

HANDBOOK OF

Web Surveys

Survey Methodology

The *Wiley Handbooks in Survey Methodology* is a series of books that present both established techniques and cutting-edge developments in the field of survey research. The goal of each handbook is to supply a practical, one-stop reference that treats the statistical theory, formulae, and applications that, together, make up the cornerstones of a particular topic in the field. A self-contained presentation allows each volume to serve as a quick reference on ideas and methods for practitioners, while providing an accessible introduction to key concepts for students. The result is a high-quality, comprehensive collection that is sure to serve as a mainstay for novices and professional alike.

De Waal, Pannekoek, and Scholtus — *Handbook of Data Editing and Imputation*

Bethlehem, Cobben, and Schouten — *Handbook of Nonresponse in Household Surveys*

Forthcoming Wiley Handbooks in Survey Methodology

Alwin — *Handbook of Measurement and Reliability in the Social and Behavioral Sciences*

Larsen and Winkler — *Handbook of Record Linkage Methods*

Johnson — *Handbook of Health Survey Methods*

HANDBOOK OF

Web Surveys

JELKE BETHLEHEM

Statistics Netherlands
Division of Methodology and Quality
The Netherlands

SILVIA BIFFIGNANDI

University of Bergamo
Center for Statistics and Analysis of Sample Surveys
Bergamo, Italy

 **WILEY**

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2012 by John Wiley & Sons, Inc. All rights reserved.

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-0-470-60356-7

Printed in the United States of America

oBook ISBN: 978-1-118-12175-7
ePDF ISBN: 978-1-118-12172-6
ePub ISBN: 978-1-118-12174-0
eMobi ISBN: 978-1-118-12173-3

10 9 8 7 6 5 4 3 2 1

Contents

PREFACE	XI
1 THE ROAD TO WEB SURVEYS	1
1.1 Introduction, 1	
1.2 Theory, 2	
1.2.1 The Everlasting Demand for Statistical Information, 2	
1.2.2 The Dawn of Sampling Theory, 4	
1.2.3 Traditional Data Collection, 8	
1.2.4 The Era of Computer-Assisted Interviewing, 10	
1.2.5 The Conquest of the Web, 12	
1.3 Application, 21	
1.4 Summary, 31	
Key Terms, 31	
Exercises, 33	
References, 34	
2 ABOUT WEB SURVEYS	37
2.1 Introduction, 37	
2.2 Theory, 40	
2.2.1 Typical Survey Situations, 40	
2.2.2 Why On-Line Data Collection?, 45	
2.2.3 Areas of Application, 48	
2.2.4 Trends in Web Surveys, 50	
2.3 Application, 52	
2.4 Summary, 55	
Key Terms, 56	
Exercises, 56	
References, 58	
3 SAMPLING FOR WEB SURVEYS	59
3.1 Introduction, 59	
3.2 Theory, 60	

- 3.2.1 Target Population, 60
- 3.2.2 Sampling Frames, 63
- 3.2.3 Basic Concepts of Sampling, 68
- 3.2.4 Simple Random Sampling, 71
- 3.2.5 Determining the Sample Size, 74
- 3.2.6 Some Other Sampling Designs, 76
- 3.2.7 Estimation Procedures, 82
- 3.3 Application, 87
- 3.4 Summary, 92
 - Key Terms, 92
 - Exercises, 93
 - References, 94

4 ERRORS IN WEB SURVEYS 97

- 4.1 Introduction, 97
- 4.2 Theory, 103
 - 4.2.1 Measurement Errors, 103
 - 4.2.2 Nonresponse, 124
- 4.3 Application, 133
 - 4.3.1 The Safety Monitor, 133
 - 4.3.2 Measurement Errors, 134
 - 4.3.3 Nonresponse, 136
- 4.4 Summary, 138
 - Key Terms, 138
 - Exercises, 140
 - References, 143

5 WEB SURVEYS AND OTHER MODES OF DATA COLLECTION 147

- 5.1 Introduction, 147
 - 5.1.1 Modes of Data Collection, 147
 - 5.1.2 The Choice of the Modes of Data Collection, 149
- 5.2 Theory, 152
 - 5.2.1 Face-To-Face Surveys, 152
 - 5.2.2 Telephone surveys, 158
 - 5.2.3 Mail Surveys, 164
 - 5.2.4 Web surveys, 169
- 5.3 Application, 174
- 5.4 Summary, 182
 - Key Terms, 183
 - Exercises, 185
 - References, 187

6	DESIGNING A WEB SURVEY QUESTIONNAIRE	189
6.1	Introduction, 189	
6.2	Theory, 191	
6.2.1	The Road Map Toward a Web Questionnaire, 191	
6.2.2	The Language of Questions, 197	
6.2.3	Answers Types (Response Format), 200	
6.2.4	Basic Concepts of Visualization, 211	
6.2.5	Web Questionnaires and Paradata, 217	
6.2.6	Trends in Web Questionnaire Design and Visualization, 223	
6.3	Application, 226	
6.4	Summary, 228	
	Key Terms, 228	
	Exercises, 229	
	References, 231	
7	MIXED-MODE SURVEYS	235
7.1	Introduction, 235	
7.2	Theory, 238	
7.2.1	What is Mixed Mode?, 238	
7.2.2	Why Mixed Mode?, 243	
7.2.3	Methodological Issues, 248	
7.2.4	Mixed Mode for Business Surveys, 262	
7.2.5	Mixed Mode for Surveys Among Households and Individuals, 267	
7.3	Application, 272	
7.4	Summary, 274	
	Key Terms, 274	
	Exercises, 275	
	References, 277	
8	THE PROBLEM OF UNDERCOVERAGE	281
8.1	Introduction, 281	
8.2	Theory, 287	
8.2.1	The Internet Population, 287	
8.2.2	A Random Sample From the Internet Population, 288	
8.2.3	Reducing the Noncoverage Bias, 290	
8.2.4	Mixed-Mode Data Collection, 294	
8.3	Application, 295	
8.4	Summary, 299	
	Key Terms, 299	

Exercises, 300
References, 302

9 THE PROBLEM OF SELF-SELECTION 303

- 9.1 Introduction, 303
- 9.2 Theory, 306
 - 9.2.1 Basic Sampling Theory, 306
 - 9.2.2 A Self-Selection Sample from the Internet Population, 309
 - 9.2.3 Reducing the Self-Selection Bias, 314
- 9.3 Application, 319
- 9.4 Summary, 323
 - Key Terms, 323
 - Exercises, 324
 - References, 326

10 WEIGHTING ADJUSTMENT TECHNIQUES 329

- 10.1 Introduction, 329
- 10.2 Theory, 334
 - 10.2.1 The Concept of Representativity, 334
 - 10.2.2 Poststratification, 336
 - 10.2.3 Generalized Regression Estimation, 349
 - 10.2.4 Raking Ratio Estimation, 358
 - 10.2.5 Calibration Estimation, 361
 - 10.2.6 Constraining the Values of Weights, 362
 - 10.2.7 Correction Using a Reference Survey, 363
- 10.3 Application, 372
- 10.4 Summary, 378
 - Key Terms, 379
 - Exercises, 380
 - References, 383

11 USE OF RESPONSE PROPENSITIES 385

- 11.1 Introduction, 385
- 11.2 Theory, 389
 - 11.2.1 A Simple Random Sample with Nonresponse, 389
 - 11.2.2 A Self-Selection Sample, 392
 - 11.2.3 The Response Propensity Definition, 393
 - 11.2.4 Models for Response Propensities, 394
 - 11.2.5 Correction Methods Based on Response Propensities, 401
- 11.3 Application, 406
 - 11.3.1 Generation of the Population, 407
 - 11.3.2 Generation of Response Probabilities, 408

11.3.3	Generation of the Sample, 408	
11.3.4	Computation of Response Propensities, 408	
11.3.5	Matching Response Propensities, 409	
11.3.6	Estimation of Population Characteristics, 411	
11.3.7	Evaluating the Results, 412	
11.3.8	Model Sensitivity, 412	
11.4	Summary, 413	
	Key Terms, 414	
	Exercises, 414	
	References, 416	
12	WEB PANELS	419
12.1	Introduction, 419	
12.2	Theory, 422	
12.2.1	Web Panel Definition and Recruitment, 422	
12.2.2	Use of Web Panels, 426	
12.2.3	Web Panel Management, 427	
12.2.4	Response Rates, 432	
12.2.5	Representativity, 443	
12.3	Application, 449	
12.4	Summary, 451	
	Key Terms, 452	
	Exercises, 452	
	References, 454	
	INDEX	459

Preface

The survey research landscape has undergone radical changes over the last few decades. First, there was the change from traditional paper-and-pencil interviewing (PAPI) to computer-assisted interviewing (CAI). And now, face-to-face surveys (CAPI), telephone surveys (CATI), and mail surveys (CASI, CSAQ) are increasingly replaced by web surveys. The popularity of on-line research is not surprising. A web survey is a simple means of getting access to a large group of potential respondents. Questionnaires can be distributed at very low costs. No interviewers are needed, and there are no mailing and printing costs. Surveys can be launched very quickly. Little time is lost between the moment the questionnaire is ready and the start of the fieldwork. Web surveys also offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation, and movies).

At first sight, web surveys seem to have much in common with other types of surveys. It is just another mode of data collection. Questions are not asked face-to-face, by telephone, or by mail, but over the Internet. There are, however, various phenomena that can make the outcomes of web surveys unreliable. Examples of such phenomena are undercoverage, self-selection, and measurement errors. They can cause estimates of population characteristics to be biased, and therefore, wrong conclusions can be drawn from the collected data.

Undercoverage occurs if the target population is wider than just those having access to the Internet. Estimates will be biased if people with Internet access differ from people without Internet access.

Self-selection means that it is completely left to individuals to select themselves for the web survey. The survey questionnaire is simply put on the web. Respondents are those individuals who happen to have Internet access, visit the website, and decide to participate in the survey. These participants generally differ significantly from the nonparticipants.

General-population surveys that have to provide reliable and accurate statistics are traditionally conducted face-to-face or by telephone. There are interviewers to persuade people to cooperate and to help respondents in giving the right answers. Interviewer assistance is lacking for web surveys. This can have a serious impact on the quality of the collected data.

This book provides more insight into the possible use of web surveys for data collection. Web surveys promise lower data collection costs. Also, it is

expected that web surveys will increase response rates. But what about data quality? This book is devoted to many theoretical and practical aspects of web surveys. Therefore, it can be considered a handbook for those involved in practical survey research. This includes survey researchers working in official statistics (e.g., in national statistical institutes), academics, and commercial market research.

The book is written by two authors with ample expertise in survey methodology. They come from two different countries (the Netherlands and Italy) and different research organizations (a national statistical institute and a university). Therefore, they can present a broad view on various theoretical and practical aspects of web survey.

The first two chapters of the book are an introduction into web surveys. The first chapter gives a historic account of developments in survey research and shows how web surveys became a data collection tool. Chapter 2 is an overview of basic aspects of web surveys. It describes how web surveys can be used and where they can be used. Official statistics as well as research institutions, market research societies and private forums are all interested in web surveys both on households/individuals and on businesses.

Chapter 3 is about sampling aspects. It is stressed that only valid population inference is possible if some form of probability sampling is used. A proper sampling frame is required for this. Some sampling designs and estimation procedures useful for web surveys are discussed.

A researcher carrying out a survey can be confronted with many practical problems. An overview of possible errors is given in Chapter 4. Two types of errors are discussed in more detail. The first one is measurement errors. These can be caused by the lack of interviewers and by specific questionnaire design issues. The second type of problem is nonresponse. This phenomenon occurs in all surveys, but specific nonresponse aspects of web surveys require attention.

A web survey is just one mode of data collection. There are other modes like face-to-face surveys, telephone surveys, and mail surveys. Chapter 5 compares the various modes of data collection with on-line data collection. It discusses the advantages and disadvantages of each mode.

Because there are no interviewers in web surveys, the respondents are on their own when completing the survey questionnaire. This makes the design of the questionnaire crucially important. Small irregularities in the questionnaire form may have large consequences for the quality of the collected data. Questionnaire design issues are discussed in Chapter 6.

A web survey may not always be the ideal instrument for producing reliable and accurate statistics. Quality may be hampered by undercoverage problems and low response rates. An interesting alternative approach could be to set up a mixed-mode survey. This is a survey in which several modes of data collection are combined, either sequentially or concurrently. A mixed-mode survey is less expensive than a single-mode, interviewer-assisted survey (face-to-face or by telephone) and solves undercoverage problems, but also it introduces a new problem of mode effects. All these aspects of mixed-mode surveys are discussed in Chapter 7.

Chapter 8 is devoted to the problem of undercoverage. This still is a substantial problem in many countries because of low Internet coverage. It is also made clear that Internet access is often unevenly distributed over the population. It is shown how this can cause survey estimates to be biased. Some techniques are discussed that may be able to reduce undercoverage bias.

Chapter 9 is about self-selection. The proper, scientifically well-founded, principle is to use probability sampling to select people for a survey. Only then can reliable estimates of population characteristics be computed. It is nowadays easy to set up a web survey. Even people without any knowledge of or experience with surveys can do it with websites available for this purpose. Many of these web surveys do not apply probability sampling but rely on self-selection of respondents. This causes serious estimation problems. Self-selection and the consequences for the survey results are discussed in this chapter. It is also shown that correction techniques do not always work.

There can be many reasons why web-survey-based estimates are biased. Nonresponse, undercoverage, and self-selection are typical examples. Adjustment weighting is often applied in surveys to reduce a possible bias. Various weighting techniques are described in Chapter 10: poststratification, generalized regression estimation, and raking ratio estimation. It is explored whether these techniques can be effective for reducing a bias caused by undercoverage or self-selection.

Chapter 11 introduces the concepts of response probabilities. It is described how they can be estimated by means of response propensities. If response probabilities can be estimated accurately, they can be used to correct biased estimates. Two general approaches are described: response propensity weighting and response propensity stratification. The first approach attempts to adjust the original selection probabilities, and the second approach is a form of poststratification.

The final chapter is devoted to web panels. Particularly in the area of commercial market research, there are many such panels. A crucial aspect is how the panel members (households, individuals, firms, and shoppings) are recruited for such a panel. This can be done by means of a proper probability sample or by means of self-selection. This has consequences for the validity of the results of the specific surveys that are conducted using the panel members. Several quality indicators are discussed.

The accompanying website, www.web-survey-handbook.com, contains the survey data set of the General Population Survey (GPS). This data set has been used for many examples and applications in the book. The data set is available in SPSS (SPSS Corporation, Chicago, IL) format.

JELKE BETHLEHEM
SILVIA BIFFIGNANDI

The Road to Web Surveys

1.1 Introduction

Web surveys are a next step in the evolution of survey data collection. Collecting data for compiling statistical overviews is already very old, almost as old as humankind. All through history statistics have been used by rulers of countries to take informed decisions. However, new developments in society always have had their impact on the way the data were collected for these statistics.

For a long period, until the year 1895, statistical data collection was based on complete enumeration of populations. The censuses were mostly conducted to establish the size of the population, to determine tax obligations of the people, and to measure the military strength of the country. The idea of sampling had not emerged yet.

The year 1895 marks a fundamental change. Populations had grown bigger and bigger. It was the period of industrialization. Centralized governments required more and more information. The time was ripe for sample surveys. The first ideas emerged around 1895. A lot of discussion took place between 1895 and 1934 about how samples should be selected: by means of probability sampling or some other sample selection technique.

