readers and other methodologists as they leverage and expand on their ideas.

Paul P. Biemer
Edith de Leeuw
Stephanie Eckman
Brad Edwards
Frauke Kreuter
Lars E. Lyberg
N. Clyde Tucker
Brady T. West

Section 1

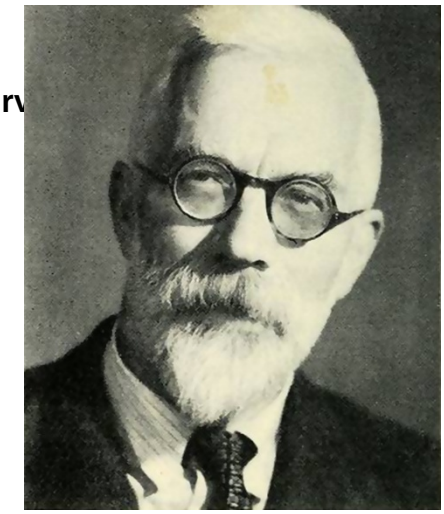
The Concept of TSE and the TSE Paradigm

1

The Roots and Evolution of the Total Surv

Lars E. Lyberg[1] *and Diana Maria Stukel*[2]

[1] *Inizio, Stockholm, Sweden*
[2] *FHI 360, Washington, DC, USA*



Sir Ronald Fisher

Jerzy Neyman

1.1 Introduction and Historical Backdrop

In this chapter, we discuss the concept of total survey error (TSE), how it originated and developed both as a mindset for survey researchers and as a criterion for designing surveys. The interest in TSE has fluctuated over the years. When Jerzy Neyman published the basic sampling theory and some of its associated sampling schemes in 1934 onward, it constituted the first building block of a theory and methodology for surveys. However, the idea that a sample could be used to represent an entire population was not new. The oldest known reference to estimating a finite population total on the basis of a sample dates back to 1000 BC and is found in the Indian epic Mahabharata (Hacking, 1975; Rao, 2005). Crude attempts at measuring parts of a population rather than the whole had been used in England and some other European countries quite extensively between 1650 and 1800. The methods on which these attempts were based were referred to as *political arithmetic* (Fienberg and Tanur, 2001), and they resembled ratio estimation using information of birth rates, family size, and average number of persons living in selected buildings and other observations. In 1895, at an International Statistical Institute meeting, Kiaer argued for developing a representative or partial investigation method

Total Survey Error in Practice, First Edition. Edited by Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, and Brady T. West.

(Kiaer, 1897). The representative method aimed at creating a sample that would reflect the composition of the population of interest. This could be achieved by using balanced sampling through purposive selection or various forms of random sampling. During the period 1900–1920, the representative method was used extensively, at least in Russia and the U.S.A. In 1925, the International Statistical Institute released a report on various aspects of random sampling (Rao, 2005, 2013; Rao and Fuller, 2015). The main consideration regarding sampling was likely monetary, given that it was resource-intensive and time-consuming to collect data from an entire population. Statistical information compiled using a representative sample was an enormous breakthrough. But it would be almost 40 years after Kiaer's proposal before Neyman published his landmark paper from 1934 "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." At this time, there existed some earlier work by the Russian statistician Tschuprow (1923a, b) on stratified sampling and optimal allocation. It is not clear whether Neyman was aware of this work when he started to develop the sampling theory in the 1920s (Fienberg and Tanur, 1996) since he did not mention Tschuprow's work when discussing optimal allocation. Neyman definitely had access to Ronald Fisher's (1925) ideas on randomization (as opposed to various kinds of purposive selection) and their importance for the design and analysis of experiments, and also to Bowley's (1926) work on stratified random sampling.

Prasanta Mahalanobis

Morris Hansen

The sampling methods proposed by Neyman were soon implemented in agencies such as the Indian Statistical Institute and the U.S. Bureau of the Census (currently named the U.S. Census Bureau). Prasanta Mahalanobis, the founder of the Indian Statistical Institute, and Morris Hansen and colleagues at the U.S. Census Bureau, became the main proponents of scientific sampling in a number of surveys in the 1940s. The development was spurred on by Literary Digest's disastrously inaccurate prediction in the 1936 U.S. presidential election poll that was based on a seriously deficient sampling frame. However, Neyman's sampling theory did not take into account nonsampling errors and relied on the assumption that sampling was the only major error source that affected estimates of population parameters and associated calculations of confidence intervals or margins of error. However, Neyman and his peers understood that this was indeed an unrealistic assumption that might lead to understated margins of error. The effect of nonsampling errors on censuses was acknowledged and discussed in a German textbook on census methodology relatively early on (Zizek, 1921). The author discussed what he called *control of*

contents and coverage. In addition, Karl Pearson (1902) discussed observer errors much earlier than that. An early example of interviewer influence on survey response was the study on the consumption of hard liquor during the prohibition days in the U.S.A., where Rice (1929) showed that interviewers who were prohibitionists tended to obtain responses that mirrored their own views and that differed from those of respondents that were interviewed by other interviewers.

Edwards Deming

In 1944, Edwards Deming published the first typology of sources of error beyond sampling. He listed 13 factors that he believed might affect the utility of a survey. The main purpose of the typology was to demonstrate the need for directing efforts to all potential sources in the survey planning process while considering the resources available. This first typology included some error sources that are not frequently referenced today, such as bias of the auspices (i.e., the tendency to indicate a particular response because of the organization sponsoring the study). Others, to which more attention is currently given, such as coverage error, were not included, however. Even though Deming did not explicitly reference TSE, he emphasized the limitations of concentrating on a few error sources only and highlighted the need for theories of bias and variability based on accumulated experience.

Rapid development of the area followed shortly thereafter. Mahalanobis (1946) developed the *method of interpenetration*, which could be used to estimate the variability generated by interviewers and other data collectors. Another error source recognized early on was nonresponse. Hansen and Hurwitz (1946) published an article in the *Journal of the American Statistical Association* on follow-up sampling from the stratum of initial nonrespondents. While the basic assumption of 100% participation in a follow-up sample was understood not to be realistic, at the time, there were relatively small nonresponse rates, and it was possible to estimate, at least approximately, the characteristics of those in the nonresponse stratum.

Even though it is not explicitly stated, TSE has its roots in cautioning against sole attention focused on sampling error along with possibly one or two other error sources, rather than the entire scope of potential errors. In response, two lines of strategic development occurred. One strategy entailed the identification of *specific error sources*, coupled with an attempt to control them or at least minimize them. The other strategy entailed the development of the so-called *survey error models*, where the TSE was decomposed and the magnitude of different error components, and ultimately the combination of them (i.e., the TSE), could be estimated. The two strategies were intertwined in the sense that a survey model could be applied not only on the entire set of survey operations but also on a subset of specific survey operations.

1.2 Specific Error Sources and Their Control or Evaluation

Apart from that of Deming (1944), there are a number of typologies described in the survey literature. Examples include Kish (1965), Groves (1989), Biemer and Lyberg (2003), Groves et al. (2009), Smith (2011), and Pennell et al. (Chapter 9 in this volume). Some of them are explicitly labeled TSE, while others consist of listings of different types of errors; however, all are incomplete. In some cases, known error sources (as well as their interactions with other error

sources) are simply omitted, and in other cases, all possible error sources are not known or the sources defy expression. For instance, new error structures have emerged when new data collection modes or new data sources, such as Big Data (see, e.g., Chapter 3 in this volume), have become popular—but the comprehension and articulation of the associated error structures have lagged in time.

Leslie Kish

Early on, the work toward the treatment of specific error sources followed two separate types of strategies: control and evaluation.

Related to the first strategy of control, one line of thinking was that statistical agencies were "data factories" that produced tables and limited analyses as their outputs. As such, they resembled an industrial assembly line. Therefore, the application of methods for industrial quality control (QC) was deemed suitable. Several statistical agencies adopted this approach for some of their operations, and the U.S. Census Bureau was clearly at the forefront. Most of these QCs were focused toward manual operations such as enumeration and interviewing, listing, coding, card punching, and editing, although it was also possible to use QC to check automatic operations such as scanning, which at the time was implemented through Film Optical Sensing Device for Input to Computers (FOSDIC). For the manual operations, the main control method was verification, where one operator's work was checked by another operator. A long list of census methodologists including Morris Hansen, Bill Hurwitz, Eli Marks, Edwards Deming, Ross Eckler, Max Bershad, Leon Pritzker, Joe Waksberg, Herman Fasteau, and George Minton made very significant contributions to this QC development. Contributions included those of Deming et al. (1942), Hansen and Steinberg (1956), Hansen et al. (1962), and the U.S. Bureau of the Census (1965).

These QC schemes were adapted from their industrial applications, and therefore were called "administrative applications of statistical QC." One example of this kind of scheme related to the coding of variables with respect to Industry and Occupation (Fasteau et al., 1964). During that era, a coder's work was typically verified by one or more coders in a dependent or independent way. To protect the users of data, acceptance sampling schemes were applied. Under such schemes, coding cases were bundled together in lots and sample inspection took place. If the number of coding errors was below or equal to an acceptance number, the lot was accepted. However, if the number of coding errors exceeded the acceptance number, the lot underwent 100% inspection, after which a decision was made that a coder should either remain on sampling control or remain under total control until results improved. An added complication was the institution of a point system that was imposed on the coders. Under the point system, the coder was given an initial allotment of three points. When a favorable quality decision was made, the coder received one more point. Otherwise, he or she lost one point. When the accumulated point balance reached zero, remedial action was taken toward the coder either in the form of additional training or dismissal from the operation. To avoid excessive accumulation of points that might culminate during a long period and that might mask substandard coding, the accumulated score was adjusted after every 10th decision. If the accumulated score was above 3 after the 10th decision it was reduced to 3. If the accumulated score was 3, 2, or 1, the coders maintained their current score (Minton, 1970).

One element that was often lacking with this factory approach was productive feedback, because at the time, root cause analysis was not really seen as a priority and "rework" was the prescription. Acceptance sampling was later vigorously criticized by Deming (1986), who claimed that under such a system, continuous improvement could not be achieved.

During the next decade, these schemes became increasingly complicated and were eventually abandoned in place of automated systems (Minton, 1972). It should be mentioned, though, that coding errors could be and still remain to this day quite substantial. Even today, gross errors, the difference between a production code and a verification code, in the range of 10–20% are not unusual. In present day systems, coding is often performed by software, but the error source itself is still basically neglected in most statistical agencies (Groves and Lyberg, 2010). The contributing factors are, in part, due to lack of software upgrades and minimal control of residual manual coding.