By 1934, it was clear that only surveys based on probability sampling could provide reliable and accurate estimates. Such surveys were accepted as a scientific method of data collection. In the period from the 1940s to the 1970s, most sample surveys were based on probability sampling. Questionnaires were printed on paper forms. They were completed in face-to-face, telephone, or mail surveys.

Somewhere in the 1970s another significant development began. The fast development of microcomputers made it possible to introduce computer-assisted

interviewing. This made survey data collection faster, cheaper, and easier, and it increased data quality. It was a time when acronyms like CATI (computer-assisted telephone interviewing) and CAPI (computer-assisted personal interviewing) emerged.

The next major development was the creation of the Internet around 1982. When more and more persons and companies received access to the Internet, it became possible to use this network for survey data collection. The first Internet surveys were e-mail surveys. In 1989, the World Wide Web was introduced. This software allowed for much friendlier graphical user interfaces for Internet users. The first browsers emerged, and the use of the Internet exploded. In the middle of 1990s, the World Wide Web became widely available and e-mail surveys were increasingly replaced by web surveys.

Web surveys are attractive because they have several advantages. They allow for simple, fast, and cheap access to large groups of potential respondents. Not surprisingly, the number of conducted web surveys has increased rapidly over time. There are, however, also potential methodological problems. Ample examples of web surveys are not based on probability sampling. Therefore, generalization of survey results to the population is questionable.

This chapter describes the historical developments that have led to the emergence of web surveys. As an illustration, Section 1.3 shows how these developments were implemented at Statistics Netherlands and led to new software for survey data collection.

1.2 Theory

1.2.1 THE EVERLASTING DEMAND FOR STATISTICAL INFORMATION

The history of data collection for statistics goes back in time thousands of years. As far back as Babylonian times, a census of agriculture was carried out. This already took place shortly after the art of writing was invented. The same thing happened in China. This empire counted its people to determine the revenues and the military strength of its provinces. There are also accounts of statistical overviews compiled by Egyptian rulers long before Christ. Rome regularly took censuses of people and of property. The collected data were used to establish the political status of citizens and to assess their military and tax obligations to the state.

Censuses were rare in the Middle Ages. The most famous one was the census of England taken by the order of William the Conqueror, King of England. The compilation of his *Domesday Book* started in the year 1086 AD. The book records a wealth of information about each manor and each village in the country. Information was collected about more than 13,000 places. More than 10,000 facts were recorded for each county.

To collect all this data, the country was divided into several regions. In each region, a group of commissioners was appointed from among the greater lords. Each county within a region was dealt with separately. Sessions were organized in

each county town. The commissioners summoned all those required to appear before them. They had prepared a standard list of questions. For example, there were questions about the owner of the manor; the number of free men and slaves; the area of woodland, pasture, and meadow; the number of mills and fishponds, to the total value; and the prospects of getting more profit. The *Domesday Book* still exists, and many county data files are available on CD-ROM and the Internet.

Another interesting example of the history of official statistics can be found in the Inca Empire that existed between 1000 and 1500 AD. Each Inca tribe had its own statistician, called the *Quipucamayoc*. This man kept records of the number of people, the number of houses, the number of llamas, the number of marriages, and the number of young men that could be recruited for the army. All these facts were recorded on *quipus*, a system of knots in colored ropes. A decimal system was used for this. At regular intervals, couriers brought the quipus to Cusco, the capital of the kingdom, where all regional statistics were compiled into national statistics. The system of Quipucamayocs and quipus worked remarkably well. The system vanished with the fall of the empire.

An early census also took place in Canada in 1666. Jean Talon, the intendant of New France, ordered an official census of the colony to measure the increase in population since the founding of Quebec in 1608. Name, age, sex, marital status, and occupation were recorded for every person. It turned out 3,215 people lived in New France.

The first censuses in Europe were conducted in the Nordic countries: The first census in Sweden-Finland took place in 1749. Not everyone welcomed the idea of a census. In particular, religious people believed that people should not be counted. They referred to the census ordered by King David in biblical times, which was interrupted by a terrible plague and never completed. Others said that a population count would reveal the strengths and weaknesses of a country to foreign enemies. Nevertheless, censuses were conducted in more and more countries. The first census in Denmark-Norway was done in 1769. In 1795, at the time of the Batavian Republic under Napoleon's influence, the first count of the population of the Netherlands took place. The new centralized administration wanted to gather quantitative information to devise a new system of electoral constituencies (see Den Dulk & Van Maarseveen, 1990).

In the period until the late 1880s, there were some applications of *partial investigations*. These were statistical inquiries in which only part of a complete human population was investigated. The way the persons were selected from the population was generally unclear and undocumented.

In the second half of the 19th century, so-called *monograph studies* became popular. They were based on Quetelet's idea of the average man. According to Quetelet, many physical and moral data have a natural variability. This variability can be described by a normal distribution around a fixed, true value. He assumed the existence of something called the *true value*. Quetelet introduced the concept of *average man* (*l'homme moyenne*) as a person of which all characteristics were equal to the true value, (see Quetelet, 1835, 1846).

The period of the 18th and 19th centuries is also called the era of the Industrial Revolution. It led to important changes in society, science, and

technology. Among many other things, urbanization started from industrialization and democratization. All these developments created new statistical demands. The foundations for many principles of modern social statistics were laid. Several central statistical bureaus, statistical societies, conferences, and journals were established soon after this period.

1.2.2 THE DAWN OF SAMPLING THEORY

The first ideas about survey sampling emerged in the world of official statistics. If a starting year must be chosen, 1895 would be a good candidate. Anders Kiaer, the founder and first director of Statistics Norway, started in this year a fundamental discussion about the use of sampling methods. This discussion led to the development, acceptance, and application of sampling as a scientific method.

Kiaer (1838–1919) was the founder and advocate of the survey method that is now widely applied in official statistics and social research. With the first publication of his ideas in 1895, he started the process that ended in the development of modern survey sampling theory and methods. This process is described in more detail by Bethlehem (2009).

It should be noted that earlier examples of scientific investigations have been based on samples, but they were lacking proper scientific foundations. The first known attempt of drawing conclusions about a population using only information about part of it was made by the English merchant John Graunt (1662). He estimated the size of the population of London. Graunt surveyed families in a sample of parishes where the registers were well kept. He found that on average there were three burials per year in 11 families. Assuming this ratio to be more or less constant for all parishes, and knowing the total number of burials per year in London to be about 13,000, he concluded that the total number of families was approximately 48,000. Putting the average family size at eight, he estimated the population of London to be 384,000. As this approach lacked a proper scientific foundation, John Graunt could not say how accurate his estimates were.

More than a century later, the French mathematician Pierre Simon Laplace realized that it was important to have some indication of the accuracy of his estimate of the French population. Laplace (1812) implemented an approach that was more or less similar to that of John Graunt. He selected 30 departments distributed over the area of France in such a way that all types of climate were represented. Moreover, he selected departments in which accurate population records were kept. Using the central limit theorem, Laplace proved that his estimator had a normal distribution. Unfortunately, he disregarded the fact that sampling was purposively and not at random. These problems made application of the central limit theorem at least doubtful.

In 1895, Kiaer (1895, 1997), the founder and first director of Statistics Norway, proposed his *representative method*. It was a partial inquiry in which a large number of persons was questioned. Selection of persons was such that a “miniature” of the population was obtained. Anders Kiaer stressed the importance of *representativity*. He argued that, if a sample was representative with respect to variables for which the population distribution was known, it would also be representative with respect to the other survey variables.

EXAMPLE 1.1 The representative method of Anders Kiaer

Anders Kiaer applied his representative method in Norway. His idea was to survey the population of Norway by selecting a sample of 120,000 people. Enumerators (hired only for this purpose) visited these people and filled in 120,000 forms. Approximately 80,000 of the forms were collected by the representative method and 40,000 forms by a special (but analogue) method in areas where the working class people lived.

For the first sample of 80,000 respondents, data from the 1891 census were used to divide the households in Norway into two strata. Approximately 20,000 people were selected from urban areas and the rest from rural areas.

Thirteen representative cities were selected from the 61 cities in Norway. All five cities having more than 20,000 inhabitants were included, as were eight cities representing the medium-sized and small towns. The proportion of selected people in cities varied: In the middle-sized and small cities, the proportion was greater than in the big cities. Kiaer motivated this choice by the fact that the middle-sized and small cities did not represent only themselves but a larger number of similar cities. In Kristiania (nowadays Oslo) the proportion was $1/16$; in the medium sized towns, the proportion varied between $1/12$ and $1/9$; and in the small towns, it was $1/4$ or $1/3$ of the population.

Based on the census, it was known how many people lived in each of the 400 streets of Kristiania, the capital of Norway. The streets were sorted in four categories according to the number of inhabitants. A selection scheme was then specified for each category: The whole adult population was enumerated in 1 out of 20 for the smallest streets. In the second category, the adult population was enumerated in half of the houses in 1 out of 10 of streets. In the third category, the enumeration concerned $1/4$ of the streets and every fifth house was enumerated; and in the last category of the biggest streets, the adult population was enumerated on half of the streets and in 1 out of 10 houses in them.

In selecting the streets, their distribution over the city was taken into account to ensure the largest possible dispersion and the “representative character” of the enumerated areas. In the medium-sized towns, the sample was selected using the same principles, although in a slightly simplified manner. In the smallest towns, the whole adult population in three or four houses was enumerated.

The number of informants in each of the 18 counties in the rural part of Norway was decided on the basis of census data. To obtain representativeness, municipalities in each county were classified according to their main industry, either as agricultural, forestry, industrial, seafaring, or fishing municipalities. In addition, the geographical distribution was taken into account. The total number of the representative municipalities

amounted to 109, which is 6 in each county on average. The total number of municipalities was 498.

The selection of people in a municipality was done in relation to the population in different parishes, and so that all different municipalities were covered. The final step was to instruct enumerators to follow a specific path. In addition, enumerators were instructed to visit different houses situated close to each other. That is, they were supposed to visit not only middle-class houses but also well-to-do houses, poor-looking houses, and one-person houses.

Kiaer did not explain in his papers how he calculated estimates. The main reason probably was that the representative sample was constructed as a miniature of the population. This made computations of estimates trivial: The sample mean is the estimate of the population mean, and the estimate of the population total could be attained simply by multiplying the sample total by the inverse of sampling fraction.

A basic problem of the representative method was that there was no way of establishing the precision of population estimates. The method lacked a formal theory of inference. It was Bowley (1906, 1926) who made the first steps in this direction. He showed that for large samples, selected at random from the population, estimates had an approximately normal distribution. From this moment on, there were two methods of sample selection:

- Kiaer's representative method, based on purposive selection, in which representativity played an essential role, and for which no measure of the accuracy of the estimates could be obtained;
- Bowley's approach, based on simple random sampling, and for which an indication of the accuracy of estimates could be computed.

Both methods existed side by side until 1934. In that year, the Polish scientist Jerzy Neyman published his famous paper (1934). Neyman developed a new theory based on the concept of the confidence interval. By using random selection instead of purposive selection, there was no need any more to make prior assumptions about the population. The contribution of Neyman was not only that he proposed the confidence interval as an indicator for the precision of estimates. He also conducted an empirical evaluation of Italian census data and proved that the representative method based on purposive sampling could not provide satisfactory estimates of population characteristics. He established the superiority of random sampling (also referred to as *probability sampling*) over purposive sampling. Consequently, use of purposive sampling was rejected as a scientific sampling method.

Gradually probability sampling found its way into official statistics. More and more national statistical institutes introduced probability sampling for

official statistics. However, the process was slow. For example, a first test of a real sample survey using random selection was carried out by Statistics Netherlands only in 1941 (CBS, 1948). Using a simple random sample of size 30,000 from the population of 1.75 million taxpayers, it was shown that estimates were accurate.

The history of opinion polls goes back to the 1820s, in which period American newspapers attempted to determine the political preference of voters just before the presidential election. These early polls did not pay much attention to sampling. Therefore, it was difficult to establish the accuracy of the results. Such opinion polls were often called *straw polls*. This expression goes back to rural America. Farmers would throw a handful of straws into the air to see which way the wind was blowing.

It took until the 1920s before more attention was paid to sampling aspects. Lienhard (2003) describes how George Gallup worked out new ways to measure interest in newspaper articles. Gallup used *quota sampling*. The idea was to investigate a group of people that could be considered representative for the population. Hundreds of interviewers across the country visited people. Interviewers were given a quota for different groups of respondents. They had to interview so many middle-class urban women, so many low-class rural men, and so on. In total, approximately 3,000 interviews were conducted for a survey.

Gallup's approach was in great contrast with that of *Literary Digest* magazine, which was at that time the leading polling organization. This magazine conducted regular "America Speaks" polls. It based its predictions on returned questionnaire forms that were sent to addresses taken from telephone directory books and automobile registration lists. The sample size for these polls was in the order of two million people. So the sample size was much larger than that of Gallup's polls.

The presidential election of 1936 turned out to be decisive for both methods. This is described by Utts (1999). Gallup correctly predicted Franklin D. Roosevelt to be the new president, whereas *Literary Digest* predicted that Alf Landon would beat Franklin D. Roosevelt. The prediction based on the very large sample size turned out to be wrong. The explanation was that the sampling technique of *Literary Digest* did not produce representative samples. In the 1930s, cars and telephones were typically owned by middle-and upper class people. These people tended to vote Republican, whereas lower class people were more inclined to vote Democrat. Consequently, Republicans were overrepresented in the *Literary Digest* sample.

As a result of this historic mistake, opinion researchers learned that they should rely on more scientific ways of sample selection. They also learned that the way a sample is selected is more important than the size of the sample.

The classic theory of survey sampling was more or less completed in 1952. Horvitz and Thompson (1952) developed a general theory for constructing unbiased estimates. Whatever the selection probabilities are, as long as they are known and positive, it is always possible to construct a useful estimate. Horvitz and Thompson completed this classic theory, and the random sampling approach was almost unanimously accepted. Most classic books about sampling also were published by then (Cochran, 1953; Deming, 1950; Hansen, Hurwitz, & Madow, 1953; Yates, 1949).

1.2.3 TRADITIONAL DATA COLLECTION

There were three modes of data collection in the early days of survey research: face-to-face interviewing, mail interviewing, and telephone interviewing. Each mode had its advantages and disadvantages.

Face-to-face interviewing was already used for the first censuses. So, it is not a surprise it was also used for surveys. Face-to-face interviewing means that interviewers visit the persons selected in the sample. Well-trained interviewers will be successful in persuading reluctant persons to participate in the survey. Therefore, response rates of face-to-face surveys are usually higher than surveys not involving interviewers (for example, mail surveys). Interviewers can also assist respondents in giving the right answers to the questions. This often results in better data. However, the presence of interviewers can also be a drawback. Research suggests that respondents are more inclined to answer sensitive questions properly if no interviewers are present.

Survey agencies often send a letter announcing the visit of the interviewer. Such a letter also can give additional information about the survey, explain why it is important to participate, and assure that the collected information is treated confidentially. As a result, the respondents are not taken by surprise by the interviewers.

The response rate of a face-to-face survey is usually high and so is the quality of the collected data. But a price has to be paid literally: Face-to-face interviewing is much more expensive. A team of interviewers has to be trained and paid. They also have to travel, which costs time and money.