Another source of nonsampling error that received a lot of attention over the years is unit nonresponse. In the 1950s and 1960s, nonresponse was seen as catastrophic in terms of the ability to ensure high quality of survey results. Even modest nonresponse rates could trigger very unrealistic reactions, where fears that all nonrespondents might have values different from the respondents were prevalent. For instance, in the measurement of the unemployment rate, if in the extreme, all nonrespondents are assumed to be either employed or unemployed, it would then be possible to create max–min intervals that produced a much exaggerated picture of the risk and impact of nonresponse (Dalenius, 1961). This rigid view was later replaced by adjustment methods (Kalton and Kasprzyk, 1986), and theories and methods for missing data (Rubin, 1976). In addition, monographs on nonresponse and missing data (Groves et al., 2002; Madow et al., 1983) were written, as were textbooks on specific treatments of nonresponse such as multiple imputation (Rubin, 1987), theories of survey participation (Groves and Couper, 1998), and nonresponse in the international European Social Survey (Stoop et al., 2010). Brick (2013) reviewed various adjustment and compensation methods for unit nonresponse including the formation of weighting classes based on response propensities in different groups, as well as calibration methods, such as poststratification. In 1990, an international workshop on household survey nonresponse was initiated by Robert Groves, and this workshop still convenes annually; materials from the workshop are found on its website www.nonresponse.org.

Despite the development of methods for dealing with nonresponse, nonresponse rates increased considerably over the years in most countries. For instance, in 1970, the nonresponse rate in the Swedish Labor Force Survey was 2%, and currently in 2016 it is approximately 40%. However, a high nonresponse rate in isolation is not a solid indication of high nonresponse bias, since bias is also a function of the differences between respondents and nonrespondents with regard to the variables under study. As such, it is understood that sometimes nonresponse rates matter and sometimes not (Groves and Peychteva, 2008). Over the years, considerable energy has been devoted to developing methods that can help control the nonresponse rates and compensate for any residual nonresponse. Regardless, it is unlikely that in the foreseeable future, there will be any major declines in nonresponse rates, particularly given the recent proliferation of inexpensive high-technology modes of data collection.

Tore Dalenius

Two common methods of compensating for item nonresponse were developed—imputation and multiple imputation, both of which replace missing data with modeled or simulated data. For instance, simple forms of "hot deck imputation" were first introduced at the U.S. Census Bureau in 1947. The principles for early uses of various imputation procedures are described in Ogus et al. (1965), but these principles differ considerably from those used today. Initially, the justification for using imputation methods was to create rectangular data sets by filling in the holes generated by missing data, since it was considered very difficult to handle missing data computationally.[1] Consequently, the Census Bureau instituted very strict rules regarding the level of permissible imputations, whereby at most 2% item imputation was allowed, but if there were high demands on timeliness, this limit could be stretched to 5%. This is, of course, a far cry from today's use of imputation where allowable rates are much higher given the increased sophistication and resulting accuracy of present-day methods.

Yet another source of nonsampling error that was identified early on was that survey staff such as interviewers, enumerators, and coders could generate both systematic and variable errors. Mahalanobis (1946) invented the method of interpenetration for estimating interviewer effects by suggesting the assignment of a random subsample of all interviews to each interviewer rather than an assignment based on practical considerations (i.e., assigning the interviews for all selected individuals in a primary sampling unit). For field interviewing, interpenetration was, of course, more costly than assignments based on practicality, but studies showed that individual interviewer styles could introduce a substantial cluster effect that could not be ignored. Interpenetration methods demonstrated that respondents within an interviewer assignment tended to answer in ways that were intrinsic to that specific interviewer's style. Examples of style variation might include systematic deviations from the actual question wording or systematically inappropriate feedback on the part of the interviewers. Such errors could result in a correlated variance that is typically integrated as part of the total response variance but is not reflected in general variance estimates. Other operations mentioned earlier, such as coding, can also generate similar unaccounted for correlated variance, although they typically tend not to be large.

The topic of correlated variance is treated at length in Hansen et al. (1961) (see Section 1.3). Kish (1962) proposed an ANOVA model to estimate interviewer variance, and Bailar and Dalenius (1969) proposed basic study schemes to estimate the correlated variance components, of which interviewer effects is one (often substantial) part. It has been acknowledged that if survey conditions do not lend themselves to the ability to control interviewer errors, the effects can be dramatic. For instance, the World Fertility Survey Program has included cases of estimates whose variances were underestimated to an order of magnitude of 10 times, leading to strikingly understated margins of error (O'Muircheartaigh and Marckward, 1980). Unaccounted for correlated variance, such as in the aforementioned example, is the reason that standardized procedures have been instituted. Standardized procedures strive to ensure that interviewers, coders, and other groups work in the same way, thereby minimizing "cluster effects." Observing interviewers in the field and monitoring of telephone interviews are means to control deviations from the standardized protocol.

Despite the ability to standardize procedures to minimize interviewer effects, other measurement errors were also prevalent and remain a concern. These measurement errors include errors due to questionnaire wording and questionnaire topics, general cognitive phenomena associated with memory and mode of data collection, and errors in field manuals. In fact, phenomena such as telescoping, memory decay, social desirability bias, comprehension, and

1 It is acknowledged that listwise deletion methods (where an entire record is excluded from analysis if any single value is missing) are rather easy to implement but suffer from bias and variance issues.

respondent fatigue were acknowledged relatively early on and discussed in the survey literature (Belson, 1968; Neter and Waksberg, 1964; Sudman and Bradburn, 1974).

Even though most data collection agencies were aware that both measurement errors and processing errors could affect the quality of survey estimates, a substantial breakthrough did not occur until the release of the Jabine et al. (1984) report on Cognitive Aspects of Survey Methodology (CASM). The report emphasized the importance of measurement errors and their contributions to TSE, and defined response process models that have illuminated how some types of errors occur and how they can be mitigated. A response process model lays out the various cognitive steps a respondent undergoes from the survey participation request through to the delivery of his or her response. By disentangling these steps, it is possible to identify where the biggest risks are and how they should be dealt with. Response process models exist both for establishment surveys (Biemer and Fecso, 1995; Edwards and Cantor, 1991; Willimack and Nichols, 2010) and for surveys of individuals (Tourangeau et al., 2000).

The discussions and developments on controlling errors have followed different lines of thought over the years. For a large agency, such as the U.S. Census Bureau, rigorous controls of specific error sources were strongly advocated in the past. At the same time, there was a realization that extensive controls were expensive and their use had to be balanced against other needs. To the U.S. Census Bureau and other large producers of statistics, this imbalance was most obvious in the editing operation, which itself is a QC operation. Large amounts of resources were allocated to editing, which remains the case even today (de Waal et al., 2011). The purpose of these rigorous controls was to reduce biases and correlated variances, so that the TSE would consist mainly of sampling variance and simple response variance, both of which could be calculated directly from the data. This general strategy of controlling errors reduced survey biases to some extent. For example, nonresponse adjustments that take into account various response classes led to decreased nonresponse bias. Adherence to appropriate questionnaire design principles led to decreased measurement biases. Standardized interviewing, monitoring, and national telephone interviewing led to decreased correlated interviewer variance. But there still remain many biases that are generally not taken into account in current-day survey implementation.

The strategy of focusing on specific error sources to minimize the impact on the TSE has some inherent issues associated with it. First, rigorous controls are expensive and time-consuming, and additional control processes make most sense when the underlying survey process is under reasonable control to begin with. Second, the practice of investigating one error source at a time can be suboptimal. Some errors are more serious than others and this relative importance varies across surveys and uses. Third, all errors cannot be simultaneously minimized, since they are interrelated. For instance, in an attempt at reducing the nonresponse rate, we might induce an increased measurement error. Recent work on TSE has concentrated more on the simultaneous treatment of two or more error sources. For instance, West and Olson (2010) discuss whether or not some of the interviewer variance should really be attributable to nonresponse error variance. Also, Eckman and Kreuter (Chapter 5 in this volume) discuss interrelations between undercoverage and nonresponse. Fourth, in addition to survey errors and costs, Weisberg (2005) points out that sometimes errors cannot be minimized because correct design decisions are unknowable. For instance, asking question A before question B may affect the answers to question B, and asking question B before question A may affect the responses to question A. Therefore, it may be impossible to remove question order effects regardless of resources spent.

Thus, the approach aiming at reducing specific error sources is very important, but the error structures are more complicated than previously believed. Therefore, the inherent issues mentioned need to be addressed in more detail.

The second strategy toward the treatment of specific error sources uses evaluation studies as a means of quantifying the various sources of errors. Typically, evaluation studies are conducted

after the survey or census has concluded, and are a means of estimating the size of the total error or the error of the outcome of a specific survey operation, such as coding. Most well-known evaluation studies have been conducted in connection with U.S. censuses. A census is an ideal vehicle for studying survey processes and survey errors. The main methodology used is a comparison of the outcome of the regular survey or census compared to the outcome of a sample using preferred (but financially, methodologically, or administratively resource intensive) procedures or gold standard methodologies. Assuming that the gold standard is correct, the difference between the two is an estimate of the TSE, even though the difference is likely to either understate or overstate the true TSE. ASPIRE is a recent innovation used to evaluate TSE. It is an approach based on a mix of quality management ideas, as well as quantitative and qualitative assessments of the magnitude of TSE. This approach is further discussed in Section 1.4.

The evaluation programs conducted as part of the U.S. population censuses in 1940, 1950, and 1960 revealed important error and process problems, which led to significant procedural changes in future censuses and surveys. For instance, findings regarding the adverse effects of correlated variance induced by census enumerators as well as large costs associated with the enumeration process led to the decision to use more self-administration by mail in the census (U.S. Bureau of the Census, 1965). Currently, there is considerably diminished engagement in large evaluation studies mostly because of the enormous financial investments needed, but also because they are typically implemented long after they can be really helpful in any improvement work. For instance, the results of the evaluation of the coding operation in the 1970 U.S. Census were released in 1974 (U.S. Bureau of the Census, 1974b). Postenumeration surveys are still conducted in the U.S.A. to estimate the coverage error, and most importantly, how many people were missed in the census, since this so-called undercount can have a great impact on the distribution of funds to different regions in the country. In this case, the gold standard is a partial re-enumeration on a sample basis, where the estimation procedure resembles capture–recapture sampling (Cantwell et al., 2009).

1.3 Survey Models and Total Survey Design

During the period 1950–1970, much development was devoted to survey models aimed at providing expressions of the TSE as a combination of mean-squared-error (MSE) components. The U.S. Census Bureau survey model is perhaps the best known of these. In that model, the MSE of an estimate x, MSE(x), is decomposed into sampling variance, simple response variance, correlated response variance, an interaction term, and the squared bias. In some versions of the model, there is also a component reflecting the relevance, which is the difference between the survey's operational goal and its ideal goal. For instance, there is an operational definition of being employed used by official statistics agencies, which differs from an ideal definition that is more relevant but unattainable. The purpose of the survey model is to articulate the relative contribution to TSE from different components and to be able to estimate TSE more easily using components that can be added together. The model is described in numerous papers including Eckler and Hurwitz (1958), Hansen et al. (1961, 1964). The main issue with this model is its incompleteness in the sense that it does not reflect all the main error sources, most conspicuously, nonresponse and noncoverage. The model focuses solely on measurement errors and sampling errors. This is an obvious deficiency, specifically discussed by Cochran (1968). However, the model offers the opportunity to estimate errors beyond those induced by sampling and simple response variance. Although the above papers offer suggestions on how to estimate these components, Bailar and Dalenius (1969) provide a more comprehensive list of basic study schemes that could be used to estimate all components of error. The schemes use replication,

interpenetration, or combinations thereof. Some of these schemes are, however, rather elaborate and unrealistic. One scheme prescribes repeated reinterviews, which would be very difficult to implement given typical survey resource constraints. The estimation from these models of design effects due to variances associated with interviewers, crew leaders, supervisors, and coders has been particularly useful and has led to radical changes in census data collection procedures, as well as standardization and automation of other survey processes. Interviewer variance studies are relevant to many surveys, and more sophisticated schemes for its estimation are presented in Biemer and Stokes (1985).

The literature on survey models extends beyond that which comes from the U.S. Census Bureau. For instance, Fellegi (1964) introduced covariance components that Hansen and colleagues had assumed to be zero, including correlation of response deviations obtained by different enumerators (e.g., arising from specific features of training procedures) and correlation of sampling and response deviations *within enumerators* (e.g., the tendency for the same enumerator to induce different responses from elderly respondents than from young respondents).

Following Kish (1962), Hartley and Rao (1978) used mixed linear models to estimate nonsampling variance, and Lessler and Kalsbeek (1992) expanded the U.S. Census Bureau survey model by including also a component reflecting nonresponse. Bailar and Biemer (1984) made a similar attempt earlier but did not suggest specific estimates due to complexities relating to interaction terms.

In principle, the survey models and information about specific error sources can be used as inputs to survey designs. In this case, the aim is to develop a design such that the MSE is minimized given a fixed budget and any other constraints, assuming that all major sources of error are taken into account. A good design elucidates information about the relative impact of different error sources on the estimates, as well as the costs associated with reducing these effects. However, designs may vary for different survey estimates, and therefore, the use of the MSE should be considered as a planning criterion only. As Dalenius (1967) points out, "there is as yet no universally accepted survey design formula that provides a solution to the design problem and no formula is in sight." Such a formula would have to take into account activities such as pretesting, implementation of operations, controlling operations, and documenting results. A formula did not exist in 1967 and still does not exist today.

A design approach toward the *partial* treatment of TSE suggested by Dalenius (1967) and Hansen et al. (1967) contained a number of steps that included the following:

- Specifying the ideal survey goal, which would permit an assessment of the relevance component;
- Developing a small number of alternative designs based on a thorough analysis of the survey objectives and the general survey conditions;
- Evaluating design alternatives with a view to understanding their respective preliminary contributions to the key components of the MSE, as well as their associated costs;
- Choosing an alternative design or some modified version of a design—or deciding not to conduct the survey at all;
- Developing an "administrative design" including components such as feasibility testing, a process signal system (currently called "paradata"[2]), a design document, and a backup plan.

2 The **paradata** of a survey are data about the process by which the survey data were collected. The term first appears in Couper's presentation of his paper (Couper, 1998) although the term is not present in the actual paper. Examples of paradata topics include the times of day interviews were conducted, interview duration, how many times there were contacts with each interviewee or number of attempts to contact the interviewee, and the mode of communication (such as phone, the web, email, or face to face). This definition is easily extended to encompass other survey processes (Lyberg and Couper, 2005) such as editing and coding.

Another approach to the treatment of TSE was suggested by Leslie Kish during an interview (Frankel and King, 1996). Influenced by the strong Bayesian-focused cadre at the University of Michigan in the 1960s, Kish suggested that Bayesian models be used to quantify some of the error components. Kish drew on the contributions by researchers such as Ericson (1969) and Edwards et al. (1963) regarding the use of Bayesian methods in survey sampling and psychometrics. Kish suggested that judgment estimates of biases could be combined with sampling variances to achieve more realistic and less understated estimates of TSE. Kish did not rule out the possibilities of using nonprobability sampling and Bayesian modeling to shed light on certain survey phenomena. Dalenius was also open to what he called "neo-Bayesian" ideas in survey sampling, and one paper he wrote discussed the use of diffuse priors in sample surveys (Dalenius, 1974). He commissioned Lyberg to write a review paper on the use of neo-Bayesian ideas in surveys (Lyberg, 1973).

Although Dalenius (1967) held a concept of total survey design that encompassed all known error sources, and Kish (1965) contemplated Bayesian ideas as input to survey design, these ideas did not materialize into a methodology that could be fully used at the time. This is because the treatment of all sources of error held too many unknowns and because Bayesian modeling was considered very demanding from a computational point of view at the time. Therefore, the TSE perspective lost some of its attraction during a relatively long period (between 1975 and 2000), because the survey model approach proved to be complicated, its components were computationally intractable, and the models were incomplete. No agency really attempted to estimate TSE, with the exception of Mulry and Spencer (1993), who tried to estimate the total MSE of the 1990 U.S. Census. Instead survey organizations continued to work on methods that could reduce specific error sources as a consequence of a rapid development of new modes, combinations of modes, and methods for handling cognitive aspects of surveys. Near the end of the era of "disinterest," Forsman (1987) expressed disappointment with the small role that survey models had played in survey implementation to date. At roughly the same time, Biemer and Forsman (1992) showed that basic reinterview schemes did not work as intended, and Dillman (1996) was concerned about the lack of innovation within the U.S. Federal Statistical System with respect to addressing these issues. Finally, Platek and Särndal (2001) posed the following question: "Can a statistician deliver?" voicing their concern regarding the theoretical foundations of survey methodology, which included the topic of TSE. The Platek and Särndal article came to serve as a wake-up call for parts of the survey industry. A new workshop, the International Total Survey Error Workshop (ITSEW), convened its first meeting in 2005 and has, since 2008, met annually. The purpose of the workshop is to promote TSE thinking as well as to encourage studies that aim at joint investigations of more than one error source.

1.4 The Advent of More Systematic Approaches Toward Survey Quality

Around 1970, there was general agreement among prominent survey organizations that all main error sources ought to be taken into account when designing surveys. A few years earlier Hansen, Cochran, Hurwitz, and Dalenius had decided to write a textbook on total survey design, but the plan was abandoned due to the sudden demise of Hurwitz in 1968 (T. Dalenius, Personal communication with Lars Lyberg, 1968). Eventually, Groves (1989) wrote a seminal textbook along these lines.

One of the problems with the work on survey errors during that era was the absence of a process perspective and a consideration of continuous improvement. For instance, improvement

work was concentrated on measuring and decreasing errors, often without considering a process perspective. The user of statistics was a rather obscure player and even though there were user conferences (Dalenius, 1968), information about errors and problems flowed in one direction, namely from producer to user. Users were rarely asked to provide feedback to producers in this regard. Statisticians sometimes "role-played" as subject-matter specialists during the design phase of surveys, rather than engaging such specialists directly. Even though industrial process control had been used extensively at the U.S. Census Bureau and other places, no real process thinking was embedded in the strategies to reduce errors. Some consideration was given to process signal systems functioning as early warning systems, much in the same vein as paradata do today. However, continuous improvement of survey processes was not well developed, and when problems occurred, "rework" was a common remedy.

In roughly 1980, quality management and quality thinking become popular in organizations. Quality management developed as a science (Drucker, 1985) and quality systems such as total quality management (TQM) and Six Sigma entered the scene. Statistical organizations jumped on the bandwagon for two reasons.

First, there was pressure to recognize the user in more formalized ways, because of the acknowledgment that for statistics to be relevant they had to be used (Dalenius, 1985). Previously, attempts such as the "U.S. Standard for Presentation of Errors" (U.S. Bureau of the Census, 1974a) and error declarations in connection with the release of survey reports were quite technical and were developed without much contact with users. The era had arrived when the user was recognized as the customer or the representative of a paying customer, both who had the right to achieve value for money. The second reason for introducing quality management principles was cost. The production of statistics was expensive, and without process changes that resulted in more cost-effective outputs, competitors might take over.

There are several activities related to the principles of quality management, which became important in the production of statistics. Flow-charting of processes, plotting of paradata on control charts, and using cause-and-effect diagrams are examples of activities that became popular within the process improvement paradigm. There was an acknowledgment of a complementarity between survey quality and survey errors. It was recognized that accuracy could not be considered the sole indicator of survey quality, in the same way that the nonresponse rate cannot be considered the only indicator of accuracy of a survey. Dimensions other than relevance and accuracy were identified as important to users, most notably the dimensions of accessibility and timeliness, in an acknowledgment that accurate statistics might have limited utility if difficult to access or received too late. Considerable development was invested in a number of quality frameworks that articulated the various dimensions of quality. The first framework was produced by Statistics Sweden (Felme et al., 1976), and since then a number have followed. For instance, the Organisation for Economic Co-operation and Development's (OECD) 2011 framework has eight dimensions: relevance, accuracy, timeliness, credibility, accessibility, interpretability, coherence, and cost-efficiency. Eurostat, the statistical agency of the European Statistical System, has developed a Code of Practice that contains 15 different dimensions that relate to quality (Eurostat, 2011).

Statistical organizations have changed as a result of the global quality movement. Many organizations now use customer satisfaction surveys, process control via paradata (Kreuter, 2010), organizational quality assessment using excellence models such as Six Sigma (Breyfogle, 2003), quality improvement projects (Box et al., 2006), and current best methods (Morganstein and Marker, 1997). In 2008, Statistics New Zealand submitted a proposal for a new Generic Statistical Business Process Model (GSBPM) (Statistics New Zealand, 2008), which defines phases and subprocesses of the statistical lifecycle. The model has gradually been refined and its fifth version was released in 2013 (see Figure 1.1). The GSBPM is intended to apply to all activities

Quality management/metadata management							
Specify needs	Design	Build	Collect	Process	Analyze	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set-up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collections	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree on action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalize collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalize outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalize production system		5.7 Calculate aggregates			
				5.8 Finalize data files			

Figure 1.1 The generic statistical business process model. Source: Statistics New Zealand (2008).

undertaken by producers of official statistics. It can be used to describe and assess process quality independent of data sources used. A more complete description of the impact of quality management principles on survey organizations is given in Lyberg (2012).

Biemer (2010) formally defined the TSE paradigm as part of a larger design strategy that sought to optimize total survey quality (TSQ) and that included dimensions of quality beyond accuracy. The dimensions under consideration could be user-driven, and could be adopted from an official framework of the kind mentioned above or from any quality vector specified by the user. The basic elements of the TSQ paradigm include: design, implementation, evaluation, and the assessment of the effects of errors on the analysis. In the design phase, information on TSE is compiled, perhaps through quality profiles, which are documents containing all information that is known on the survey quality. From this, the major contributors to TSE are identified and resources are allocated to control these errors. During the implementation phase, processes for modifying the design are entertained as a means of achieving optimality. The evaluation part of the process allows for the routine embedding of experiments in ongoing surveys to obtain data that can inform future survey designs.