Mail interviewing is much less expensive than face-to-face interviewing. Paper questionnaires are sent by mail to persons selected in the sample. They are invited to answer the questions and to return the completed questionnaire to the survey agency. A mail survey does not involve interviewers. Therefore, it is a cheap mode of data collection. Data collection costs only involve mailing costs (letters, postage, and envelopes). Another advantage is that the absence of interviewers can be experienced as less threatening for potential respondents. As a consequence, respondents are more inclined to answer sensitive questions properly.

The absence of interviewers also has several disadvantages. There are no interviewers to explain questions or to assist respondents in answering them. This may cause respondents to misinterpret questions, which has a negative impact on the quality of the collected data. Also, it is not possible to use show cards. A *show card* is typically used for answering closed questions. Such a card contains the list of all possible answers to a question. Respondents can read through the list at their own pace and select the answer corresponding to their situation or opinion. Mail surveys put high demands on the design of the paper questionnaire. For example, it should be clear to all respondents how to navigate through the questionnaire and how to answer questions.

As the persuasive power of the interviewers is absent, the response rates of mail surveys tend to be low. Of course, reminder letters can be sent, but this is often not very successful. More often survey questionnaire forms end up in the pile of old newspapers.

In summary, the costs of a mail survey are relatively low, but often a price has to be paid in terms of data quality: Response rates tend to be low and the quality

of the collected data is also often not very good. Dillman (2007) believes, however, that good results can be obtained by applying his tailored design method. This set of guidelines is used for designing and formatting mail survey questionnaires. It pays attention to all aspects of the survey process that may affect response rates or data quality.

Face-to-face interviewing was preferred in the early days of survey interviewing in the Netherlands. The idea was in the 1940s that poor people had poor writing skills, and moreover, they were not interested in the topics of the surveys. Therefore, they had a smaller probability of completing mail questionnaires. People completing and returning questionnaire forms were assumed to be more interested in the survey topics because their intelligence and social-economic position was above average.

A third mode of data collection is *telephone interviewing*. Interviewers are needed to conduct a telephone survey, but not as many as for a face-to-face survey because they do not have to travel from one respondent to the next. They can remain in the call center of the survey agency and can conduct more interviews in the same amount of time. Therefore, interviewer costs are less. An advantage of telephone interviewing over face-to-face interviewing is that respondents may be more inclined to answer sensitive questions because the interviewer is not present in the room. A drawback in the early days of telephone surveys was that telephone coverage in the population was small. Not every respondent could be contacted by telephone.

EXAMPLE 1.2 The first telephone survey in the Netherlands

The first telephone survey was conducted in the Netherlands on June 11, 1946. See NIPO (1946) for a detailed description. A few hundred owners of telephones in Amsterdam were asked to answer a few questions about listening to the radio. The people were called between 20:00 and 21:30 hours on a Tuesday night. Some results are given in Table 1.1.

TABLE 1.1 The first telephone survey in the Netherlands

Are you listening to the radio at this moment?	Percentage
Was listening	24%
Was not listening	38%
Line busy	5%
No answer	31%
Did not have a radio	2%

People listening to the radio also were asked which program they were listening to. It turned out that 85% was listening the “Bonte Dinsdagavondtrein,” a very famous radio show at that time.

Telephone interviewing has some limitations. Interviews cannot last too long, and questions may not be too complicated. Another problem may be the lack of a proper sampling frame. Telephone directories may suffer from severe undercoverage because many people do not want their phone number to be listed in the directory. Another, new, development is that increasingly people replace their landline phone by a mobile phone. Mobile phone numbers are not listed in directories in many countries. For example, according to Fannic Cobben and Jelke G. Bethlehem (2005), only between 60% and 70% of the Dutch population can be reached through a telephone dictionary. For more information about the use of mobile phones for interviewing, see Kuusela, Vehovar and Callegaro (2006).

A way to avoid the undercoverage problems of telephone directories is to apply *random digit dialing* (RDD) to generate random phone numbers. A computer algorithm computes valid random telephone numbers. Such an algorithm can generate both listed and unlisted numbers. So, there is complete coverage. An example of an algorithm used in the United Kingdom is to take a number from a directory and replace its last digit by a random digit. Random digit dialing also has drawbacks. In some countries, it is not clear what an unanswered number means. It can mean that the number is not in use. This is a case of overcoverage. No follow-up is needed. It also can mean that someone simply does not answer the phone, which is a case of nonresponse, that has to be followed up. Another drawback of RDD is that there is no information at all about nonrespondents. This makes correction for nonresponse very difficult (see also Chapter 10 about weighting adjustments).

The choice of the mode of data collection is not any easy one. It is usually a compromise between quality and costs. In large countries (like the United States) or sparsely populated countries (like Sweden), it is almost impossible to collect survey data by means of face-to-face interviewing. It requires so many interviewers that have to do so much traveling that the costs would be very high. Therefore, it is not surprising that telephone interviewing emerged here as a major data collection mode. In a very small and densely populated country like the Netherlands, face-to-face interviewing is much more attractive. The coverage problems of telephone directories and the low response rates also play a role in the choice for face-to-face interviewing. More about data collection issues can be found in the study by Couper et al. (1998).

1.2.4 THE ERA OF COMPUTER-ASSISTED INTERVIEWING

Collecting survey data can be a costly and time-consuming process, particularly if high-quality data are required, the sample is large, and the questionnaire is long and complex. Another problem of traditional data collection is that the completed paper questionnaire forms may contain many errors. Substantial resources must therefore be devoted to cleaning the data. Extensive data editing is required to obtain data of acceptable quality.

Rapid developments in information technology since the 1970s have made it possible to reduce these problems. This was accomplished by introducing microcomputers for data collection. The paper questionnaire was replaced by a

computer program asking the questions. The computer took control of the interviewing process, and it checked answers to the questions. Thus, *computer-assisted interviewing* (CAI) emerged.

Computer-assisted interviewing comes in different modes of data collection. The first mode of data collection that emerged was *computer assisted telephone interviewing* (CATI). Couper and Nicholls (1998) describe its development in the United States in the early 1970s. The first nationwide telephone facility for surveys was established in 1966. The idea at that time was not implementation of computer-assisted interviewing but simplification of sample management. The initial systems evolved in subsequent years into full featured CATI systems. Particularly in the United States, there was a rapid growth of the use of these systems. CATI systems were little used in Europe until the early 1980s.

Interviewers in a CATI survey operate a computer running interview software. When instructed to do so by the software, they attempt to contact a selected person by telephone. If this is successful and the person is willing to participate in the survey, the interviewer starts the interviewing program. The first question appears on the screen. If this is answered correctly, the software proceeds to the next question on the route through the questionnaire.

Call management is an important component of many CATI systems. Its main function is to offer the right telephone number at the right moment to the right interviewer. This is particularly important in cases in which the interviewer has made an appointment with a respondent for a specific time and date. Such a call management system also has facilities to deal with special situations like a busy number (try again after a short while) or no answer (try again later). This all helps to increase the response rate. More about the use of CATI in the United States can be found in the study by Nicholls and Groves (1986).

Small portable computers came on the market in the 1980s. This made it possible for the interviewers to take computers with them to the respondents. This is the computer-assisted form of face-to-face interviewing. It is called *computer-assisted personal interviewing* (CAPI). After interviewers have obtained cooperation from the respondents, they start the interviewing program. Questions are displayed one at a time. Only after the answer has been entered, will the next question appear on the screen.

At first, it was not completely clear whether computers could be used for this mode of data collection. There were issues like the weight and size of the computer, the readability of the screen, battery capacity, and the size of keys on the keyboard. Experiments showed that CAPI was feasible. It became clear that computer-assisted interviewing for data collection has three major advantages:

- It simplifies the work of interviewers. They do not have to pay attention any more to choosing the correct route through the questionnaire. The computer determines the next question to ask. Interviewers can concentrate more on asking questions and on helping respondents give the proper answers.
- It improves the quality of the collected data. Answers can be checked by the software during the interview. Detected errors can be corrected immediately. The respondent is there to provide the proper information. This is much

more effective than having to do data editing afterward in the survey agency and without the respondent

- Data are entered into the computer immediately during the interview. Checks are carried out straightaway, and detected errors are corrected. Therefore, the record of a respondent is “clean” after completion of the interview. No more subsequent data entry and/or data editing is required. Compared with the old days of traditional data collection with paper forms, this considerably reduces the time needed to process the survey data. Therefore, the timeliness of the survey results is improved.

More information about CAPI in general can be found in the study by Couper et al. (1998).

The computer-assisted mode of mail interviewing also emerged. It was called *computer-assisted self-interviewing* (CASI) or sometimes also *computer assisted self-administered questionnaires* (CSAQ). The electronic questionnaire program is sent to the respondents. They run the software, which asks the questions and stores the answers. After the interview has been completed, the data are sent back to the survey agency. Early CASI applications used diskettes or a telephone and modem to transmit the questionnaire and the answers to the question. Later it became common practice to use the Internet as a transport medium.

A CASI survey is only feasible if all respondents have a computer on which they can run the interview program. As the use of computers was more widespread among companies than households in the early days of CASI, the first CASI applications were business surveys. An example is the production of Fire Statistics in the Netherlands in the 1980s. Because all fire brigades had a microcomputer at that time, data for these statistics could be collected by means of CASI. Diskettes were sent to the fire brigades. They ran the questionnaire on their MS-DOS computers. The answers were stored on the diskette. After having completed the questionnaire, the diskette was returned to Statistics Netherlands.

An early application in social surveys was the *Telepanel*, which was set up by Saris (1998). The Telepanel started in 1986. It was a panel of 2,000 households that agreed to complete questionnaires regularly with the computer equipment provided to them by the survey organization. A home computer was installed in each household. It was connected to the telephone with a modem. It also was connected to the television set in the household so that it could be used as a monitor. After a diskette was inserted into the home computer, it automatically established a connection with the survey agency to exchange information (downloading a new questionnaire or uploading answers of the current questionnaires). Panel members completed a questionnaire each weekend. The Telepanel was in essence very similar to the web panels that are frequently used nowadays. The only difference was the Internet did not exist yet.

1.2.5 THE CONQUEST OF THE WEB

The development of the Internet started in the early 1970s. The first step was to create networks of computers. The U.S. Department of Defense decided to connect computers across research institutes. Computers were expensive. A

network made it possible for these institutes to share each other's computer resources. This first network was called ARPANET.

ARPANET became a public network in 1972. Software was developed to send messages over the network. Thus, e-mail was born. The first e-mail was sent in 1971 by Ray Tomlinson of ARPANET.

The Internet was fairly chaotic in the first decade of its existence. There were many competing techniques and protocols. In 1982, the TCP/IP set of protocols was adopted as the standard for communication of connected networks. This can be seen as the real start of the Internet.

Tim Berners-Lee and scientists at CERN, the European Organization for Nuclear Research in Geneva, were interested in making it easier to retrieve research documentation over the Internet. This led in 1989 to the *hypertext* concept. This is text containing references (hyperlinks) to other texts the reader can immediately access. To be able to view these text pages and navigate to other pages through the hyperlinks, Berners-Lee developed computer software. He called this program a *browser*. This first browser was named the *World Wide Web*. This name is now used to denote the whole set of linked hypertext documents on the Internet.

In 1993, Mark Andreessen and his team at the National Center for Supercomputing Applications (NCSA, IL) developed the browser *Mosaic X*. It was easy to install and use. This browser had increased graphic capabilities. It already contained many features that are common in current browsers. It became a popular browser, which helped to spread the use of the World Wide Web across the world.

The rapid development of the Internet led to new modes of data collection. Already in the 1980s, prior to the widespread introduction of the World Wide Web, e-mail was explored as a new mode of survey data collection. Kiesler and Sproull (1986) described an early experiment conducted in 1983. They compared an e-mail survey with a traditional mail survey. They showed that the costs of an e-mail survey were much less than those of a mail survey. The response rate of the e-mail survey was 67%, and this was somewhat smaller than the response rate of the mail survey (75%). The turnaround time of the e-mail survey was much shorter. There were less socially desirable answers and less incomplete answers. Kiesler and Sproull (1986) noted that limited Internet coverage restricted wide-scale use of e-mail surveys. In their view, this type of data collection was only useful for communities and organizations with access to, and familiarity with, computers. These were relatively well-educated, urban, white collar, and technologically sophisticated people.

Schaefer and Dillman (1998) also compared an e-mail surveys with mail surveys. They applied knowledge about mail surveys to e-mail surveys and developed an e-mail survey methodology. They also proposed mixed-mode surveys for populations with limited Internet coverage. They pointed out some advantages of e-mail surveys. In the first place, e-mail surveys could be conducted very fast, even faster than telephone surveys. This was particularly the case for large surveys, where the number of available telephones and interviewers may limit the number of cases that can be completed each day. In the second place, e-mail surveys were inexpensive because there were no mailing, printing, and interviewers costs.

The experiment of Schaefer and Dillman (1998) showed that the response rates of e-mail and mail surveys were comparable, but the completed questionnaires of the e-mail survey were received much quicker. The answers to open questions were, on average, longer for e-mail surveys. This did not come as a surprise because of the relative ease of typing an answer on a computer compared with writing an answer on paper. There was lower item nonresponse for the e-mail survey. A possible explanation was that moving to a different question in an e-mail survey is much more difficult than moving to a different question on a paper form.

Couper, Blair, and Triplett (1999) found lower response rates for e-mail surveys in an experiment with a survey among employees of statistical agencies in the United States. They pointed out that nonresponse can partly be explained by delivery problems of the e-mails and not by refusal to participate in the survey. For example, if people do not check their e-mail or if the e-mail with the questionnaire does not pass a spam filter, people will not be aware of the invitation to participate in a survey.

Most e-mail surveys could not be considered a form of computer-assisted interviewing. It was merely the electronic analogue of a paper form. There was no automatic routing and no error checking. See Figure 1.1 for a simple example of an e-mail survey questionnaire. It is sent to the respondents. They are asked to reply to the original message. Then they answer the questions in the questionnaire in the reply message. For closed questions, they do that by typing an "X" between the brackets of the option of their choice. The answer to an open question is typed between the corresponding brackets. After completion, they send the e-mail message to the survey agency.

Use of e-mail imposes substantial restrictions on the layout. Because of the e-mail software of the respondent and the settings of the software, the

<ol style="list-style-type: none">1. What is your age? []2. Are you male or female? [] Male [] Female3. What is your marital status? [] Married [] Not married4. Do you have a job? [] Yes [] No5. What kind of job do you have? []6. What is your yearly income? [] Less than \$20,000 [] Between \$20,000 and \$40,000 [] More than \$40,000
--

FIGURE 1.1 Example of an e-mail survey questionnaire

questionnaire may look different to different respondents. For example, to avoid problems caused by line wrapping, Schaefer and Dillman (1998) advise a line length of at most 70 characters.

Schaefer and Dillman (1998) also noted another potential problem of e-mail surveys: the lack of anonymity of e-mail. If respondents reply to the e-mail with the questionnaire, it is difficult to remove all identifying information. Some companies have the possibility to monitor the e-mails of their employees. If this is the case, it may become difficult to obtain high response rates and true answers to the questions asked.

Personalization may help to increase response rates in mail surveys. Therefore, this principle should also be applied to e-mail surveys. An e-mail to a long list of addresses does not help to create the impression of personal treatment. It is probably better to send a separate e-mail to each selected person individually.