In relation to the first two pillars of the paradigm (design and implementation), a number of strategies have been developed that allow for design modification or adaptation during implementation to control costs and quality simultaneously. The activities in support of these strategies are conducted in real time and the strategies include continuous quality improvement, responsive design, Six Sigma, and adaptive total design.

The first strategy, "continuous quality improvement," is based on the continuous analysis (throughout implementation) of process variables, process metrics, or paradata that have been chosen because stable values of them are critical to quality. As a result of the analysis, specific interventions might be deemed necessary to ensure acceptable cost and quality.

A second strategy, called "responsive design" (Groves and Heeringa, 2006), was developed to reduce nonresponse bias. It is similar to continuous quality improvement but includes three phases: experimentation, data collection, and special methods to reduce nonresponse bias.

A third strategy is the use of the Six Sigma excellence model. It emphasizes decision making based on data analysis using a rich set of statistical methods and tools to control and improve processes. Six Sigma is an extreme version of continuous improvement.

A fourth and final strategy, called "adaptive total design and implementation," is a monitoring process which is adaptive in the sense that it combines features of the previous three strategies.

In all these strategies, the analysis of metrics is crucial. The theory and methods for industrial QC can be used (Montgomery, 2005) in the same way as they were during the U.S. Census Bureau operations in the 1960s. However, what differs is the treatment of different kinds of variations. Process variation used to be attributed solely to operators, for instance, while the current prevailing philosophy is that it is the underlying processes themselves that more often have to change.

The third pillar of the paradigm is the TSE evaluation. Such an evaluation can address any dimension of survey quality and is essential to long-term quality improvement. Examples include nonresponse bias studies and measurement bias studies. Of particular importance is the consideration of the joint effects of error sources and their interactions, rather than just single sources of error such as nonresponse.

The fourth pillar is the assessment of the effects of errors on the analysis. This is a neglected area but has been discussed in the literature by Biemer and Stokes (1991), Koch (1969), and Biemer and Trewin (1997). (See also Chapter 23 in this volume.) The effects of errors depend on the kind of parameter that is estimated and also on the specific use of the deliverables.

It was mentioned earlier that both users and producers of statistics alike have problems understanding the complexity of TSE and its components. Some types of errors are difficult to explain,

and therefore there is a tendency to emphasize errors and concepts that are easily understood, such as nonresponse. Furthermore, this lack of understanding is exacerbated by the fact that statistical agencies do not attempt to estimate TSE at all. However, recently the ASPIRE system (A System for Product Improvement, Review, and Evaluation) was developed at Statistics Sweden by Paul Biemer and Dennis Trewin in an attempt to assist management and data users in assessing quality in a way that can be easily understood. In this system, the MSE is decomposed into error sources. A number of somewhat subjective criteria on (among other things) risk awareness, compliance with best practice, and improvement plans are defined and quality rating guidelines are defined for each criterion. Rating and scoring rules are defined, and risk assessments as well as an evaluation process are performed. ASPIRE is described in Biemer et al. (2014) and has been successfully used for the 10 most critical products at Statistics Sweden; the quality of these products has improved over the four rounds conducted thus far.

Moving beyond the concept of TSQ, the concept of total research quality (TRQ) was introduced recently by Kennet and Shmueli (2014). The authors penned the term "InfoQ" to describe attempts at assessing the utility of a particular data set for achieving a given analysis goal by employing statistical analysis or data mining.

1.5 What the Future Will Bring

The survey landscape is currently transforming quickly. Because traditional surveys are costly and time-consuming, they are being replaced or complemented by other types of information sources.

"Opt-in online panels" based on nonprobability sampling methods borrowed from the presampling era are used to create representative miniature populations and have become quite common, especially in marketing and polling firms. The panels consist of individuals who have been recruited by banners on a website, or by email—and who have provided their email addresses to the implementing firm. Double opt-in online panels means that the recruited individuals receive a response from the firm and are asked to confirm their willingness to participate as well as to provide their email address and other personal information. Sometimes those who join receive an incentive. There is even an ISO (2009) standard for using and maintaining such panels, sometimes called "access panels," but as of the present, there is no theory to back the use of such panels. However, it is not uncommon to find that the results based on these panels produce outcomes that are quite similar to those using probability sampling (AAPOR, 2010; Wang et al., 2015), although it is often impossible to disentangle the magnitude of the differences. Online panels based on opt-in and double opt-in are likely here to stay, but data quality issues in relation to these have yet to be resolved. The use of Bayesian modeling (Gelman et al., 2014) is a possible route to explore, as well as the sensible adjustments of nonprobability samples using multilevel regression and poststratification, as demonstrated by Wang et al. (2015) in election predictions.

Some research fields use survey procedures without adopting a TSE perspective. Big Data allow for the harvesting and analysis of sensor data, transaction data, and data from social media. As shown in the recent AAPOR (2015) task force on Big Data and in Chapter 3 in this volume, it is possible to develop a TSE framework for Big Data. Hard-to-sample populations and international comparative surveys are other examples of survey areas that have their own research traditions (Chapter 9 in this volume; Tourangeau, 2014) that could benefit from a TSE perspective, and such work is underway. The use of administrative data also needs its own TSE framework (Wallgren and Wallgren, 2007). Even data disclosure limitation can be viewed from a TSE perspective (Chapter 4 in this volume).

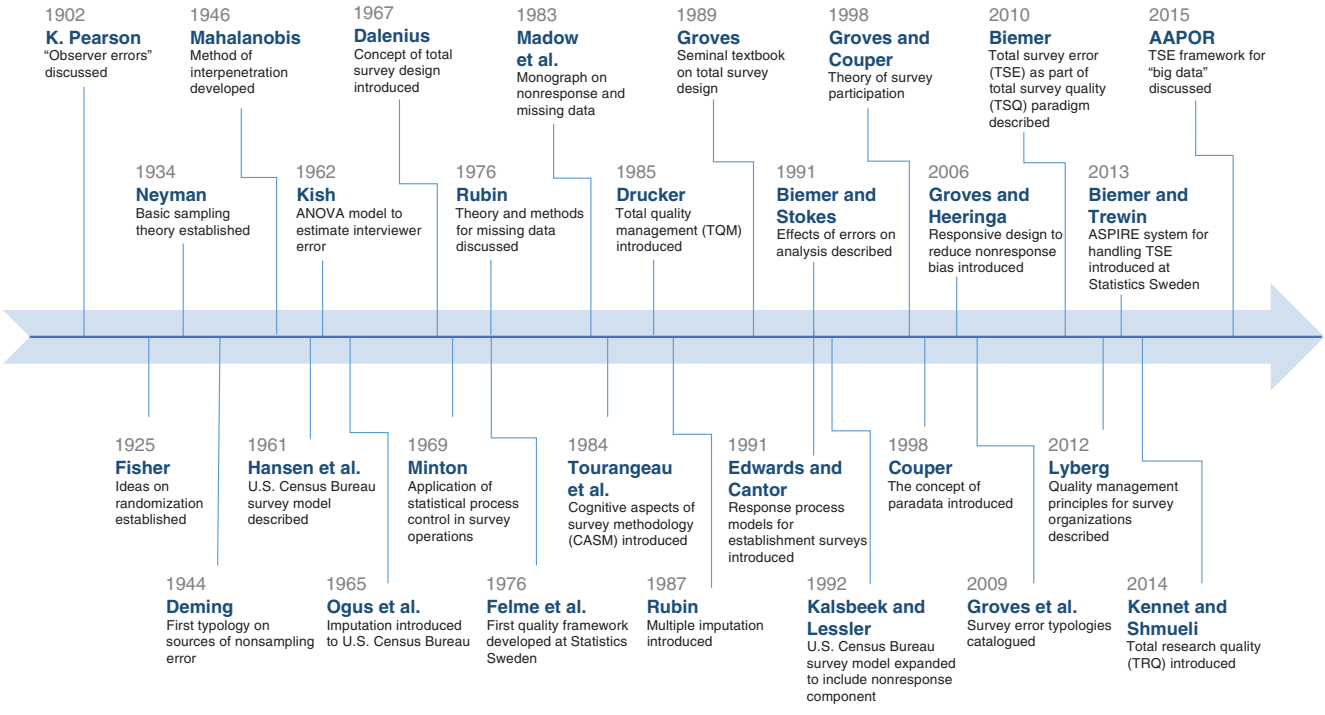


Figure 1.2 Subjective sample of events in the evolution of the concept of TSE.

It is heartening to see that quality issues have resurfaced as an area of interest for survey methodologists and data users alike. Recently, media outlets, who are important users of data, have developed publication guidelines including criteria on response rate, question wording, sampling method, and sponsorship. *The New York Times*, *The Washington Post*, and *Radio Sweden* are examples of such outlets. This is part of a greater trend toward data-driven journalism that is based on analyzing and filtering large data sets for the purpose of creating news stories based on high-quality data.

A new survey world that uses multiple data sources, multiple modes, and multiple frames is at our disposal, and it is essential that quality considerations keep pace with such developments to the extent possible. Indeed, promoting and defending ideas on data quality and sources of error is an important, albeit daunting task.

In closing, Figure 1.2 provides the authors' subjective summary timeline of some of the most important developments in TSE research from 1902 to present day.

References

AAPOR (2010). Online panel task force report. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOROnlinePanelsTFReportFinalRevised1.pdf (accessed July 15, 2016).

AAPOR (2015). Big data task force report. https://www.aapor.org/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf (accessed July 15, 2016).

Bailar, B. and Biemer, P. (1984). Some methods for evaluating nonsampling error in household censuses and surveys. In P.S.R.S. Rao and J. Sedransk (eds) *W.G. Cochran's impact on statistics*, 253–274. New York: John Wiley & Sons, Inc.

Bailar, B. and Dalenius, T. (1969). Estimating the response variance components of the U.S. Bureau of the Census' survey model. *Sankhya, Series B*, 31, 341–360.

Belson, W.A. (1968). Respondent understanding of survey questions. *Polls*, 3, 1–13.

Biemer, P. (2010). Overview of design issues: Total survey error. In P. Marsden and J. Wright (eds) *Handbook of survey research*, Second edition. Bingley: Emerald Group Publishing Limited.

Biemer, P. and Fecso, R. (1995). Evaluating and controlling measurement error in business surveys. In B. Cox, D. Binder, B.N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds) *Business survey methods*, 257–281. New York: John Wiley & Sons, Inc.

Biemer, P. and Forsman, G. (1992). On the quality of reinterview data with applications to the Current Population Survey. *Journal of the American Statistical Association*, 87, 420, 915–923.

Biemer, P. and Lyberg, L. (2003). *Introduction to survey quality*. New York: John Wiley & Sons, Inc.

Biemer, P. and Stokes, L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 158–166.

Biemer, P. and Stokes, L. (1991). Approaches to the modeling of measurement error. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds) *Measurement error in surveys*, 487–516. New York: John Wiley & Sons, Inc.

Biemer, P. and Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds) *Survey measurement and process quality*, 603–632. New York: John Wiley & Sons, Inc.

Biemer, P., Trewin, D., Bergdahl, H., and Japec, L. (2014). A system for managing the quality of official statistics. *Journal of Official Statistics*, 30, 3, 381–415.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv. 1, 6–62.

Box, G. and Friends (2006). *Improving almost anything: Ideas and essays*. Hoboken: John Wiley & Sons, Inc.

Breyfogle, F. (2003). *Implementing six sigma*, Second edition. New York: John Wiley & Sons, Inc.
Brick, M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 3, 329–353.
Cantwell, P., Ramos, M., and Kostanich, D. (2009). Measuring coverage in the 2010 U.S. Census. *American Statistical Association, Proceedings of the Social Statistics Section*, Alexandria, VA, 43–54.
Cochran, W. (1968). Errors of measurement in statistics. *Technometrics*, 10, 637–666.
Couper, M. (1998). Measuring survey quality in a CASIC environment. Paper presented at the Joint Statistical Meetings, American Statistical Association, Dallas, TX, August 9–13.
Dalenius, T. (1961). Treatment of the non-response problem. *Journal of Advertising Research*, 1, 1–7.
Dalenius, T. (1967). Nonsampling errors in census and sample surveys. Report no. 5 in the research project Errors in Surveys. Stockholm University.
Dalenius, T. (1968). Official statistics and their uses. *Review of the International Statistical Institute*, 26, 2, 121–140.
Dalenius, T. (1974). Ends and means of total survey design. Report from the research project Errors in Surveys. Stockholm University.
Dalenius, T. (1985). Relevant official statistics. *Journal of Official Statistics*, 1, 1, 21–33.
De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. Hoboken: John Wiley & Sons, Inc.
Deming, E. (1944). On errors in surveys. *American Sociological Review*, 9, 359–369.
Deming, E. (1986). *Out of the crisis*. Cambridge: MIT.
Deming, E., Tepping, B., and Geoffrey, L. (1942). Errors in card punching. *Journal of the American Statistical Association*, 37, 4, 525–536.
Dillman, D. (1996). Why innovation is difficult in government surveys. *Journal of Official Statistics*, 12, 2, 113–198 (with discussions).
Drucker, P. (1985). *Management*. New York: Harper Colophone.
Eckler, A.R. and Hurwitz, W.N. (1958). Response variance and biases in censuses and surveys. *Bulletin of the International Statistical Institute*, 36, 2, 12–35.
Edwards, S. and Cantor, D. (1991). Toward a response model in establishment surveys. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds) *Measurement errors in surveys*, 211–233. New York: John Wiley & Sons, Inc.
Edwards, W., Lindman, H., and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
Ericson, W. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 2, 195–233.
Eurostat (2011). *European statistics Code of Practice*. Luxembourg: Eurostat.
Fasteau, H., Ingram, J., and Minton, G. (1964). Control of quality of coding in the 1960 censuses. *Journal of the American Statistical Association*, 59, 305, 120–132.
Fellegi, I. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016–1041.
Felme, S., Lyberg, L., and Olsson, L. (1976). *Kvalitetsskydd av data. (Protecting Data Quality.)* Stockholm: Liber (in Swedish).
Fienberg, S.E. and Tanur, J.M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review*, 64, 237–253.
Fienberg, S.E. and Tanur, J.M. (2001). History of sample surveys. In N.J. Smelser and P.B. Baltes (eds) *International encyclopedia of social and behavioral sciences*, Volume 20, 13453–13458. Amsterdam/New York: Elsevier Sciences.
Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Forsman, G. (1987). Early survey models and their use in survey quality work. *Journal of Official Statistics*, 5, 41–55.

Frankel, M. and King, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 1, 65–87.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian data analysis*. Boca Raton: Chapman and Hall.

Groves, R. (1989). *Survey errors and survey costs*. New York: John Wiley & Sons, Inc.

Groves, R.M. and Couper, M.P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons, Inc.

Groves, R. and Heeringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 439–457.

Groves, R. and Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly*, 74, 5, 849–879.

Groves, R. and Peychteva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72, 2, 167–189.

Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds) (2002). *Survey nonresponse*. Hoboken: John Wiley & Sons, Inc.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2009). *Survey methodology*, Second edition. Hoboken: John Wiley & Sons, Inc.

Hacking, I. (1975). *The emergence of probability*. London/New York: Cambridge University Press.

Hansen, M. and Hurwitz, W. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517–529.

Hansen, M. and Steinberg, J. (1956). Control of errors in surveys. *Biometrics*, 12, 462–474.

Hansen, M., Hurwitz, W., and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 32nd Session, 38, Part 2, 359–374.

Hansen, M., Fasteau, H., Ingram, J., and Minton, G. (1962). Quality control in the 1960 Censuses. *American Society for Quality Control. Proceedings of the Middle Atlantic Conference*. Milwaukee, WI, 323–339.

Hansen, M., Hurwitz, W., and Pritzker, L. (1964). The estimation and interpretation of gross differences and simple response variance. In C.R. Rao (ed.) *Contributions to Statistics*, 111–136. Oxford: Pergamon Press.

Hansen, M., Hurwitz, W., and Pritzker, L. (1967). Standardization of procedures for the evaluation of data: Measurement errors and statistical standards in the Bureau of the Census. Paper presented at the 36th session of the International Statistical Institute, Sydney, Australia, August 28 to September 7.

Hartley, H.O. and Rao, J. (1978). Estimation of nonsampling variance components in sample surveys. In N. Namboodiri (ed.) *Survey sampling and measurement*, 35–43. New York: Academic Press.

ISO (2009). Access panels in market, opinion and social research. Standard 26362. International Organization for Standardization.

Jabine, T., Straf, M., Tanur, J., and Tourangeau, R. (1984). *Cognitive aspects of survey methodology: Building a bridge between disciplines. Report of the advanced research seminar on cognitive aspects of survey methodology*. Washington, DC: National Academy of Sciences Press.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1–16.

Kennet, R. and Shmueli, G. (2014). On information quality. *Journal of the Royal Statistical Society, Series A*, 177, 1, 3–38.

Kiaer, A. (1897). The representative method of statistical surveys. *Kristiania Videnskapsselskabets Skrifter, Historik-filosofiske Klasse*, 4, 37–56 (in Norwegian).

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 297, 92–115.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons, Inc.
Koch, G. (1969). The effect of nonsampling errors on measures of association in 2 x 2 contingency tables. *Journal of the American Statistical Association*, 64, 852–853.
Kreuter, F. (ed.) (2010). *Improving surveys with paradata*. Hoboken: John Wiley & Sons, Inc.
Lessler, J. and Kalsbeek, W. (1992). *Nonsampling error in surveys*. New York: John Wiley & Sons, Inc.
Lyberg, L. (1973). The use of neo-Bayesian ideas in survey sampling. The research project Errors in Surveys, Report no. 66. Stockholm University (in Swedish).
Lyberg, L. (2012). Survey quality. *Survey Methodology*, 2, 107–130.
Lyberg, L. and Couper, M. (2005). The use of paradata in survey research. Invited paper, International Statistical Institute, Sydney, Australia, April 5–12.
Madow, W.G., Nisselson, H., and Olkin, I. (eds) (1983). *Incomplete data in sample surveys*, Volumes 1–3. New York: Academic Press.
Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325–378.
Minton, G. (1970). Some decision rules for administrative applications of quality control. *Journal of Quality Technology*, 2, 2, 86–98.
Minton, G. (1972). Verification error in single sampling inspection plans for processing survey data. *Journal of the American Statistical Association*, 67, 337, 46–54.
Montgomery, D. (2005). *Introduction to statistical quality control*, Fifth edition. New York: John Wiley & Sons, Inc.
Morganstein, D. and Marker, D. (1997). Continuous quality improvement in statistical agencies. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds) *Survey measurement and process quality*, 475–500. New York: John Wiley & Sons, Inc.
Mulry, M.H. and Spencer, B.D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of the American Statistical Association*, 88, 1080–1091.
Neter, J. and Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 305, 18–55.
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, 558–606.
OECD (2011). *Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities*. Paris: OECD.
Ogus, J., Pritzker, L., and Hansen, M.H. (1965). Computer editing methods-some applications and results. *Bulletin of the International Statistical Institute*, 35, 442–466.
O'Muircheartaigh, C. and Marckward, A.M. (1980). An assessment of the reliability of World Fertility Study data. *Proceedings of the World Fertility Survey Conference*, 3, 305–379. International Statistical Institute, The Hague, the Netherlands.
Pearson, K. (1902). On the mathematical theory of errors of judgment. *Philosophical Transactions of the Royal Society, London, Series A*, 198, 235–299.
Platek, R. and Särndal, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, 17, 1, 1–20 and Discussion, 21–127.
Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 117–138.
Rao, J.N.K. (2013). Impact of sample surveys on social sciences. Paper presented at the Catholic University of the Sacred Heart, Piacenza, Italy, March 13.
Rao, J.N.K. and Fuller, W. (2015). Sample survey theory and methods: Past, present and future directions. Invited paper presented at the ISI meetings in Rio de Janeiro, International Statistical Institute, July 26–31.
Rice, S.A. (1929). Contagious bias in the interview. *American Journal of Sociology*, 35, 420–423.

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
Smith, T.W. (2011). Refining the total survey error perspective. *International Journal of Public Opinion Research*, 23, 4, 464–484.
Statistics New Zealand (2008). Proposal for a new generic statistical business process model. Paper presented at the joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Luxembourg, April 9–11.
Stoop, I., Billiet, J., Koch, A., and Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. Chichester: John Wiley & Sons, Ltd.
Sudman, S. and Bradburn, N. (1974). *Response effects in surveys*. Chicago: Aldine.
Tourangeau, R. (2014). Defining hard-to-survey populations. In R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, and N. Bates (eds) *Hard-to-survey populations*, 3–20. New York: Cambridge University Press.
Tourangeau, R., Rips, L.J., and Rasinski, K.A. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
Tschuprow, A. (1923a). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (Chapters I–III). *Metron*, 2, 461–493.
Tschuprow, A. (1923b). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (Chapters IV–VI). *Metron*, 2, 646–680.
U.S. Bureau of the Census (1965). *Quality control of preparatory operations, microfilming, and coding*. Washington, DC: U.S. Government Printing Office.
U.S. Bureau of the Census (1974a). *Standards for discussion and presentation of errors in data*. Washington, DC: U.S. Department of Commerce, Bureau of the Census.
U.S. Bureau of the Census (1974b). *Coding performance in the 1970 census, evaluation and research program PHC(E)-8*. Washington, DC: U.S. Government Printing Office.
Wallgren, A. and Wallgren, B. (2007). *Register-based statistics. Administrative data for statistical purposes*. Hoboken: John Wiley & Sons, Inc.
Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31, 3, 980–991.
Weisberg, H. (2005). *The total survey error approach*. Chicago: The University of Chicago Press.
West, B. and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 5, 1004–1026.
Willimack, D.K. and Nichols, E. (2010). A hybrid response process model for business surveys. *Journal of Official Statistics*, 26, 3–24.
Zizek, F. (1921). *Grundriß der Statistik*. München/Leipzig: Duncker & Humblot (in German).