EXAMPLE 1.3 The first e-mail survey at Statistics Netherlands

The first test with an e-mail survey at Statistics Netherlands was carried out in 1998. At the time, Internet browsers and HTML were not sufficiently developed and used to make a web survey feasible.

The objective of the test was to explore to what extent e-mail could be used to collect data for the Survey on Short Term Indicators. This was a noncompulsory panel survey, where companies answered a small number of questions about production expectations, order-books, and stocks.

The traditional mode of data collection for this survey was a mail survey.

The test was conducted in one wave of the survey. A total of 1,600 companies were asked to participate in the test. If they did, they had to provide their e-mail address. Approximately 190 companies agreed to participate. These were mainly larger companies with a well-developed computer infrastructure.

A simple text form was sent to these companies by means of e-mail. After activating the reply-option, respondents could fill in answers in the text. It was a software-independent and platform-independent solution, but it was primitive from a respondent's point of view.

The test was a success. The response rate among the participating companies was almost 90%. No technical problems were encountered. Overall, respondents were positive. However, they considered the text-based questionnaire old-fashioned and not very user-friendly.

More details about this first test with an e-mail survey at Statistics Netherlands can be found in the study by Roos, Jaspers, and Snijkers (1999).

It should be noted that e-mail also can be used in a different way to send a questionnaire to a respondent. An electronic questionnaire can be offered as an executable file that is attached to the e-mail. The respondents download this interview program on their computers and run it. The advantage of this approach is that such a computer program can have a much better graphical user interface. Such a program also can include routing instructions and checks. This way of data collection is sometimes called CASI.

EXAMPLE 1.4 The production statistics pilot at Statistics Netherlands

In October 2004, Statistics Netherlands started a pilot to find out whether a CASI approach could be used to collect data for yearly production statistics.

One of the approaches tested is denoted by electronic data reporting (EDR). It was a system for responding companies to manage interviewing programs (generated by the Blaise System) on their own computers. The EDR software could be sent to respondents on CD-ROM, or respondents could download the software from the Internet. After the software had been installed, new survey interviews could be sent to respondents by e-mail. These electronic questionnaires were automatically imported in the EDR environment. A simple click would start the interview. After off-line completion of the interview, the entered data were automatically encrypted and sent to Statistics Netherlands.

The pilot made clear that downloading the software was feasible. It should be preferred over sending a CD-ROM because it was simpler to manage and less expensive. Some companies experiences problems with downloading and installing the software because security settings of their computer systems and networks prevented them of doing so. User-friendliness and ease of navigation turned out to be important issues for respondents.

For more information about this pilot, see Snijkers, Tonglet, and Onat (2004, 2005).

This form of CASI also has disadvantages. It requires respondents to have computer skills. They should be able to download and run the interviewing program. Couper et al. (1999) also note that problems may be caused by that fact that different users may have different operating systems on their computers or different versions of the same operating system. This may require different versions of the interviewing program. And it must be known in advance which operating system a respondent has. Moreover, the size of an executable file may be substantial, which may complicate sending it by e-mail.

E-mail surveys had the advantages of speed and low costs. Compared with computer-assisted interviewing, it had the disadvantages of a poor user interface

and lack of adequate editing and navigation facilities. An e-mail questionnaire was just a paper questionnaire in an e-mail. The Internet became more interesting for survey data collection after HTML 2.0 was introduced in 1995. HTML stands for HyperText Markup Language. It is the markup language for web pages. The first version of HTML was developed by Tim Berners-Lee in 1991. Version 2 of HTML included support for forms. This made it possible to transfer data from a user to the web server. Web pages could contain questions, and the answers could be collected by the server.

EXAMPLE 1.5 Designing questions in HTML 2.0

Version 2.0 of HTML made it possible to implement questions on a web page. The `<input>` tag can be used to define different types of questions. With `type=radio`, this tag becomes a *radio button*. A *closed question* is defined by introducing a radio button for each possible answer. See Figure 1.2 for an example. Not more than one radio button can be selected. This corresponds to a closed question for which only one answer must be selected.

Survon - Surveys Online ✓

Labor Force Survey Question 7 of 9

What is your yearly income?

Less than 20,000 euro

Between 20,000 and 40,000 euro

More than 40,000 euro

Previous Next

FIGURE 1.2 A closed question in HTML

Sometimes respondents must be offered the possibility to select more than one answer, like in Figure 1.3. Respondents are asked for their means of transport to work. Some people may use several transport means. For

Survon - Surveys Online ✓

Labor Force Survey Question 8 of 9

How do you travel to work?

Walking

By bicycle

By car

By public transport

Other means of transport

Previous Next

FIGURE 1.3 A check-all-that-apply question in HTML

example, a person may first take a bicycle to the railway station, and then continues by train. Such a closed question is sometimes also called a *check-all-that-apply* question. It can be implemented in HTML by means of a series of *checkboxes*. A checkbox is obtained by setting the type of the `<input>` tag to `checkbox`.

Figure 1.4 shows the implementation of an open question. Any text can be entered in the input field. A limit may be set to the length of the text. An open question is defined with `type=text` for the `<input>` tag.

The screenshot shows a web survey interface for 'Survion - Surveys Online'. The survey title is 'Labor Force Survey' and it is 'Question 6 of 9'. The question is 'What kind of job do you have?'. The answer field contains the text 'Statistician'. There are 'Previous' and 'Next' navigation buttons at the bottom.

FIGURE 1.4 An open question in HTML

If an input field is preferred that allows for more lines of text to be answered, the `<textarea>` tag can be used for this.

There are no specific types of the `<input>` tag for other types of questions. However, most of these question types can be implemented with the input field of an open question. For example, Figure 1.5 shows an numeric question. The question is basically an open question, but extra checks on the answer only allow numbers to be entered within certain bounds.

The screenshot shows a web survey interface for 'Survion - Surveys Online'. The survey title is 'Labor Force Survey' and it is 'Question 2 of 9'. The question is 'What is your age?'. The prompt is 'Enter a number between 0 and 99:'. The answer field contains the number '37'. There are 'Previous' and 'Next' navigation buttons at the bottom.

FIGURE 1.5 A numeric question in HTML

Date questions can be specified as a set of three input fields: one for the day, one for the month, and one for the year.

In the first years of the World Wide Web, use of web surveys was limited by the low penetration of the Internet. Internet penetration was higher among establishments than among households. Therefore, it is not surprising that first experiments tested the use of web business surveys. Clayton and Werking (1998)

describe a pilot carried out in 1996 for the Current Employment Statistics (CES) program of the U.S. Bureau of Labor Statistics. They expected the World Wide Web to offer a low-cost survey environment. Because it was a form of true on-line data collection, an immediate response to the answers of the respondents was possible. This could improve data quality. They also saw the great flexibility of web survey questionnaires. They could be offered in a form layout or in a question-by-question approach. The drawback was the limited number of respondents having access to the Internet. Only 11% of CES respondents had access to Internet and a compatible browser.

Roos and Wings (2000) conducted a test with Internet data collection at Statistics Netherlands for the construction industry. Respondents could choose among three modes:

- Completing a form off-line. The form was sent as an HTML-file that was attached to an e-mail. The form is downloaded, completed off-line, and returned by e-mail.
- Completing a form on-line. The Internet address of an on-line web form was sent by e-mail. The form was completed online.
- Completing an e-mail form. An e-mail is sent containing the questionnaire in plain text. Respondents clicked the reply button, answered the questions, and sent the e-mail back.

A sample of 1,500 companies was invited to participate in the experiment. Overall, 188 companies were willing and able to participate. Of those, 149 could surf the Internet and 39 only had e-mail. Questionnaire completion times of all three modes were similar to that of a paper form. Respondents preferred the form-based layout over the question-by-question layout. The conclusion of the experiment was that web surveys worked well.

General-population web surveys were rare in the first period of existence of the Internet. This was attributed to the low Internet penetration among households. This prevented conducting representative surveys. However, there were polls on the Internet. Recruitment of respondents was based on self-selection and not on probability sampling. Users could even create their own polls on websites like *Survey Central*, *Open Debate* and *Internet Voice*; see O'Connell (1998).

Also in 1998, the *Survey 2000* project was carried out. This was a large self-selection web survey on the website of the National Geographic Society. This was a survey on mobility, community, and cultural identity. In a period of two months over 80,000 respondents completed the questionnaire. See Witte, Amoroso, and Howard (2000), for more details about this project.

It seems to be typical for this type of self-selection web surveys that they make it possible to collect data about a large number of respondents in a relatively short time. Other examples are given by Bethlehem and Stoop (2007). The survey *21minuten.nl* has been conducted several times in the Netherlands. This survey was supposed to supply answers to questions about important

problems in Dutch society. Within a period of 6 weeks in 2006, approximately 170,000 people completed the online questionnaires. A similar survey was conducted in Germany. It is called *Perspektive Deutschland*. More than 600,000 participated in this survey in 2005/2006.

It should be noted that these large sample sizes are no guarantee for proper statistical inference. Because of undercoverage (not everyone has access to the Internet) and self-selection (no proper random sampling) estimates can be biased. This bias is independent of the sample size.

Internet penetration is still low in many countries, making it almost impossible to conduct a general population web survey. Because data collection costs can be reduced if the Internet is used, other approaches are sought. One such approach is *mixed-mode data collection*. A web survey is combined with one or more other modes of data collection, like a mail survey, a telephone survey, or a face-to-face survey. Researchers first attempt to collect as much data as possible with the cheapest mode of data collection (web). Then, the nonrespondents are reapproached in a different (next cheapest) mode, and so on.

EXAMPLE 1.6 An experiment with a mixed-mode survey

Beukenhorst and Wetzels (2009) describe a mixed-mode experiment conducted by Statistics Netherlands. They used the Dutch Safety Monitor for this experiment. This survey asks questions about feelings of security, quality of life, and level of crime experienced. The sample for this survey was selected from the Dutch population register. All sample persons received a letter in which they were asked to complete the survey questionnaire on the Internet. The letter also included a postcard that could be used to request a paper questionnaire. Two reminders were sent to those that did not respond by web or mail. If still no response was obtained, nonrespondents were approached by means of CATI, if a listed telephone number was available. If not, these nonrespondents were approached by CAPI.

To be able to compare this four-mode survey with a traditional survey, also a two-mode survey was conducted for an independent sample. Sampled persons were approached by CATI if their telephone number was listed in the directory, and otherwise, they were approached by CAPI.

The response rate for the four-mode survey turned out to be 59.7%. The response rate for the two-mode survey was 63.5%. So, introducing more modes did not increase the overall response rate. However, more than half of the response (58%) in the four-mode survey was obtained with a self-administered mode of data collection (web or paper). Therefore, the costs of the survey were much lower. Interviewers were deployed in only 42% of the cases. For more detail, see Beukenhorst and Wetzels (2009) or Bethlehem, Cobben, and Schouten (2011).

1.3 Application

The historic developments with respect to surveys as described in the previous section also took place in the Netherlands. In particular, the rapid developments in computer technology have had a major impact on the way Statistics Netherlands collected its data. Efforts to improve the process of collecting and processing survey data in terms of costs, timeliness, and quality have led to a powerful software system called Blaise. This system emerged in the 1980s, and it has evolved over time so that it is now also able to conduct web surveys and mixed-mode surveys. The section gives an overview of the developments at Statistics Netherlands leading to Internet version of Blaise.

All statistics published by Statistics Netherlands in the first half of the 20th century were based on a complete enumeration. Either data were collected by means of a population census or the data were obtained from a population register. One of the first real applications of sampling took place in 1947 with respect to income statistics. In 1946, a complete enumeration had been carried out. It meant processing data on 4 million tax administration cards. As the quality of the data on the cards was not very good, substantial manual editing was required. To reduce the magnitude of this immense operation, it was decided to use sampling methods for subsequent years.

Statistics Netherlands started using sample survey methods for agricultural statistics in the same period. Surveys were carried out from 1947 to estimate agricultural production. Samples were selected from a sampling frame consisting of a list of addresses of farms. These lists were compiled in the agricultural census (a complete enumeration) that was conducted every year in the month of May. A stratified sample was selected, where strata were formed based on province and size of farms. Within each stratum, systematic samples were selected. The total sample size was 10,000 to 20,000 farms. The surveys allowed for early estimates for the type and size of agricultural production.

Collecting and processing statistical data was a time-consuming and expensive process. Data editing was an important component of this work. The aim of these data editing activities was to detect and correct errors in the individual records, questionnaires, or forms. This should improve the quality of the results of surveys. Since statistical offices attached much importance to this aspect of the survey process, a large part of human and computer resources are spent on it.

To obtain more insight into the effectiveness of data editing, Statistics Netherlands carried out a Data Editing Research Project in 1984. Bethlehem (1987) describes how survey data were processed at that time. After all paper forms had been collected, subject-matter specialists checked them for completeness. If necessary and possible, skipped questions were answered, and obvious errors were corrected on the forms. Sometimes, the data on a form were manually copied to a new form to allow for the subsequent step of fast data entry. Next, the forms were transferred to the data entry department. Data typists entered the data in the computer at high speed without error checking. The computer was a dedicated data entry system. After data entry, the files were

transferred to the mainframe computer system. On the mainframe, an error detection program was run. Usually, this was a dedicated program written in the Cobol language. This program produced a list of detected errors. This list was sent to the subject-matter department. Specialists investigated the error messages, located and consulted corresponding forms, and corrected errors on the lists. Corrected forms were sent to the data entry department, and data typists entered the corrections in the data entry computer. The file with corrections was transferred to the mainframe computer. Corrected records and already present correct records were merged. The cycle of batch-wise error detection and manual correction was repeated until the number of detected errors was sufficiently small.

The Data Editing Research Project discovered several problems:

- Various people from different departments were involved. Many people dealt with the information: respondents, subject-matter specialists, data typists, and computer programmers. Transfer of material from one person/department to another could be a source of error, misunderstanding, and delay.
- Different computer systems were involved. Most data entry was carried out on Philips P7000 minicomputer systems, and data editing programs ran on a CDC Cyber 855 mainframe. Furthermore, there was a variety of desktop (running under MS-DOS) and other systems. Transfer of files from one system to another caused delay, and incorrect specification and documentation could produce errors.
- Not all activities were aimed at quality improvement. Time was also spent on just preparing forms for data entry and not on correcting errors.
- The process was going through cycles, from one department to another, and from one computer system to another. The cycle of data entry, automatic checking, and manual correction was in many cases repeated three times or more. Because of these cycles, data processing was very time consuming.
- The structure and nature of the data (the metadata) had to be specified in nearly every step of the data editing process. Although essentially the same, the “language” of this meta-data specification could be completely different for every department or computer system involved.

The conclusions of the Data Editing Research Project led to general redesign of the survey processes of Statistics Netherlands. The idea was to improve the handling of paper questionnaire forms by integrating data entry and data editing tasks. The traditional batch-oriented data editing activities, in which the complete data set was processed as a whole, was replaced by a record-oriented process in which records (forms) were completely dealt with one at a time.

The new group of activities was implemented in a so-called CADI system. CADI stands for *computer-assisted data input*. The CADI system was designed for use by the workers in the subject-matter departments. Data could be processed in two ways by this system:

- *Heads-up data entry*. Subject-matter employees worked through a pile of forms with a microcomputer, processing the forms one by one. First, they

entered all data on a form, and then they activated the check option to test for all kinds of errors. Detected errors were reported on the screen. Errors could be corrected by consulting forms or by contacting the suppliers of the information. After elimination of all errors, a “clean” record was written to file. If employees did not succeed in producing a clean record, they could write the record to a separate file of “dirty” records. A specialist could deal with these hard cases later on, also with a CADI system.

- *Heads-down data entry.* Data typists used the CADI system to enter data beforehand without much error checking. After completion, the CADI system checked in a batch run all records and flagged the incorrect ones. Then subject-matter specialists handled these “dirty” records one by one and corrected the detected errors.

To be able to introduce CADI on a wide scale in the organization, a new standard package was developed in 1986. The name of this standard package was *Blaise*. The basis of the Blaise System was the Blaise language, which was used to create a formal specification of the structure and contents of the questionnaire.

The first version of the Blaise System ran on microcomputers (or networks of microcomputers) under MS-DOS. It was intended for use by the people of the subject-matter departments; therefore, no computer expert knowledge was needed to use the Blaise system.

In the Blaise philosophy, the first step in carrying out a survey was to design a questionnaire in the Blaise language. Such a specification of the questionnaire contains more information than a traditional paper questionnaire. It did not only describe questions, possible answers, and conditions on the route through the questionnaire but also relationships between answers that had to be checked.

Figure 1.6 contains an example of a simple paper questionnaire. The questionnaire contains one route instruction: Persons without job are instructed to skip the questions about the type of job and income.

Figure 1.7 contains the specification of this questionnaire in the Blaise System. The first part of the questionnaire specification was the *Fields section*. It contains the definition of all questions that can be asked. A question consists of an identifying name, the text of the question as presented to the respondents, and a specification of valid answers. For example, the question about age has the name *Age*, the text of the question is “*What is your age?*”, and the answer must be a number between 0 and 99. The question *JobDes* is an open question. Any text not exceeding 20 characters is accepted. *Income* is a closed question. There are three possible answer options. Each option has a name (for example, *Less20*) and a text for the respondent (for example, *Less than 20,000*).

The second part of the Blaise specification is the *Rules section*. Here, the order of the questions is specified and the conditions under which they are asked. According to the rules section in Figure 1.7, every respondent must answer the questions *SeqNum*, *Age*, *Sex*, *MarStat*, and *Job* in this order. Only persons with a job (*Job = Yes*) have to answer the questions *JobDes* and *Income*.

The rules section can also contain checks on the answers of the questions. Figure 1.7 contains such a check. If people are younger than 15 years (*Age < 15*), then their marital status can only be not married (*MarStat = NotMar*).

1. Sequence number of the interview

2. What is your age?
 years

3. Are you male or female?
 Male
 Female

4. What is your marital status?
 Married
 Not married

5. Do you have a paid job?
 Yes
 No —————> END of questionnaire

6. What kind of job do you have?

7. What is your yearly income?
 Less than 20,000
 Between 20,000 and 40,000
 More than 40,000

FIGURE 1.6 A simple paper questionnaire

The check also contains texts that are used to display the error message on the screen. (*If respondent is younger than 15, then helshe is too young to be married!*)

The rules section also may contain computations. Such computations could be necessary in complex routing instructions or checks, or to derived new variables.

The first version of Blaise used a questionnaire specification to generate a CADI-program. Figure 1.8 shows how the computer screen of this MS-DOS program looked like for the Blaise questionnaire in Figure 1.7.

As this program was used by subject-matter specialists, only question names are shown on the screen. Additional information could be displayed through special keys.

Note that the input fields for the questions *Age* and *MarStat* contain error counters. These error indicators appeared because the answers to the questions *Age* (2) and *MarStat* (*Married*) did not pass the check.

```

DATAMODEL LFS "The Labour Force Survey";

FIELDS
  SeqNum "Sequence number of the interview?": 1..1000
  Age    "What is your age?": 0..99
  Sex    "Are you male or female?": (Male, Female)
  MarStat "What is your marital status?":
    (Married "Married",
     NotMar "Not married")
  Job    "Do you have a job?": (Yes, No)
  JobDes "What kind of job do you have?": STRING[20]
  Income "What is your yearly income?":
    (Less20 "Less than 20,000",
     Upto40 "Between 20,000 and 40,000",
     More40 "More than 40,000")

RULES
  SeqNum Age Sex MarStat Job
  IF Job = Yes THEN
    JobDes Income
  ENDIF

  IF Age < 15 "respondent is younger than 15" THEN
    MarStat = NotMar "he/she is too young to be married!"
  ENDIF

ENDMODEL

```

FIGURE 1.7 A simple Blaise questionnaire specification

CBS	BLAISE 1.11	CADI	LFS	Error(s) in form
SeqNum		11		
Age	1	2		
Sex		1	Male	
MarStat	1	1	Married	
Job		1	Yes	
JobDes		1	Programmer	
Income		1	Less20	

PAGING F1 = Help; F2 = Edit; ↑F2 = Store record; F3 = Check

FIGURE 1.8 A Blaise CADI program

After Blaise had been in use for a while, it was realized that such a system could be made much more powerful. The questionnaire specification in the Blaise system contained all knowledge about the questionnaire and the data needed for survey processing. Therefore, Blaise should be capable of handling computer-assisted interviewing.

Implementing computer-assisted interviewing means that the paper questionnaire is replaced by a computer program containing the questions to be asked. The computer takes control of the interviewing process. It performs two important activities:

- *Route control.* The computer program determines which question is to be asked next and displays that question on the screen. Such a decision may depend on the answers to previous questions. Hence, it relieves the interviewer of the task of taking care of the correct route through the questionnaire. As a result, it is not possible anymore to make route errors.
- *Error checking.* The computer program checks the answers to the questions that are entered. Range checks are carried out immediately after the answer has been entered and consistency checks after entry of all relevant answers. If an error is detected, the program produces an error message, and one or more of the answers concerned has to be modified. The program will not proceed to the next question until all detected errors have been corrected.

Application of computer-assisted data collection has three major advantages. In the first place, it simplifies the work of interviewer (no more route control); in the second place, it improves the quality of the collected data; and in the third place, data are entered in the computer during the interview resulting in a clean record, so no more subsequent data entry and data editing is necessary.

Version 2 of Blaise was completed in 1988. It implemented one form of computer-assisted interviewing: computer-assisted personal interviewing (CAPI). It is a form of face-to-face interviewing in which interviewers use a small laptop or notebook computer to ask the questions and to record the answers.

Figure 1.9 shows an example of a screen of a CAPI program generated by Blaise. The screen was divided into two parts. The upper part contains the current question to be answered (*What kind of a job do you have?*). After an answer had been entered, this question was replaced by the next question on the route.

Just displaying one question at a time gave the interviewers only limited feedback on where they are in the questionnaire. Therefore, the lower part of the screen displayed (in a very compact way) the current page of the questionnaire.

Statistics Netherlands started full-scale use of CAPI in a regular survey in 1987. The first CAPI survey was the Labor Force Survey. Each month, approximately 400 interviewers equipped with laptops visited 12,000 addresses. After a day of interviewing, they returned home and connected their computers to the power supply to recharge the batteries. The laptop also was connected to a telephone and modem. The collected data were automatically transmitted to the office at night. In return, new addresses were sent to the interviewers. The next



FIGURE 1.9 A Blaise CAPI program

morning, the batteries were recharged and the interviewing software was prepared for a new day of interviewing.

Another mode of computer-assisted interviewing was included in 1990: computer-assisted telephone interviewing (CATI). The interviewing program was installed on desktop computers. Interviewers called respondents from a central unit (call center), and they conducted interviews by telephone. The interviewing program for CATI was the same as that for CAPI. An important new tool for CATI was a call scheduling system. This system took care of proper handling of busy numbers (try again shortly), no-answers (try again later), appointments, and so on.

In the very early 1990s, nearly all household surveys of Statistics Netherlands had become CAPI or CATI surveys. Surveys using paper forms had almost become extinct. Table 1.2 lists all major and regular household surveys at that time together with their mode of interviewing.

In the middle of the 1990s, the MS-DOS operating system on micro-computers was gradually replaced by Windows (Microsoft Corporation, Redmond, WA). Particularly, the release of *Windows 95* was a success. It marked the start of the use of graphical user interfaces. Early versions of Microsoft's Internet browser *Internet Explorer* were included in this operating system.

The change of operating systems also had consequences for the Blaise System. Blaise 4 was the first production version of Blaise for Windows. It was released in 1999. The functionality of Blaise did not change, but the graphical user interface offered many more possibilities for screen layout. Figure 1.10 gives an example of a screen of the Blaise 4 CAPI program.

When more and more people and companies were connected to the Internet, web surveys became more and more a popular mode of data collection among researchers. The main reasons for this popularity were the high response

TABLE 1.2 Household surveys carried out by Statistics Netherlands in the early 1990s

Surveys	Modes	Interviews per year
Survey on Quality of Life	CAPI	7,500
Health Survey	CAPI	6,200
Day Recreation Survey	CAPI	36,000
Crime Victimization Survey	CAPI	8,000
Labour Force Survey	CAPI	150,000
Car Use Panel	CATI	8,500
Consumer Sentiments Survey	CATI	24,000
Social-Economic Panel	CATI	5,500
School Career Survey	CATI	4,500
Mobility Survey	CATI / CADI	20,000
Budget Survey	CADI	2,000

The Labour Force Survey

Forms Answer Navigate Options Help

What kind of job do you have?

Enter a text of at most 20 characters

SeqNum 123

Age 56

Sex 1 **Male**

MarStat 2 **NotMar**

Job 1 **Yes**

JobDes

Income

New 1/1 Modified Dirty Navigate LFS

FIGURE 1.10 The screen of a CAPI program in Blaise 4

speed, the possibility to provide feedback to respondents about the meaning of questions and possible errors, and the freedom for the respondents to choose their own moment to fill in the questionnaire.

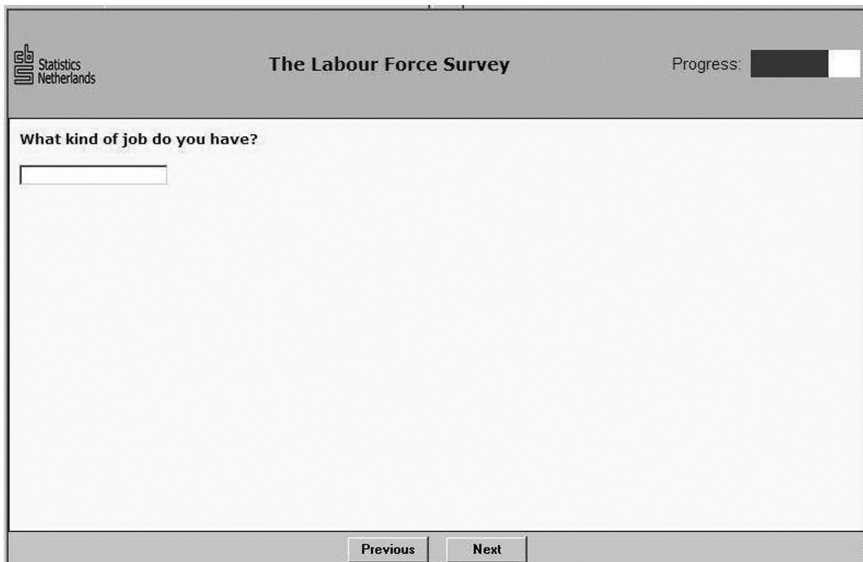
An attractive property of web surveys is that the questionnaire is completed in a browser. Respondents are familiar with browsers because they also use them for all their other activities on the Internet. So there was no need to explain the graphical user interface. The user interfaces for CAPI and CATI software were not that straightforward. Therefore, use of these software tools for computer-assisted self-interviewing was much more demanding.

The possibility of conducting web surveys was included in version 4.6 of Blaise that was released in 2003. Blaise provided two different approaches for web surveys: the interview approach and the form approach.

The *interview approach* is typically used for long and complex questionnaires that contain routing (skip) instructions and checks on the given answers. The respondent completes the questionnaire on-line because continuous interaction is required between the computer of the respondent and the software on the Internet server.

The Internet questionnaire is divided into pages. Each page may contain one or more questions. After the respondent has answered all the questions on a page, the answers are submitted to the Internet server. The answers are checked, and a new page is returned to the respondent. The contents of this page may depend on the answers to previous questions.

Figure 1.11 shows a simple example of a screen of a web survey. In this case, the page contains only one question. A progress indicator in the upper right corner keeps respondents informed about their progress in the questionnaire.



The screenshot shows a web browser window displaying a questionnaire page. At the top left is the logo for 'Statistics Netherlands'. The title 'The Labour Force Survey' is centered at the top. On the top right, there is a 'Progress:' label followed by a horizontal progress bar that is approximately 25% filled. The main content area contains the question 'What kind of job do you have?' followed by a single-line text input field. At the bottom of the page, there are two buttons labeled 'Previous' and 'Next'.

FIGURE 1.11 A web questionnaire using the interview approach

<p>4. What is your marital status?</p> <p><input type="radio"/> Married</p> <p><input type="radio"/> Not married</p>
<p>5. Do you have a job?</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p>
<p>6. What kind of job do you have?</p> <p><input type="text"/></p>
<p>7. What is your yearly income?</p> <p><input type="radio"/> Less than 20,000</p> <p><input type="radio"/> Between 20,000 and 40,000</p> <p><input type="radio"/> More than 40,000</p>

FIGURE 1.12 A web questionnaire using the form approach

Blaise 4.6 also supports a *form approach*. This approach is suitable for short and simple questionnaires with straightforward data entry without question routing. The Internet form is just like a paper form. There is no extra functionality. The questionnaire comprises one web page that can be scrolled up and down to answer the questions.

The web questionnaire form may be delivered in several ways. One way is to offer its URL on a website, and another way is to send a form to respondents as an attachment to an e-mail.

All questions are presented in a fixed sequence. Respondents can browse through the form and answer questions in any order. They can fill out questions off-line. There is no need for continuous contact between respondents and the Internet server while they answer questions. However, this does not mean the form cannot be filled out on-line. Figure 1.12 contains an example of a form-based web questionnaire in Blaise.

Because there is no contact between the respondent and the server, no routing or checking is possible during the interview. However, the client computer does execute range checks for entered answers.

When respondents have completed their questionnaire, they submit their form to the server (after making an Internet connection if none is present). Then the answers are sent to the Internet server. The server can execute the checking mechanism if desired and can store the data in a Blaise database. The respondent receives a confirmation from the server.

More about the development of the Blaise System and its underlying philosophy can be found in the study by Bethlehem and Hofman (2006).

1.4 Summary

Web surveys are a next step in the evolution process of survey data collection. Collecting data for compiling statistical overviews is already very old, almost as old as humankind. All through history, statistics have been used by rulers of countries to take informed decisions. However, new developments in society always have had their impact on the way the data were collected for these statistics.

For a long period, until the year 1895, statistical data collection was based on complete enumeration of populations. The censuses were mostly conducted to establish the size of the population, to determine the tax obligations of the people, and to measure the military strength of the country.

The first ideas about sampling emerged around 1895. There was a lot of discussion between 1895 and 1934 about how samples should be selected: by means of probability sampling or some other sample selection technique. By 1934 it was clear that only surveys based on probability sampling could provide reliable and accurate estimates. Such surveys were accepted as a scientific method of data collection.