2

Total Twitter Error

Decomposing Public Opinion Measurement on Twitter from a Total Survey Error Perspective

Yuli Patrick Hsieh and Joe Murphy

Survey Research Division, RTI International, Chicago, IL, USA

2.1 Introduction

2.1.1 Social Media: A Potential Alternative to Surveys?

Social scientists investigating public opinion trends typically begin their research by seeking national estimates from representative surveys, such as the General Social Survey (GSS), or those conducted in the U.S.A. by the Pew Research Center for the People and the Press and Gallup. These reputable sources are useful if a national estimate and broad trends are sufficient for their analyses. However, surveys are limited in their ability to produce very timely and rapid estimates in response to current events, since large-scale survey data collection is very time- and resource-intensive. Additional challenges have arisen in recent years that have made it more costly and difficult to obtain accurate survey estimates (i.e., the erosion of landline telephone coverage and declining response rates). Such limitations may also indicate an emerging need to look for alternative methods to study public opinion.

As new information and communication technologies (ICTs) like mobile phones and social media become widely adopted and deeply integrated into contemporary daily routines, they are changing the nature of the public sphere—many users share thoughts and information to express their attitudes and opinions about ongoing events spontaneously, instantaneously, and often publicly across services and platforms. Consequently, such information expressed in online social spaces provides researchers potential alternative resources and data for studying public opinion. For example, a researcher can access a repository of posts made on Twitter, define search terms to retrieve relevant tweets, and then interpret what those tweets reveal about public sentiment on a given topic. Because these data are produced "organically" as opposed to the "designed" nature of surveys (Groves and Lyberg, 2010), they are available on a much more frequent basis. These data can also be relatively inexpensive to retrieve and the sheer volume of these "big data" can provide an enticing potential alternative source of data on attitudes and opinions.

However, social media data also come with their own limitations. On the surface, the lack of representation is already a well-known criticism. For instance, only 23% of online American

Total Survey Error in Practice, First Edition. Edited by Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, and Brady T. West.

adults use Twitter, one of the most commonly accessed social media platforms for research (Duggan et al., 2015).[1] Beyond the obvious concerns with coverage, there are many additional limitations that manifest themselves as one delves deeper into the process of analyzing and making sense of social media postings. When evaluating survey statistics, researchers benefit from frameworks such as the total survey error (TSE) (e.g., Biemer, 2010; Groves and Lyberg, 2010) as theoretical underpinnings to identify and estimate potential errors while constructing statistical measures of public opinion from survey data. However, researchers currently do not have a systematic error framework to guide the quality assessment of social media data.

2.1.2 TSE as a Launching Point for Evaluating Social Media Error

The TSE framework presents a structural approach to the procedural and statistical errors of survey estimates with the goal of ensuring the data quality for subsequent analysis and inferences. It theorizes the properties of different types of errors and develops statistical techniques to estimate the magnitude of such errors (Biemer and Lyberg, 2003; Groves et al., 2004). Being able to account for the errors stemming from the survey process makes the aforementioned reputable survey statistics accountable estimates of public opinion. So, when it comes to studying public opinion on social media, a natural question becomes whether and to what extent the TSE or a similar framework can be used to conceptualize and discern the error sources of nonprobabilistic, organically generated, and passively collected social media data.

To accept this premise, we must also accept that "error" is an appropriate concept in the realm of social media analysis. That is, we must believe that there is some true value—whether it can be known or not—that the analysis of social media strives to measure. For instance, one may ask "In December 2012, what proportion of individuals in Colorado were in favor of marijuana legalization?" Twitter may be an appropriate data source to examine this question *if* a representative proportion of the population of Colorado was on Twitter, tweeting about their opinions on this topic, in a truthful manner, and at a consistent rate. As we will see later in this chapter, these assumptions may be very unrealistic, and the traditional approach of formulating an unchanging research question prior to starting analysis is ill-suited to the content and infrastructure of Twitter. Regardless, another benefit of exploring the applicability of TSE to social media analysis of public opinion is the potential to evaluate estimates from surveys and social media in common terms. A common error framework would provide a valuable basis for the comparative quality of research based on each method.

Through the lens of TSE, we seek to conceptualize the errors that can result from the common practice of social media data extraction and analysis, identifying the trade-offs between data and errors across queries. In completing this exercise, we have arrived at a general error framework for Twitter opinion research comprising three broad and interrelated but exhaustive and mutually exclusive error sources: *coverage error*, *query error*, and *interpretation error*. **Coverage error** concerns various sources of over- and under-coverage of both Twitter users and posts regardless of the unit of analysis. It is the difference between the target population and units available for analysis on Twitter. In the data extraction process, **query error** occurs when a researcher misspecifies the search queries to extract the proper data for analysis. For example, if researchers just include "pot" in their query to extract tweets about marijuana use, then they are likely to miss many relevant tweets containing the terms "marijuana," "weed," and so forth. They are also likely to obtain off topic tweets about gardening and cooking. **Interpretation error** arises after the tweets are extracted when a researcher uses human or machine methods to infer (i) the

1 See the Pew Research Center's Social Networking Fact Sheet (http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/) for the latest updates on social media usage statistics.

sentiment of the extracted data or (ii) missing information about users' characteristics. This error can be defined as the extent to which the true meaning or value differs from that determined by the researcher.

In the remainder of the chapter, we provide an overview of the literature describing how the architecture and the user-generated content of Twitter may reflect public opinion. Next, we discuss the sources of coverage, query, and interpretation errors associated with Twitter data in more detail, relating, where possible, to the similar TSE concepts. To further demonstrate our error framework, we provide examples of these error sources by walking through a common method of accessing, querying, and interpreting Twitter data for marijuana legalization and abortion rights. We focused on these issues given the availability of opinion estimates from nationally representative surveys, allowing us to compare survey estimates to those from Twitter. We include in these examples our rationales and process of topical keyword selection and search query specification, and compare the extraction results between various queries for both topics, linking the findings to the error components of our framework. Last, we discuss the implications of this research, limitations for comparing surveys and social media in common terms, and suggestions for future research in this area.

2.2 Social Media: An Evolving Online Public Sphere

The proliferation of social network sites, or more generally social media, is one of the most significant cultural phenomena of the new millennium (for a detailed review of the history and definition of social media, see boyd (2007), boyd and Ellison (2007), and Ellison and boyd (2013)). Social media users primarily seek to stay connected with their social circles by publicly sharing *status updates*—information about their current thoughts and behaviors in their everyday life. The streams of content distributed across users' social networks become the center of organization on social media (Ellison and boyd, 2013), whether in the form of trending posts on the social news site reddit and Facebook's newsfeed, or as the landing page on Instagram, Tumblr, and Twitter.

2.2.1 Nature, Norms, and Usage Behaviors of Twitter

As of 2015, Twitter is one of the most popular social network sites. Twitter's architecture affords flexibility and brevity in expression. It allows users to post a message, or a "tweet" with a maximum length of 140 characters. Twitter's service requires users to create a username (i.e., handle) and invites users to create a profile containing a brief introductory description, name, and location information with the options of uploading photos for the account header and the profile headshots. Other demographic information such as gender, education, income, and race are not collected nor stored in the Twitter metadata. The default setting of Twitter's service is to make profiles and user-generated content public unless users explicitly change their privacy settings. Users can find and follow any other users with a public profile on Twitter without reciprocation. Although content is expected to be publicly accessible to encourage interaction, Twitter does not require users to submit real personal information to their profiles, allowing people to maintain their privacy and control their self-disclosure with creativity (i.e., listing locations like "Here" or "Hogwarts").

Twitter's default settings facilitate a unique social environment that enables the construction of sparsely knit networks suitable for self-disclosure and information dissemination (boyd et al., 2010; Kwak et al., 2010; Naaman et al., 2010; Walton and Rice, 2013). This particular design

allows users to connect with celebrities, public figures, or other personal contacts beyond users' social circles to exchange information. At the same time, it also preserves conversation opportunities between users when they reciprocate the "following" connection for further engagement.

When posting on Twitter, users tend to employ shorthand, symbols, and emoticons to share their updates. Some of these practices have become cultural norms and site functionalities. The usage of the "at" sign (@) in combination with a username is a syntax to address the tweet to a specific user, whereas using a hashtag (#) followed by a topical keyword will classify the tweet's topic and associate the message with all other tweets using the same identifying hashtag. Additionally, tweets are expected to be shared and rebroadcasted (boyd et al., 2010). Users often "retweet" (RT) the posts shared by others to their own followers. The retweeting practice encourages fast-paced information sharing in an online public sphere with some degree of privacy and anonymity.

While Twitter's service was originally designed for sharing a "short burst of inconsequential information" online (Sarno, 2009), it has profound implications for social interaction and civic engagement. Social media like Twitter comprises more than just individual users; corporations, government agencies, news media, nonprofit organizations, celebrities, and public figures are using Twitter for disseminating information, news or personal anecdotes; promoting products and services; or organizing and raising money for causes. A growing body of research has shown that Twitter has played a critically enabling role as an alternative news circulating and resource mobilizing venue during the Arab Spring and other political protests and social movements (Chaudhry, 2014; Gleason, 2013; González-Bailón et al., 2013; Lotan et al., 2011; Papacharissi and de Fatima Oliveira, 2012; Thorson et al., 2013; Tufekci, 2013; Wilson and Dunn, 2011). Conversely, some social media user accounts are even set up as nonhuman "spamming bots"—malicious programs automatically generating and spreading a massive amount of fraudulent or useless information for the purpose of mischief.

2.2.2 Research on Public Opinion on Twitter

Early scholarship exploring public opinion on Twitter mainly focused on tracking conversational trends about newsworthy events. For example, Twitter data have been used to detect breaking news and disease outbreaks (Achrekar et al., 2011; Bandari et al., 2012; Ciulla et al., 2012; Hu et al., 2012; Lanagan and Smeaton, 2011; Petrovic et al., 2013; Sakaki et al., 2010). The promising predictive accuracy reported in these studies seems to suggest that the various measures constructed from tweet volume and content sentiment can be indicative of breaking news developments and the actual outcomes of social events.