Somewhere in the 1970s another significant development started. The fast development of microcomputers made it possible to introduce computer-assisted interviewing. This made survey data collection faster, cheaper, and easier, and it increased data quality. It was a time when acronyms like CATI and CAPI emerged.

The next major development was the creation of the Internet around 1982. When more and more persons and companies got access to the Internet, it became possible to use this network for survey data collection. The first Internet surveys were e-mail surveys. In 1989 the World Wide Web was developed. In the middle of the 1990s, web surveys became popular.

Web surveys are attractive because they allow for simple, fast, and cheap access to large groups of potential respondents. There are, however, also potential methodological problems. There are ample examples of web surveys that are not based on probability sampling. It is not always easy to distinguish good from bad surveys.

KEY TERMS

Blaise: A software package for computer-assisted interviewing and survey processing developed by Statistics Netherlands.

Census: A way of gathering information about a population in which every element in the population has to complete a questionnaire form.

Computer-assisted interviewing (CAI): A form of interviewing in which the questionnaire is not printed on paper. Questions are asked by a computer program.

Computer-assisted personal interviewing (CAPI): A form of face-to-face interviewing in which interviewers use a laptop computer to ask the questions and to record the answers.

Computer-assisted self-administered questionnaires (CSAQ): A form of data collection in which respondents complete the questionnaires on their own computer. See also CASI.

Computer-assisted self-interviewing (CASI): A form of data collection in which respondents complete the questionnaires on their own computer. See also CSAQ.

Computer-assisted telephone interviewing (CATI): A form of telephone interviewing in which interviewers use a computer to ask the questions and to record the answers.

E-mail survey: A form of data collection via the Internet in which respondents are sent a questionnaire that is part of the body text of an e-mail. The questionnaire is completed by returning the e-mail after answering the questions in the text.

Face-to-face interviewing: A form of interviewing where interviewers visit the homes of the respondents (or another location convenient for the respondent). Together, the interviewer and the respondent complete the questionnaire.

Mail survey: A form of data collection where paper questionnaire forms are sent to the respondents. After completion of the questionnaires, they are returned to the research organization.

Purposive sampling: A form of nonprobability sampling in which the selection of the sample is based on the judgment of the researcher as to which elements best fit the criteria of the study.

Probability sampling: A form of sampling where selection of elements is a random process. Each element must have a positive and known probability of selection.

Quota sampling: A form of purposive sampling in which elements are selected from the population in such a way that the distribution of some auxiliary variables matches the population distribution of these variables.

Random digit dialing (RDD): A form of sample selection for a telephone survey where random telephone numbers are generated by some kind of computer algorithm.

Representative method: A method proposed by Anders Kiaer in 1896 to select a sample from a population in such a way that it forms a “miniature” of the populations.

Straw poll: An informal survey conducted to measure a general feeling of a population. Sample selection is such that it usually does not allow us to draw conclusions about the population as a whole.

Survey: A way of gathering information about a population in which only a sample of elements from the population has to complete a questionnaire form.

Telephone interviewing: A form of interviewing in which interviewers call selected persons by telephone. If contact is made with the proper person, and this person wants to cooperate, the interview is started and conducted over the telephone.

Web survey: A form of data collection via the Internet in which respondents complete the questionnaires on the World Wide Web. The questionnaire is accessed by means of a link to a web page.

EXERCISES

Exercise 1.1. Which of the following options is not an advantage of computer-assisted interviewing (CAI) as compared with traditional modes of data collection

- a. Data quality is higher because of included checks.
- b. The software is in charge of routing through the questionnaire.
- c. CAI leads to higher response rates.
- d. Data can be processed quicker.

Exercise 1.2. What is an advantage of an e-mail survey over a traditional mail survey?

- a. Data quality is higher because of included checks.
- b. There is less undercoverage.
- c. Response rates are higher.
- d. It has better facilities for navigation through the questionnaire.

Exercise 1.3. Why were the first surveys on the Internet e-mail surveys and not an the web surveys?

- a. E-mail surveys were cheaper.
- b. The World Wide Web did not exist yet.
- c. E-mail surveys are more user-friendly.
- d. E-mail surveys require less data communication over the Internet.

Exercise 1.4. When should the form-based approach be preferred over the question-by-question approach in a web survey?

- a. The questionnaire is very long.
- b. The questionnaire contains route instructions and edits.
- c. All questions fit on one screen.
- d. The survey is a business survey.

Exercise 1.5. Which of the four features is typically an advantage of web surveys;

- a. There is no undercoverage.
- b. The sample size is always large.
- c. A survey can be designed and conducted very quickly.
- d. Accurate estimates can always be computed.

Exercise 1.6. How can the problem of undercoverage in a general population web survey be avoided?

- a. Conduct a mixed-mode survey.
- b. Increase the sample size.
- c. Conduct a self-selection web survey.
- d. Replace the web survey by an e-mail survey.

REFERENCES

- Bethlehem, J. G. (1987), The Data Editing Research Project of the Netherlands Central Bureau of Statistics, *Proceedings of the Third Annual Research Conference of the US Bureau of the Census*, U.S. Bureau of the Census, Washington, DC, pp. 194–203.
- Bethlehem, J. G., Cobben, F., & Schouten, B. (2011), *Handbook on Nonresponse in Household Surveys*. John Wiley & Sons, Hoboken, NJ.
- Bethlehem, J. G. (2009), *The Rise of Survey Sampling*. Discussion Paper 09015, Statistics Netherlands, The Hague/Heerlen, the Netherlands.
- Bethlehem, J. G. & Hofman, L. F. M. (2006), Blaise—Alive and Kicking for 20 Years. *Proceedings of the 10th International Blaise Users Conference*, Arnhem, the Netherlands, pp. 61–86.
- Bethlehem, J. G. & Stoop, I. A. L. (2007), Online Panels—A Theft of Paradigm? The Challenges of a Changing World. *Proceedings of the Fifth International Conference of the Association of Survey Computing*, Southampton, U. K., pp. 113–132.
- Beukenhorst, D. & Wetzels, W. (2009), *A Comparison of Two Mixed-mode Designs of the Dutch Safety Monitor: Mode Effects, Costs, Logistics*. Technical Paper DMH 206546, Statistics Netherlands, Methodology Department, Heerlen, the Netherlands.
- Bowley, A. L. (1906), Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society*, 69, pp. 548–557.
- Bowley, A. L. (1926), Measurement of the Precision Attained in Sampling. *Bulletin of the International Statistical Institute*, XII, Book 1, pp. 6–62.
- CBS (1948), Enige Beschouwingen over Steekproeven. Reprint from *Statistische en Economische Onderzoekingen* 3, Statistics Netherlands, The Hague, the Netherlands.
- Clayton, R. L. & Werking, G. S. (1998), Business Surveys of the Future: The World Wide Web as a Data Collection Methodology. In: Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls, W. L., & O'Reilly, J. (eds.), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.
- Cobben, F. & Bethlehem, J. G. (2005), *Adjusting Undercoverage and Non-response Bias in Telephone Surveys*. Discussion Paper 05006. Statistics Netherlands, Voorburg/Heerlen, the Netherlands.

- Cochran, W. G. (1953), *Sampling Techniques*. John Wiley & Sons, New York.
- Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls II, W. L., & O'Reilly, J. M. (eds.) (1998), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.
- Couper, M. P., Blair, J., & Triplett, T. (1999), A Comparison of Mail and E-mail for a Survey of Employees in U. S. Statistical Agencies. *Journal of Official Statistics*, 15, pp. 39–56.
- Couper, M. P. & Nicholls, W. L. (1998), The History and Development of Computer Assisted Survey Information Collection Methods. In: Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls, W. L., & Reilly, J. (eds.), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.
- Deming, W. E. (1950), *Some Theory of Sampling*. John Wiley & Sons, New York.
- Den Dulk, K. & Van Maarseveen, J. (1990), The Population Censuses in The Netherlands. In: Van Maarseveen, J. & Gircour, M. (eds.), *A Century of Statistics, Counting, Accounting and Recounting in The Netherlands*. Statistics Netherlands, Voorburg, the Netherlands.
- Dillman, D. A. (2007), *Mail and Internet Surveys: The Tailored Design Method*. John Wiley & Sons, Hoboken, NJ.
- Graunt, J. (1662), *Natural and Political Observations upon the Bills of Mortality*. Martyn, London, U. K.
- Hansen, M. H., Hurvitz, W. N., & Madow, W. G. (1953), *Survey Sampling Methods and Theory*. John Wiley & Sons, New York.
- Horvitz, D. G. & Thompson, D. J. (1952), A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, pp. 663–685.
- Kiaer, A. N. (1895), Observations et Expériences Concernant des Dénombréments Représentatives. *Bulletin of the International Statistical Institute*, IX, Book 2, pp. 176–183.
- Kiaer, A. N. (1997 reprint), Den Repräsentative Undersøkelsesmetode. *Christiania Videnskabselskabets Skrifter. II. Historiskfilosofiske klasse, Nr 4* (1897). English translation: *The Representative Method of Statistical Surveys*, Statistics Norway, Oslo, Norway.
- Kiesler, S. & Sproull, L.S. (1986), Response Effects in the Electronic Survey. *Public Opinion Quarterly*, 50, pp. 402–413.
- Kuusela, V., Vehovar, V., & Callegaro, M. (2006), *Mobile phones—Influence on Telephone Surveys*. Paper presented at Second International Conference on Telephone Survey Methodology, Miami, FL.
- Laplace, P. S. (1812), *Théorie Analytique des Probabilités. Oeuvres Complètes*, Vol. 7, Gauthier-Villar, Paris, France.
- Lienhard, J. H. (2003), *The Engines of Our Ingenuity, An Engineer Looks at Technology and Culture*. Oxford University Press, Oxford, U. K.
- 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, pp. 558–606.
- Nicholls, W. L. & Groves, R. M. (1986), The Status of Computer Assisted Telephone Interviewing. *Journal of Official Statistics*, 2, pp. 93–134.
- NIPO, (1946), Eerste Telefonische Enquête in Nederland Verricht Door NIPO. *De Publieke Opinie*, 1^e jaargang, No. 4, p. 1.

- O'Connell, P. L. (1998), Personal Polls Help the Nosy Sate Curiosity. *New York Times*, June 18.
- Quetelet, L. A. J. (1835), *Sur l'Homme et le Développement de ses Facultés, Essai de Physique Sociale*. Paris, France.
- Quetelet, L. A. J. (1846), *Lettre à S.A.R. le Duc Régant de Saxe Coburg et Gotha sur la Théorie des Probabilités, Appliquée aux Sciences Morales et Politiques*. Brussels, Belgium.
- Roos, M., Jaspers, L., & Snijkers, G. (1999), *De Conjunctuurtest via Internet*. Report H4350–99-GWM, Statistics Netherlands, Data Collection Methodology Department, Heerlen, the Netherlands.
- Roos, M. & Wings, H. (2000), Blaise Internet Services Put to the Test: Web-surveying the Construction Industry. *Proceedings of the 6th International Blaise Users Conference*, Kinsale, Ireland.
- Saris, W. E. (1998), Ten Years of Interviewing Without Interviewers: the Telepanel. In: Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls II, W. L., & O'Reilly, J. M. (eds.), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York, pp. 409–430.
- Schaefer, D. R. & Dillman, D. A. (1998), Development of a Standard E-mail Methodology: Results of an Experiment. *Public Opinion Quarterly*, 62, pp. 378–397.
- Snijkers, G., Tonglet, J., & Onat, E. (2004), *Projectplan Pilot e-PS*. Internal Report H3424–04-BOO, Development and Support Department, Division of Business Statistics, Statistics Netherlands, Heerlen, the Netherlands.
- Snijkers, G., Tonglet, J., & Onat, E. (2005), *Naar een Elektronische Vragenlijst voor Productiestatistieken*. Internal Report, Development and Support Department, Division of Business Statistics, Statistics Netherlands, Heerlen, the Netherlands.
- Utts, J. M. (1999), *Seeing Through Statistics*. Duxbury Press, Belmont, CA.
- Witte, J. C., Amoroso, L. M., & Howard, P. E. N. (2000), Method and Representation in Internet-based Survey Tools. *Social Science Computer Review*, 18, pp. 179–195.
- Yates, F. (1949), *Sampling Methods for Censuses and Surveys*. Charles Griffin & Co, London, U. K.

About Web Surveys

2.1 Introduction

The Internet is one data collection tool that is available for conducting surveys. It is a relatively new method. At first sight, it is an attractive means of data collection because it offers a possibility of collecting a large amount of data in a short period of time at a low cost. Therefore, web surveys have quickly become popular.

The methodology of web surveys has not yet been fully developed. One should realize that only after a sound, scientifically valid theory of web surveys has been established, can one determine whether the advantages outweigh the potential problems. Therefore, the statistical studies described in this handbook are of critical importance for the future of web surveys.

Traditionally, surveys can be carried out using various modes of data collection:

- By mail using paper questionnaire forms.
- By telephone. The interviewer can use a paper form or a computer program for computer-assisted interviewing (CATI, computer-assisted telephone interviewing).
- Face-to-face. The interviewer can use a paper form or a computer program for computer-assisted interviewing (CAPI, computer-assisted personal interviewing).

Web surveys resemble mail surveys. Both modes of data collection rely on visual information transmission. Note that telephone surveys and face-to face surveys

use oral information transmission. Furthermore, no interviewers are involved in data collection. Data collection is based on self-administered interviews.

Of course, a web survey is a computer-assisted form of data collection (like CAPI and CATI). Therefore, it is sometimes also called computer-assisted web interviewing (CAWI). Web survey questionnaire forms can include features like automatic routing through the questionnaire and automatic checking for inconsistencies. These features are not possible for mail survey questionnaires.

Like for any other survey, web survey respondents also have to be contacted first and invited to participate in the survey. In general, the following approaches are possible:

- Send an e-mail with a link to the website containing the survey questionnaire. The link may include a unique identification code. The unique code ensures that a respondent will complete the questionnaire only once. It also ensures that only selected individuals complete the questionnaire.
- Send a letter by ordinary mail inviting the potential respondents to go to the survey website. The letter contains the address (URL) of the website and a unique code. Again, this guarantees that only the proper individuals participate in the survey.
- Catch potential survey participants on the Internet when they are visiting a website. They are invited to click on a link or button to start the survey. They may be directed to a different website containing the survey, or the survey starts as a newly opened window (popup window) on the screen. The web survey may also be embedded in a website visited by the individual.

The third approach is a very simple way to conduct a survey. No e-mails or letters have to be sent. However, it has the disadvantage that no proper sampling procedure is used. This may lead to a response that lacks representativity. Sometimes such surveys may allow a respondent to complete the questionnaire more than once. Moreover, there is no guarantee that each respondent is a member of the intended survey population. Finally, technical aspects, like popup blockers, may prevent starting the survey questionnaire.

EXAMPLE 2.1 A web survey on technological communication and links between enterprises

This survey was carried out within the survey research activities of the Centro Analisi Statistiche e Indagini (CASI) at the Department of Mathematics, Statistics and Informatics of the Faculty of Economics of the University of Bergamo. The survey topics were the use of e-commerce, collaboration with other enterprises, and/or the belonging to groups, markets, and employment. The questionnaire was kept simple (6 pages of

40 substantial items, 1 welcome page, and 1 final page) and asked mainly for qualitative answers and percentage data.

The survey was directed toward approximately 2,000 firms of the provinces of Bergamo (used as the pilot province), Brescia, Lecco, Varese, and Mantova (each province is in the Lombardy region in Italy) in the manufacturing and building sector. E-mail addresses and stratification variables have been provided from administrative databases of the Chamber of Commerce for the Bergamo province and from a Unioncamere (Union of Chambers of Commerce) database for the other provinces. The database for the e-mail list and stratification variables has been obtained from the administrative records of the same institutions.

The overall response rate was 21.9%, which is high considering that as a result of the quality of the list, 12% of the follow-up contacts were explicitly wrong e-mail addresses. Response rates by size of the firm, legal form, and economic activity did not differ very much. It is interesting to note that for small firms (less than 20 employees), the response rate was high compared with other firms' sizes.

Data collection took place in spring 2000 and was carried out according to the following steps:

- Invitation to participate in the survey was sent by e-mail (survey presentation letter, a survey report as incentive, and other related advantages were prospected).
- A link of the individual firm address for completion of the questionnaire was included in the presentation letter. Therefore, no identification code (id) and no password were required.
- Three e-mail reminders were sent (the first reminder was 14 days after the survey follow-up, and the remaining two reminders were sent with weekly periodicity). Mainly for research purposes a fourth e-mail reminder was sent after the end of the survey period. As described in the literature, the three reminders were effective in improving the response rate, and the fourth reminder did not have an effect.

The other chapters in this book focus mostly on the theory and application of web surveys. A web survey questionnaire consists of one or more web pages. Respondents have to visit this website to answer the survey questions. Note that there also other types of data collection that use the Internet. For example, it is possible to deploy the Internet only as a medium to transport the empty questionnaire to the respondent and to transport the completed questionnaire back from the respondent to the survey agency. For example, a simple questionnaire form is implemented in a Microsoft Excel spreadsheet (Microsoft Corporation, Redmond, WA). The respondents receive the spreadsheet as an attachment to an e-mail. After they have downloaded this electronic form, they fill it in on their

own computer. After completion, they return the form, also by e-mail. This type of survey is called an *Internet survey* because it uses the Internet in a much broader sense than just the HTML pages of the World Wide Web.

This chapter describes the various forms of online data collection, from simple e-mail surveys to advanced web surveys. It shows how web surveys can be used for different target populations, for cross-sectional data collection, and for longitudinal data collection (panels). It discusses the main reasons for online data collection, the advantages and disadvantages, areas of application, and specific related problems.

2.2 Theory

Collecting data using a web survey has much in common with other modes of data collection. There are the usual steps, like survey design, fieldwork, data processing, analysis, and publication. At each step, however, the suitability of concepts and methods taken from traditional survey approaches (face-to-face, paper, and telephone) need to be considered and, where necessary, adapted. This handbook examines the most important practical and methodological aspects of web surveys that need careful consideration. They can be translated in the following questions:

- How should the sample be selected?
- How should potential respondents be contacted?
- How should a web questionnaire be constructed?
- How can proper statistical inference be made based on a web survey data?
- What is the impact of sampling and nonsampling errors?
- How should web panels be handled?

These questions will be answered in detail in the other chapters of this handbook. The current chapter provides a general overview of different approaches of conducting web surveys. For each approach and each situation, different problems may occur and, therefore, different methodological solutions are needed.

2.2.1 TYPICAL SURVEY SITUATIONS

In this section, typical situations are identified in which a web survey can be conducted. These situations are determined by several different key aspects that lead to different survey situations:

- *Target population.* There are general population surveys (among individuals or households), business surveys, specific population surveys (among specific populations like company employees, company customers, students at a university or school, or members of club) or open population surveys (among ill-defined populations like consumers of a product or service).

- *Survey administrator.* This can be a national statistical institute (NSI) or other official statistical government body, a commercial market research company, a university, or another research institute.
- *Cross-sectional versus longitudinal data collection.* A cross-sectional web survey measures the status of a population at one specific point in time, based on a sample selected for that purpose. A longitudinal web survey (or web panel) is recruited and maintained to allow measuring change over time. Also, surveys on specific topics can be selected from the web panel.
- *Technical implementation.* The questionnaire can be designed as a website on the World Wide Web. In this case, questionnaires are completed on-line. It is also possible to use the Internet as just a vehicle to transport a questionnaire form to the respondents. For example, a form in an Excel spreadsheet can be sent as an attachment to an e-mail. In this case, the questionnaire is completed off-line. This is an example of an e-mail survey.

The choice of the *target population* for a web survey may have an impact on the magnitude of survey errors and, particularly, on nonsampling errors. Such errors may partly be caused by selection problems (for example, undercoverage) and partly by measurement errors (resulting from the lack of interviewer assistance).

If the target population is the general population (households or individuals), there is a problem with the sampling frame. Some countries have a population register. Such a register contains addresses. Therefore, it can be used as a sampling frame for a face-to-face or mail survey. Sometimes telephone numbers can be linked to addresses, which makes it possible to use it as a sampling frame for telephone surveys. Unfortunately, these registers do not contain e-mail addresses, nor can e-mail addresses be linked to it.

Internet penetration varies greatly between countries (see Chapter 8 regarding undercoverage problems). Currently, Internet coverage is relatively high, say between 60% and 90%, in several countries. These coverage rates seem to suggest that general population web surveys are possible in these countries, and that they can compete with traditional data collection modes. However, it should be noted that a large Internet penetration does not imply high Internet use. Moreover, it also does not imply that high-quality fast Internet connections are available. For example, not everyone with Internet access has broadband.

One should always bear in mind that not everyone has access to the Internet. One example is that not every employee of a company is allowed to use the Internet. Moreover, Internet access rates are substantially lower in many countries for specific subpopulations. For example, Hispanic blacks are underrepresented in the United States. Another example of underrepresented groups are people with low education and people living in rural areas. This situation is encountered in many countries. Also, the elderly are often underrepresented among Internet users. Undercoverage leads to web surveys that lack representativity. Therefore, there is a risk that wrong conclusions are drawn from the survey results.

If the target population consists of businesses, it is probable, in most countries, that each business has Internet access and therefore has an e-mail

address. Thus, the collection of business that can be sampled for a web survey is close to the target population. However, obtaining a complete list of e-mail addresses for businesses may be a difficult task. Partial lists can be sometimes found, but complete lists are often lacking. NSIs regularly contact large enterprises for surveys. Therefore, they may have a complete list of e-mail addresses for certain economic branches or specific size classes of companies. In most cases, obtaining an e-mail population list for small enterprises and businesses could well be a difficult task for NSIs. And even if such a list is available, it may also require a lot of effort to maintain it.

If the target population is a closed population (employees of a company, or students of university), there often is a sampling frame containing the e-mail addresses of all members of the population. In such situations, there is no difference between the target population and the sampling frame. There are no coverage problems. This is the ideal case for a web survey.

With respect to the *survey administrator*, differences may occur with respect to the amount of information for setting up the surveys, and the topics that are addressed in the web survey. National statistical institutes and official statistics bodies probably have the largest amount of information available on the general population (households and individuals) and businesses (or institutions). They may have access to population registers, they may have census data, and they may manage demographic databases, the business register, and other sources of information. Therefore, although this huge amount of data may be insufficient for generating a sampling frame of e-mail addresses for the target population, they may well be in a fairly good position to obtain this information in the near future if web surveys gain importance and if the Internet penetration within the population continues to increase. Currently, the advantages of the NSIs are twofold:

- They often have a sampling frame for the general population of individuals of households based on addresses. This means that they can select a suitable probability sample from the target population. They can contact the selected persons/households using an alternative mode (for example, by mail), while using the web mode as a second step in the data collection process.
- With regard to businesses, they have a full population list, at least for large businesses, together with contact points and, probably, e-mail addresses. Therefore, an adequate web sampling frame is available.

EXAMPLE 2.2 The Information and Computer Technology (ICT) Survey Pilot

Statistics Netherlands carried out a pilot with the ICT survey to find out whether it was possible to use the web for data collection. This survey collects information on the use of computers and the Internet in households and by individuals. The regular ICT survey was a CATI survey. The

survey was fairly expensive. It also suffered from undercoverage because the sample was selected from the telephone directory. Households with unlisted numbers and mobile-only households could not be selected.

The sample for the pilot was selected from the population register. So there was no undercoverage. All persons in the sample received an invitation letter by mail. The letter contained the Internet address of the survey and a unique log-in code. Respondents had the possibility of completing the questionnaire on paper. To prevent those with Internet from responding by paper, the paper questionnaire was not included in the invitation letter. People had to apply for the paper form by returning a stamped return postcard.

After one week, a postcard was sent to all nonrespondents with a reminder to complete the survey questionnaire, either by web or mail. Two weeks after receipt of the invitation letter, the remaining nonrespondents were approached again. Part of these nonrespondents received a reminder letter, and another part was called by telephone (if a telephone number was available). The telephone call was just to remind the nonrespondents and did not replace the paper/web questionnaire form.

It turned out the postcard reminders worked well. Each time they were sent, there was a substantial increase in response. The telephone reminder did not work as well as the postcard reminder. Of the people that promised by telephone to fill in the form, only 40% actually did so.

Other survey administrators, such as academic researchers, market research companies, or private businesses, may not have proper sampling frames available. One solution to this problem could be to let an NSI select the sample for them. Another could be to obtain a copy of the sampling frame after privacy-related information has been removed. Nevertheless, privacy-related laws may prevent NSIs from making sampling frame information available to third parties.

With respect to the topic of the survey, it should be borne in mind that NSIs and other government statistical bodies collect data primarily for policy decisions. There may be different surveys for different social and economic indicators. Many surveys are compulsory, which means that the contacted elements are obliged to respond. If they do not, they may be fined. Sometimes questionnaires are complex because many topics are covered. Surveys conducted by academic researchers, market research organizations, and other companies tend to be more heterogeneous, covering several different issues: product characteristics, customer satisfaction regarding products and services, employee satisfaction, trends in consumer preferences or behavior, health, use of technological products, and so on. Generally speaking, survey topics dealt with by this type of survey administrator are mainly devoted to a more or less traditionally defined target population, and therefore, an appropriate survey frame definition becomes more difficult. Surveys carried out by this type of survey administrator could often make use of simpler and shorter questionnaires.

With respect to the distinction between *cross-sectional* and *panel* data collection, it should be noted that cross-sectional surveys gather data about one moment in time, whereas panel surveys collect information at many successive points in time with the focus on investigating changes over a period of time. Panel surveys are discussed in Chapter 12. The main problem with panel surveys is the lack of representativity of the panel and of the samples selected from it. In fact, a panel is a large group of elements that is as large as the population, or representative of this population. The real world is full of panels that have been recruited by means of self-selection and therefore not representative for the population.

If a panel is used for longitudinal studies, all respondents must be tracked back to the moment they entered the panel. Therefore, when doing a longitudinal analysis of survey results, DiSogra and Callegaro (2009) recommend computing cumulative standardized response rates (taking into account different recruitment waves; i.e., rates based on a multiple recruitment approach). This approach captures the dynamics of the history of a panel member with regard to nonresponse and attrition.

With regard to *technical implementation of the questionnaire*, there are two approaches possible: on-line data collection and off-line data collection:

- *On-line data collection* is a way of data collection for which the respondents have to remain on-line during the process of answering the questions. The questionnaire is implemented as one or more web pages. The respondent has to surf to the survey website in order to start the questionnaire. The questionnaire can be question based or form based. *Question based* means that every web page contains a single question. After answering a question, the respondent proceeds to the next question that is on the next page. The page-based approach is advised if the questionnaire contains routing instructions and consistency checks. *Form based* means that there is a single web page containing all questions. This page looks like a form. Usually there are no routing instructions and no consistency checks.
- *Off-line data collection*. The electronic questionnaire form (an HTML-page, an Excel spreadsheet, or another interviewing software tool) is sent to the respondent by e-mail, or the respondent can download it from the Internet. The respondent fills in the form or spreadsheet off-line. After completing the questionnaire, it is returned (uploaded, or sent by e-mail) to the survey agency. Statistics Netherlands, for example, uses this approach for several business surveys. A computer-assisted interviewing program is sent to the selected businesses. The businesses run this program off-line and answer the questions. After completion, contact is made with the Internet again, and the data are uploaded to the survey agency.

In the case of an electronic form, the advantages of a printable questionnaire can be combined with those computer-assisted interviewing (routing and consistency checking). Note that it is also possible not to bother the respondents with

consistency checking. This means no errors will be detected during form completion. However, errors may afterward be detected by the survey agency, possibly resulting in returning the form to the respondent for error correction. This is much less efficient.

2.2.2 WHY ON-LINE DATA COLLECTION?

2.2.2.1 Advantages. Web surveys have several advantages. The three most important advantages are that web surveys are faster, simpler, and cheaper.

With respect to the time required to conduct a web survey, the following observations can be made:

- The time it takes to get in contact with the respondent can be considerably reduced if the invitation is sent by e-mail.
- Follow-ups can be carried out very quickly by e-mail. The timing of reminders can be tailored to the respondents. A typical pattern for web surveys is that many completed questionnaires are returned almost immediately. The number of returns diminishes fast after a few days. Web surveys allow for a short time lag between request and reminder than mail surveys. Biffignandi and Pratesi (2002) showed that the time interval between the first contact and the first follow-up can be shorter than in mail surveys. The intervals between successive follow-ups can also be very short, and no more than three reminders are necessary. The fourth reminder is almost ineffective. The authors suggest that ten days is an adequate time interval between first contact and first follow-up, whereas the time interval between successive reminders is around one week. Furthermore, Crawford, Couper, and Lamias (2001) showed that a quick reminder after two days works fairly well.
- The time it takes to deliver a complete questionnaire is also very short. The questionnaire can be immediately submitted as soon as it is completed. Thus, there is no time lag between the moment the respondent returns the questionnaire and the moment it is received.
- The time it takes to store the collected data is eliminated because responses are instantly submitted into a database and prepared for analysis.

To sum up, the entire data collection period is significantly shortened. All data can be collected and processed in little more than a month. There are even opinion polls on the web for which design, data collection, analysis, and publication all take place in one day.

A second advantage of web surveys is that they can be tailored to the situation. Therefore, they may make life simpler for the respondents and the researcher. Here are some examples:

- Respondents may be allowed to save a partially completed form. At a later point in time they can continue and complete the questionnaire. This is particularly

important for business surveys where sometimes different departments of the business each have to complete a specific part of the questionnaire.

- The questionnaire may be filled with already available information. The respondents only have to check them for changes. This can be useful in web panels. Examples of such preloaded data are the address of the respondent and employment status.
- There can be a facility that automatically generates an e-mail message to the survey institute if the respondent indicates he has complaints about the questionnaire. Such information can help to improve surveys and avoid future problems.
- Response rates can be monitored over time. Action can be undertaken if the response is lower than expected. For example, customized e-mail reminders are sent. However, it should be remembered that although there are no costs attached to sending reminders, a good rule is to send them at well-defined moments in time, should they be necessary. Slow respondents should not be overloaded by reminders. The literature shows that this may lead to irritation and break-off.
- The proper survey software can see to it that no respondent can complete the questionnaire more than once. Of course, this requires handing out unique identification codes to the individuals selected in the sample. Note that this does not work in case of self-selection surveys.
- Like in computer-assisted interviewing, web questionnaires may contain route instructions. These instructions see to it that respondents only answer relevant questions, and that irrelevant questions are skipped.