However, research predicting political behaviors such as election results or voting intentions via Twitter has produced mixed outcomes. Some studies have put forth methods that produce metrics highly and positively correlated to the election results and public opinion estimates (such as popularity or approval of candidates) gathered from traditional surveys (Ceron et al., 2014; O'Connor et al., 2010; Skoric et al., 2012; Tumasjan et al., 2010). Conversely, other researchers (Chung and Mustafaraj, 2011; Gayo-Avello, 2011, 2013; Jungherr et al., 2012) either found inconsistent patterns from the same data or were not able to replicate the success under different contexts using the same method (e.g., O'Connor et al., 2010; Tumasjan et al., 2010). The controversy cautioned that the correlation between the election results and the measures constructed from Twitter data may vary depending on the research design decisions ranging from data collection time frame to keyword extraction parameters.

Seeking to enable comparability across studies of opinion behaviors on Twitter, Bruns and Stieglitz (2013) proposed a set of Twitter data metrics describing the general patterns of user

activity and visibility along with the temporal changes in tweet volume. This approach can be informative in situations where a handful of power (or opinionated) users disproportionally generate a great share of the content while most users may tweet only once about the event of interest. However, this approach mainly addresses the measurement *within* Twitter. It does not directly address the broader coverage issues of the Twitter user base or the errors that stem from the data extraction process and subsequent analyses.

2.3 Components of Twitter Error

Researchers typically extract data from Twitter using keyword queries. Results are returned at the tweet level (i.e., there may be more than one tweet per user), but analysis and interpretation might occur at the tweet level, the subtweet level (e.g., count of positive or negative words per tweet), or by treating the extracted dataset as a single "corpus" of tweets (Schober et al., 2016). The typical workflow for a Twitter content analysis begins with identifying a time frame, geography, and languages of interest. Next, the researcher identifies a set of topical keywords relevant to the research inquiry and develops the search query to extract the proper data through multiple iterations. At this stage, the goal of query specification is to maximize *topic coverage* rather than *population coverage* (Schober et al., 2016). Then, the researcher selects an automated text analysis or other machine learning technique to determine the meaning of the tweets as proxy measures of public opinion. Sometimes, this involves a human review of a subset of tweets to serve as a "gold standard" for training the machine algorithm. At this stage, the goal of analysis is to achieve high predictive accuracy. Sometimes, the researcher also attempts to classify and discern demographic and geographic information about the authors of the tweets when this information is not included in standard Twitter metadata (Murphy et al., 2014). We argue that three major classes of errors, with multiple subtypes, are likely to occur during the data extraction and analysis process: coverage, query, and interpretation errors. Figure 2.1 and Table 2.1 provide readers a graphic representation and a detailed breakdown of our error framework.

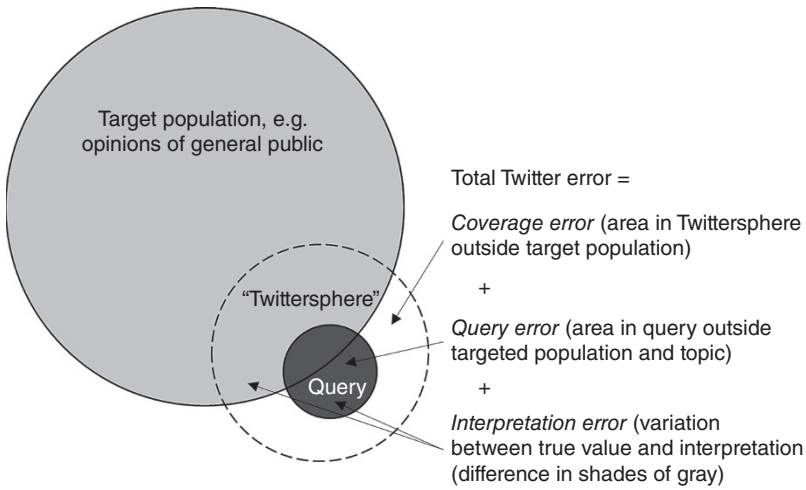


Figure 2.1 Theoretical spaces of Twitter data error.

Table 2.1 Characteristics of Twitter data error.

	Coverage error	Query error	Interpretation error
Abbreviated definition	Deviation from coverage of research population	Variation in scope of research topics of extracted tweets	Variation in inferring meaning and user information from extracted tweets
Origin	Differences between target population and the Twitter user base	Mis-specification of search queries	Selection and use of predictive modeling techniques and parameters
Examples	Mismatch between U.S. adult population and U.S. adults on Twitter; over- or under-coverage due to incorrect geography	Inappropriate inclusion or exclusion of RTs; irrelevant or missing keywords	Human error in determining positive vs. negative sentiments in tweets; machine algorithm incorrectly predicting sentiment
Related TSE components	Coverage error	Coverage error	Measurement error
		Measurement error	Modeling error
			Classification error

2.3.1 Coverage Error

The lack of general population coverage of Twitter users is well acknowledged (Graham et al., 2014; Mislove et al., 2011). Twitter users tend to be younger on average than the general population. They are also more likely to be of black non-Hispanic race/ethnicity and residing in urban areas (Duggan et al., 2015). Therefore, Twitter data suffer from **undercoverage** for the purpose of gauging representative public opinion. This undercoverage is represented in Figure 2.1 by the portion of the large target population circle that does not intersect with the "Twittersphere." Also, as described earlier, Twitter comprises both individual and "nonindividual" users on Twitter. In the context of analyzing public opinion, the "noise" produced by these nonindividuals is a form of **overcoverage**. To further assess coverage, some researchers attempt to impute the missing demographic and largely missing geographic metadata for tweets. Such a practice may allow for more detailed examination of coverage but may also introduce interpretation error. See Section 2.3.3 for more details.

2.3.2 Query Error

Another critical error source that stems from the collection of Twitter data for analysis is **query error**. This occurs when the query, or keyword search, does not provide results that well represent the topic under investigation (see Figure 2.1). Public attention to current newsworthy events on social media is highly contingent on the nature and social contexts of such events. Research has shown that the rhythm, volume, and meaning of tweets can vary significantly by events or cultures (Metaxas and Mustafaraj, 2012; Skoric et al., 2012). Not all Twitter users share information or post their opinions about newsworthy events at the same rate during a given time period. Additionally, the ways users engage in the Twittersphere may switch between information sharing and interpersonal communication over the duration of the attention span, and thus alter the trending dynamics and expressive sentiment of tweets (Jackoway et al., 2011;

Lin et al., 2014). However, such a source of potential bias due to exclusion from the extracted data has been mostly ignored in the literature (Gayo-Avello, 2013).

Consequently, unlike traditional survey estimates, analytical findings from Twitter data regarding public opinion are very sensitive to all parameters of research design. First, query error, whereby irrelevant posts are included and relevant messages are excluded due to the choices of keywords used in queries, may almost certainly occur during the data extraction when the search queries are mis-specified. The query error is similar to the "error of selectivity" in the cognitive process of response formation during surveys (Edwards and Cantor, 2004, pp. 218–219). Such a definition is also in line with the concepts of precision and recall—the quality measures of information retrieval used to evaluate the quality of search query in the computer and information sciences (Murphy et al., 2014; van Rijsbergen, 1979). *Precision* refers to the proportion of the retrieved outcomes that are relevant to the intended target of the search query, whereas *recall* denotes the proportion of relevant records that are obtained by the search query. As an example, the specification of a search query when using a search engine like Google is often a balancing act between precision and recall: a query with specific search terms may retrieve results with higher precision and lower recall than a less specific query. The metrics of precision and recall are useful tools for conceptualizing the query error, allowing researchers to evaluate the sensitivity and specificity of the queries and assess the quality of the extracted data. We raise awareness of these metrics here for context in discussing the trade-offs and considerations in attempting to minimize query error.

The definition and magnitude of the query error may vary dramatically when using different sets of keywords and time frames. Determining the relevance of extracted items is a subjective and iterative process as it requires multiple attempts to determine which query may achieve the least query error. Each query will result in different total retrieved outcomes and varying amount of relevant records based on separate sets of keywords. Therefore, the precision and recall may not be the most practical estimates of the query error since the true value of the denominator of these estimates may be unknown and the estimates may be incomparable between different queries. Researchers still need to make a subjective judgment about the best dataset with the most relevant records.

Assessing the query error of Twitter data is extremely difficult even when there are some existing survey estimates to serve as a baseline for comparison. The decision on extraction parameters and including or excluding just one keyword can dramatically affect the estimates of query error and the result of the substantive analysis (Jungherr et al., 2012; Tumasjan et al., 2010). Therefore, even if the precision and recall estimates are provided for different queries, they are constructed from essentially different sets of extracted data. It can be very problematic to compare the estimates at face value.

2.3.3 Interpretation Error

Once the Twitter data have been queried and extracted for analysis, the door opens to another source of error related to the interpretation of the content. Interpretation error may occur when the analyst infers a meaning from the Twitter content other than that intended by the tweeter. This can come in the form of human misinterpretation of the content or failure of the machine algorithm to appropriately assign sentiment or meaning to the data. It can also include error in inferring values for missing data, such as the interpretation of a user's or tweet's location based on the content contained in the tweet. Note that employing machine learning techniques still involves a subjective decision-making process similar to human coding. Researchers may decide to either (i) use a well-established "off-the-shelf" machine learning algorithm without changing

its parameters, (ii) alter an existing predictive algorithm to adjust the modeling in specific ways, or (iii) develop a new one to better fit the data at hand.

Relatedly, interpretation error may be introduced when researchers use machine learning techniques to address coverage error by filling in missing demographic and geographic information of Twitter data. Inferring user characteristics is subject to the interpretation of cultural and linguistic variation between people with different demographic traits (Graham et al., 2014; Hecht et al., 2011; Mislove et al., 2011), which also leads to interpretation error. For instance, even when the errors of the aforementioned data mining techniques can be limited to an ignorable margin, they may work well only with Western names for inferring users' gender and race and ethnicity (Mislove et al., 2011). In addition, using data mining techniques to identify spammers (e.g., Jindal and Liu, 2008) or Twitter accounts belonging to organizational entities can result in some degree of interpretation error since it is unrealistic to expect a full rate of detection accuracy.

2.3.4 The Deviation of Unstructured Data Errors from TSE

Unlike the probability sampling design applied in surveys, the bottom-up approach to the analysis of unstructured social media textual data considers these data as a form of *corpus* and seeks to achieve *topic coverage* rather than *population coverage* (Schober et al., 2016). The premise of this approach is that if the collected corpus can include as much of the opinion about a given topic as possible from a given population, then the analysis can reflect and inform the opinion landscape of the population regarding the topic. From the perspective of TSE, this analytic approach also poses a disconnection between the unit of analysis and the unit of data collection for conducting research, given that the unstructured textual data are collected primarily by entry (i.e., posts by individual users) but are analyzed in an aggregated form. In other words, the survey errors are theorized and estimated based on the same unit of analysis (i.e., survey participants) for sample construction, sampling, data collection, and analysis. However, our social media data error framework is theorized based on the process of collecting and analyzing the unstructured textual data. This is the fundamental difference in understanding the errors embedded in different research methods and analytic approaches.

It is worth noting that while we agree to the general approach to unstructured textual data and its premise identified in the literature (Schober et al., 2016), we include the traditional population coverage error in our framework given that it is necessary for researchers to think carefully about the sources of unstructured online data. For instance, collecting unstructured data from Facebook may be easier for researchers to further estimate the population coverage error since Facebook requires users to provide much more personal information in exchange of its service, whereas addressing population coverage error embedded in Twitter data with little information about the authors of tweets is extremely difficult. Additionally, the demographic profiles of social media users may vary significantly by platforms due to the differences in their services and marketing strategies. Therefore, researchers will benefit from understanding the population coverage error of the corpus during their analysis.

We further contend that both query and interpretation errors may better be seen as the systematic differences between alternative analytical procedures for discovering patterns in Twitter data. Unlike the probability sampling design in surveys, the bottom-up approach to the analysis of unstructured social media textual data seeks to achieve *topic coverage* rather than *population coverage* (Schober et al., 2016). This approach often involves identifying the most appropriate configuration to extract the data with the topic coverage for the research inquiry through multiple attempts, while the scope of the inquiry may also evolve to better match with what the potential patterns can answer. However, the changes in the parameters used for Twitter data

extraction are likely to frame the conceptual space of substantive inquiry and select the eligible observations in different ways. As a result, the batches of data may not be extracted from the same conceptual frame, and the construct being estimated by the substantive measures generated from different batches may be conceptually identical. Although the conceptual boundaries of query and interpretation errors may be clear, gauging the magnitude of these errors and comparing error estimates across procedures may still be quite difficult.

Given that demographic and geographic information are predominantly undisclosed as a typical social media behavior, missing data on these dimensions is the norm for Twitter. Therefore, we caution that interpretation error emerges, in part, when researchers employ predictive modeling techniques to address "missing" background information for producing additional analytical insights. For instance, it is difficult to discern the origins of tweets that have mentioned Springfield when the geographic metadata are unavailable, given that there are 41 cities named Springfield in the U.S.A. More importantly, Twitter users are not commonly asked to respond to any form of questions. If some Twitter users have never expressed anything about a specific matter, then their tweets will not be extracted by the search query. This is fundamentally different from surveys, where missing data are considered suboptimal responses to a standardized survey instrument presented to a group of targeted sample members selected from a carefully designed frame.

To better understand how these aforementioned errors stemming from research design decisions and contextual factors affect Twitter error, we illustrate, in Section 2.4, the major types of Twitter errors by examining public opinion on two topics using the identical research design and data extraction procedures.

2.4 Studying Public Opinion on the Twittersphere and the Potential Error Sources of Twitter Data: Two Case Studies

Abortion rights and marijuana legalization have been controversial issues for decades in the U.S.A. The morality and legality of abortion has been an ongoing debate since the landmark decision by the U.S. Supreme Court on *Roe v. Wade* in 1973. At the same time, the movement toward marijuana legalization has only gained substantial attention in recent years, with legalization in Colorado and Washington. The survey statistics from Gallup and GSS offer a glimpse into broad public opinion trends on these topics in the U.S.A.

Surveys suggest that the majority of Americans have been in favor of the legal abortion, at least under certain circumstances, since 1973. However, recent GSS results indicate a moderate decline in the approval of abortion (Smith and Son, 2013); Gallup has found that public opinion has been fairly evenly divided between pro-life and pro-choice attitudes since the late 1990s (Saad, 2015). Regarding marijuana legalization, the historical trends of both GSS and Gallup polls show that since the 1970s, the public support for legalizing marijuana in the U.S.A. has increased considerably and reached a point where as many approve as disapprove (Ingraham, 2015; Saad, 2014).

But broad national trends may not be sufficient depending on the research needs. Policy analysts may want to explore more specific events such as the public reaction to the decision by the U.S. Supreme Court on *Burwell v. Hobby Lobby Stores, Inc*., regarding the corporation's opposition to provide insurance coverage for contraception for their employees. Public health researchers may be interested in discovering potential and immediate impacts of the recent midterm election results on the attitudes toward marijuana legalization beyond Alaska, Oregon, and Washington DC. In these scenarios, the annual opinion estimates produced by national surveys

may only offer limited insights and not be able to provide timely information for such research endeavors. In contrast, a passive analysis of the timely data generated from Twitter or other social media may be an appropriate venue for answering these questions and one that can require significantly fewer resources. These historical trends of public opinion toward marijuana legalization and abortion rights can serve as examples to help illustrate potential error.

2.4.1 Research Questions and Methodology of Twitter Data Analysis

In order to investigate the utility and error inherent in the process of analyzing Twitter data to measure public opinion about marijuana legalization and abortion rights, we started by asking the question "Between 2011 and 2014, what were the patterns of opinions toward marijuana legalization and abortion rights in the U.S.A. expressed by users on Twitter?" Note that although the objective of this question is to understand the opinion of individuals, Twitter data are collected, and often analyzed, at the tweet level. Given that tweets serve as the unit for accessing, searching, and extracting the Twitter data, we considered tweets, or the post of content, as the unit of analysis when decomposing the query and interpretation errors associated with the Twitter data. It should be noted that such a mismatch in unit between surveys (individuals) and Twitter (tweets) complicates comparisons, and from a survey perspective, including potentially multiple opinions from single individuals may suggest a query error of duplication.

We collected four years of tweets, from January 1, 2011 through December 31, 2014, using Crimson Hexagon's Forsight[2] tool. This tool was selected as it represents a class of "off-the-shelf" social media analysis solutions that have become increasingly popular in research in recent years. Such systems gather data based on keyword specifications and conduct automated sentiment analysis with varying levels of guidance from the researcher. We selected this method to demonstrate what many researchers have employed to date; superior insights and flexibility may arise from directly extracting data from the Twitter Application Programming Interface (API), but at a more significant cost in terms of data access and programming time and expertise required.

The four-year window we selected provided us a time frame loosely matched to the available recent national estimates and respective news events in relation to these social issues. To filter and extract the relevant and appropriate Twitter data for the analysis, researchers need to construct a search query and then iteratively identify the optimal configuration to extract the data that achieve topic coverage (and consequently minimize query error). The Twitter query specifies the time frame, geography, language, and keywords of interest. In a way, specifying search queries can be similar to simultaneously defining the population of interest and writing survey questions. Specifying search queries is similar, in a way, to designing a sampling frame: what the users and the content researchers will collect depends, in part, on how they construct the parameters of the frame. At the same time, the information researchers will obtain from respondents depends, in part, on how they write and ask the question. Following this practical assumption, we conceive the search query specification process as somewhat similar to the survey design process, allowing us to conceptualize the potential error sources associated with social media data using the logic of TSE.

For our case studies, we used an iterative process for search query specification. First, we started by investigating whether geographic specification was necessary to reduce coverage error. Next, we constructed a *basic* query, casting a wide net by using some of the most popular and prominent keywords related to marijuana legalization and abortion rights. Given that RTs

2 See http://www.crimsonhexagon.com/PDFs/Crimson%20Hexagon%20ForSight%20Platform%20Overview%20Sheet.pdf for more information about this tool.

might be considered as repeated or duplicated observations, which may have significant implications for estimating public opinion on Twitter data, we constructed another query to assess the implications of the inclusion and exclusion of RTs. Next, we engaged in multiple iterations of query specification and identified an *expanded* query, including additional keywords that we considered most useful and appropriate for data extraction. Last, we considered interpretation error, comparing the results of our sentiment analysis on the Twitter data to survey data collected by Gallup.

In the case of public opinion about marijuana legalization, we started with a basic query informed by our observation of common terms[3] on Twitter: **(marijuana OR pot) AND (legal OR legalize OR legalization)** to extract the tweets that contain such keyword combinations within our time frame. Similarly, we began by using **prolife OR pro-life OR "pro life" OR pro-choice OR pro-choice OR "pro choice**" as our basic query to extract the tweets that may reveal the public opinion about abortion rights. With respect to the expanded queries for both topics, we specify our queries including an additional set of prominent keywords. For instance, we include other common keywords such as **weed**, **#mmj**, and **#mmot** in the expanded query to extract the tweets about marijuana legalization, and **praytoendabortion**, **stand4life**, **fem2**, and **waronwomen** in the expanded query about abortion rights. The full expanded query for marijuana legalization is **((marijuana OR pot OR 420 OR cannabis OR mmj OR weed OR hemp OR ganja OR THC) AND (legal OR legalize OR legalization)) OR mmot**. The full expanded query for abortion rights is **prolife OR pro-life OR "pro life" OR prochoice OR pro-choice OR "pro choice" OR praytoendabortion OR stand4life OR fem2 OR waronwomen**.

2.4.2 Potential Coverage Error in Twitter Examples

Our examples set the target population as U.S. adults, a common target in public opinion research. To assess the coverage of the U.S. adult population on Twitter, we turn to data gathered by the Pew Research Center Internet Project (Duggan et al., 2015). Based on representative surveys of the U.S. population, Pew found 23% of online adults using Twitter as of 2014. This rate differed by some demographic splits, as shown in Figure 2.2. The most dramatic (and statistically significant) difference is that those aged 18–29 years use Twitter at a much higher rate (37%) than other age groups. Lowest use is among the 65+, where only 10% are on Twitter. It should also be noted that 2014 represents the year of highest coverage among our four years of analysis. In 2011, only about 16% of U.S. online adults used Twitter. For any study aiming to portray the general U.S. population using Twitter data, the Pew figures serve as the best and most current source of information on the **coverage of the Twittersphere as a whole**. We know that most adults do not use Twitter and those who do are, on average, younger than the general population.

A second source of coverage error emerges when trying to determine the right geographic area within the Twittersphere or within the extracted Twitter data. In Figures 2.3 and 2.4, we examine the potential for coverage error in limiting the analysis of tweets to a specific geography for our selected examples. The solid line in Figure 2.3 presents the volume of all English-language tweets, without geographic restriction, by month that match the basic query for marijuana legalization for the years 2011 through 2014 in terms of tweets *per million overall tweets*. We use this metric rather than the raw count of tweets since Twitter use increased overall from about 3 to 21 billion tweets per month during this period and this uneven volume over time may falsely suggest a dramatic increase in discussion of these topics on Twitter. Over the course of the

3 The Forsight tool automatically includes mentions of these terms preceded by the # sign (i.e., "hashtagged" versions).