Web questionnaires may be improved by applying *usability testing*. Usability refers to the ease of use of a software application for a web questionnaire. Usability is measured with reference to the speed with which a task can be performed, the frequency of errors in performing a task, and user satisfaction with regard to an application interface, in terms of being easy to understand and use. Two techniques are especially valuable in usability testing:

- *Qualitative interviews*. Usability tests are typically carried out with small groups of individuals. A fully functional web questionnaire is created using current or proposed standards for the interface. A group of people is invited and considered typical of respondents. After completing the questionnaire, they get an in-person, in-depth interview. The case study reported in Section 2.3 is an example of using qualitative interviews for questionnaire testing.
- Analysis of paradata. *Paradata* are data concerning the actual web questionnaire completion process. User actions are collected by interviewers or respondents as they complete the questionnaire. The value of paradata in web questionnaire testing, and in an analysis of response behavior, is becoming more and more significant. Information on the characteristics of the respondent's technical environment, respondent response time, errors

made, and navigation behavior help to detect and correct problems in the questionnaire.

Survey data collection is expensive. Because many statistical agencies of governments face budget cuts and other survey organizations attempt to reduce data collections, it is worth while to consider cheaper modes of data collection. Web surveys are substantially cheaper than other modes of data collections.

Of course, a web survey requires initial investment in computers, servers, and software. Additionally, there are initial costs for the sampling design (if a probability sampling approach is adopted) and for web questionnaire design and implementation. Skilled and specialized personnel, with an understanding of usability and visual design, are needed to design and implement a web survey.

After these survey steps are over, there are no further data collection costs other than the costs of help desk personnel. Such a help desk is important to answer respondents' questions or to solve their problems (Lozar Manfreda and Vehovar, 2008).

Field data collection is relatively cost-free and not dependent on the number of questionnaires that are administered and completed. The database with survey data is automatically generated, making data input costs irrelevant as well. Time and effort related to data entry and verification is eliminated. For a comparison of the timing of return rates in mail and web surveys, see the study by Dillman, Smith, and Christian (2009).

To sum up, large numbers of completed questionnaires can be collected in a very short time and at low costs.

Web surveys also have some other attractive properties worth mentioning:

- The response burden may be easily monitored in web surveys because of the possibility of obtaining server-side and client-side information. This makes it possible to record how much time respondents need to complete the questionnaire. Analysis may show how the response burden is related to the response rate.
- The response burden can be reduced by using short questionnaires. It may help to split a large questionnaire into several small questionnaires. The small questionnaires can be offered at different moments in time. This does not increase the costs of the survey.
- Web surveys are less intrusive, and they suffer less from social desirability effects.
- Geographical boundaries are not a problem. Web surveys are not limited by geography in the same way as face-to-face interviews and mail and telephone surveys. Therefore, international target populations may be easily reached without special additional costs or time delays.

2.2.2.2 Disadvantages and Problems. A major problem of web-based surveys is sample selection. For research applications, a random sample is desirable and often essential, and researchers may simply not have a comprehensive

sampling frame of e-mail addresses for people who drink fruit juices or go to church. Despite the huge growth of the Internet, there are still many people who do not have access to, or choose not to use, the Internet. There are also wide disparities in Internet access among ethnic, socioeconomic, and demographic groups. A sampling frame, including the e-mail addresses, of all members of the target population should be available to draw a random sample. In practice, this list is rarely available. Therefore, large coverage problems develop, and this is the most relevant issue.

Sampling problems in particular may be an issue for general-population surveys. For many specific populations, there are no problems. Examples include companies collecting customer satisfaction data, employers measuring job satisfaction, educators collecting course evaluations and conducting examinations, bloggers wanting to consult with their readers, and event organizers checking proposed attendance and meal and other preferences. Although there is still a need for some caution, in terms of learning how to use the new technology with confidence, the use of web surveys is growing rapidly and will clearly continue to grow.

A disappointing aspect of web surveys is that they do not contribute to solving the problem of decreasing response rates. They usually result in low response rates, with little chance of obtaining higher rates. It should be noted that, despite low response rates, the use of server-side and client-side paradata can help to focus efforts on specific population that most need it.

2.2.3 AREAS OF APPLICATION

Web surveys may be used in any field of application provided that the elements in the population have Internet access, and that they have some basic computer skills. In some cases, as described in detail in Chapter 12, a probability-based sample from the general population is selected. Then some people without a computer may receive one (with Internet access), together with basic instructions for use. This solution typically has been adopted for general-population web panels.

If the web data collection is possible for all potential respondents, a web survey can be a very useful data collection tool, combining low costs and high quality.

Unfortunately, often not all elements in the target population have Internet access, computers with adequate processing power to process questionnaires, or sufficient computer literacy. This problem holds true for general-population surveys as well as for many others possible target populations. Even if Internet penetration is growing, differences may exist between countries and between groups within populations. Large differences in computer equipment, screen settings, and technical literacy may have a substantial impact. Thus, to carry out a good web survey, a statistical sound approach is needed that attempts to minimize a possible bias as much as possible. And if a bias cannot be avoided, there should be statistical techniques applied to correct for this bias afterward.

In practice, despite the methodological challenges, many surveys (especially commercial surveys) are conducted on the web without properly taking into

account the impact on the reliability of the results originating from the lack of Internet coverage and/or lack of computer literacy. Such surveys are administered exclusively via the web and, therefore, reach only one part of the target population. When using web survey results, one should be aware of potential problems. Therefore, it is important to assess the quality of the web survey by analyzing the methodological description in the documentation.

EXAMPLE 2.3 Reliable web surveys

Certain surveys are not affected by the instrument bias a web survey may cause. When measuring job satisfaction among high-tech workers, the bias will be minimal. Getting feedback from employees on a benefit package can have a slightly higher bias if not all employees have computer access. However, an attempt to determine what role the United States should have played in the Libyan conflict of 2011 would probably produce highly biased estimates because one would only obtain the opinions from computer-literate people with Internet access. They will not be representative for the whole population.

Web surveys may be conducted for profiling purposes. Examples are member surveys, audience profiling, and donor profiles. Web surveys may also be used for data collection by asking people to provide information about themselves. Other interesting applications are social-economic research, planning support, and social behavior studies. A web survey can also be used for attitude polls, opinion polls, program evaluation, and community cultural planning surveys. Other possible topics are economic aspects and performances, as well as market trends and customer/employees satisfaction.

EXAMPLE 2.4 The Kauffman Firm Survey (KFS)

The Kauffman Firm Survey (KFS) is a panel study of new businesses founded in 2004. They are tracked over their early years of operation. The survey focuses on the nature of new business formation activities, characteristics of the strategy, offerings, employment patterns of new businesses, the nature of the financial and organizational arrangements of these businesses, and the characteristics of their founders.

The KFS created the panel by using a random sample from a Dun & Bradstreet (D&B) database list of new businesses started in 2004. The list contained in total approximately 250,000 businesses.

The KFS oversampled these businesses based on the intensity of research and development employment in the businesses' primary industries. The KFS sought to create a panel that included new businesses created by a person or team of people, purchases of existing businesses by a

new ownership team, and purchases of franchises. To this end, the KFS excluded D&B records for businesses that were wholly owned subsidiaries of existing businesses, businesses inherited from someone else, and not-for-profit organizations. Also, previous research on new businesses showed variability in how business founders perceive the operation of tier starting businesses. Therefore, a series of questions was asked of business owners about indicators of business activity and whether these were conducted for the first time in the reference year (2004). These indicators were payment of state unemployment (UI) taxes, payment of Federal Insurance Contributions Act (FICA) taxes, presence of a legal status for the business, use of an Employer Identification Number (EIN), and use of a Schedule C to report business income on a personal tax return.

To be “eligible” for the KFS, at least one of these activities had to have been performed in 2004. A random sample of 32,469 businesses was selected for data collection on the baseline survey, which was conducted between July 2005 and July 2006. A total of 6,030 businesses were identified as eligible. Interviews were completed with principals of 4,928 businesses that started operations in 2004. A self-administered web survey and computer-assisted telephone interviewing (CATI) were used for data collection. The KFS respondents were paid \$50 to complete the interview. The CATI response accounted for 3,781 completed forms (77%) and the web for 1,147 forms (23%).

The KFS panel contains data during the 2004–2009 period on a large cohort of firms that began operations in 2004. It is planned to update the collected data until 2011. Because of panel attrition, the number of units is becoming slightly smaller each year. As only the 2004 cohort is under study, no continuous refreshment is planned.

The objective of the panel survey is to provide information about creation and development aspects on new business (especially of high-technology and women-owned businesses). Firm characteristics, revenue and expenses, profit and loss, owner characteristics, and since 2007, information about predominant markets and Internet sales are also collected.

This panel presents some challenges with respect to establishing whether or not changes are significant. It requires more research.

2.2.4 TRENDS IN WEB SURVEYS

Dillman, Smith, and Christian (2009) describe some changes in the survey environment from 1970 to 2000, focusing on the factors like human interaction, trust that the survey is legitimate, time involvement for each respondent, attention given to each respondent, respondent control over access, and respondent control over whether to respond. These observations indicates that during the 1990s, human interaction and individual contact relevance are decreasing as a result of the use of IT (i.e., computer-aided and web surveys) and massive use of e-mails.

Trust on survey relevance and legacy is very low, and the possibility of refusal and of filtering against surveys (anti-spam, disclosure rules) is very high. These observations are in line with developments with respect to web surveys.

More recently, Biffignandi (2010a, 2010b) has been focusing on major trends with respect to web surveys. The author underlines that survey methodology has recently undergone a paradigm shift: The focus has shifted to the causes of survey errors and how to prevent them. This new paradigm stresses the total error concept, which includes sampling error (the central error in the traditional approach) and nonsampling error. Nonsampling errors can be numerically larger than sampling errors. The new paradigm takes into account the following aspects:

- All kinds of events and behavior occurring during the survey process are to taken into account.
- The overall response rate is only a very simple measure of survey quality, although it is frequently used as an indicator. To some extent, this measure could be useful in identifying weak points in the process (for instance, a large amount of refusal might be a result of a bad-contact process), but it fails to consider that people who have web access, or that are respondents to a web survey, could significantly differ from other units.
- Overall response rates do not provide information regarding the response propensity of different respondent subgroups (late respondents versus early respondents, and sociodemographically different subgroups), or on respondent behavior.
- Response rates are anyway becoming very low, and there is a need to investigate the reasons for this and to find solutions.
- With respect to sampling error, because of imperfect frames in web surveys, traditional probabilistic samples are in many cases not easy to implement. Therefore, a sampling error cannot be computed for the survey results, as the theory of statistical inference does not apply.

As a consequence of this new paradigm, attention is going to be paid to:

- How to face decreasing response rates. Possible solutions may be:
 - Keeping respondents focused on the relevant parts of the computer screen, and keeping distractions to a minimum can help to complete the questionnaire. To accomplish this task, studies based on eye-tracking analysis are to be carried out.
 - An interesting strategy for improving response rates is to use mixed-mode surveys. However, new problems originate with the mixed approach because mode effects are to be taken into account in analyzing survey results. Occurrence and treatment of mixed-mode effects need further investigation.
- How to use paradata (i.e., data are collected during the interviewing process). Increased attention is going to be devoted to the analysis of this type of data. In particular, they are used to identify typologies of response behavior

explaining the potential variations in participation in web-based surveys and providing a valuable insight into understanding nonresponse and various aspects of response behavior. From the methodological point of view, behavioral analyses rely on the Cognitive Aspects of Survey Methodology movement (CASM), and in many empirical studies, the theory of planned behavior (TPB) model is applied (Ajzen, 1991). The main objective is to obtain a more comprehensible picture of how intentions are formed. For example, based on TPB, two alternative models were empirically tested, in which the roles of trust and innovativeness were theorized differently—either as moderators of the effects that perceived behavioral control and attitude have on participation intention (moderator model) or as direct determinants of the attitude, perceived behavioral control, and intention (direct effects model).

- How do you get representative web surveys and/or panels? Many access panels consist of volunteers, and it is impossible to evaluate how well these volunteers represent the general population. In any case, they represent a nonprobability sample. Recent research attempts to tackle the task of how to apply probabilistic recruitment to panels and how to draw inferences from them are present in recent literature. One approach to correct for a lack of representativity is, for example, to apply propensity scoring methodology. Propensity scores have been used to reweight web survey results. For more about propensity scores, see chapter 11.

Generally speaking, the methodology and quality of data collected in the area of socioeconomics could greatly benefit by:

- a. The development of suitable estimation methods aimed at capturing the bias and specific variance connected with the frame characteristics and participation process of this type of survey.
- b. Research, principally based on experimental designs, allowing the effects of various factors to be tested (for example, the effects of different types of question structure, various contact modes, etc.).
- c. Research, based on behavioral models, that allows response and participation processes to be analyzed and modeled in the context of the individual behavior of survey respondents.

2.3 Application

The Italcementi Group is a large Italian company. With an annual production capacity of approximately 75 million tons of cement, it is the world's fifth largest cement producer. The group has companies in 22 countries around the world. Italcementi regularly monitors the working conditions and working climate in the company. This is realized by means of a mixed-mode survey. Part of the data is collected by means of the web and part by paper questionnaire forms.

The target population consists of all employees of all companies belonging to the Italcementi Group around the globe. No sample is selected. All employees are invited to participate in the survey. So in principle it is a census and not a survey. There are no inference issues. Statistics can simply be computed from the data.

Nonresponse can be a problem in surveys. Some people may fail to complete the survey questionnaire because they could not be contacted (despite several attempts), they refused to cooperate, or were not able to answer the questions. This may lead to biased estimates of population characteristics. Nonresponse is not a big problem in the Italcementi survey. In many countries, almost everyone participates. See also Table 2.1.

The first step in the survey process is the design of the questionnaire. In fact, there are two questionnaires: one for blue-collar employees and one for white-collar employees. The questionnaire is organized into three main sections, covering the Italcementi Group as a whole, the company to which the employee belongs, and the specific location of the company. In addition, a section with questions about personal characteristics is included. The paper questionnaire for white-collar employees is 8 pages long, whereas that for blue-collar employees is 6 pages.

TABLE 2.1 Response rates of the Italcementi survey by country in 2007

Country	Overall response rate (%)	White-collar response rate (%)	Blue-collar response rate (%)
Albania	91.9		
Bulgaria	51.9	59.8	46.9
Dubai	100.0		
Egypt	48.9	66.1	50.4
France / Belgium	44.7 (F) 51.7 (B)	65.6	38.1
Gambia	90.7		
Greece	52.3	79.7	31.2
India	94.1	94.3	94.0
Italia	48.6	65.2	30.5
Kazakhstan	66.7	61.8	68.2
Morocco	64.4	67.4	62.9
Mauritania	98.7		
North America	43.3	76.3	21.9
Singapore	75.0		
Spain	53.7	66.2	45.1
Sri Lanka	65.5		
Thailand	78.2	84.1	72.7
Turkey	78.6	79.6	77.9
Total	54.7		

A new survey is first discussed by a management committee of the enterprise group. Next, the questionnaire is tested by means of cognitive interviews. After the questionnaire has been approved, fieldwork can start. The survey is announced by means of a letter on the intranet. It is signed by the Enterprise Group CEO. This letter is also distributed as part of pay packets. Moreover, it is displayed on the company notice boards.

It is a mixed-mode survey that includes both a web mode and a paper mode. Everyone with an e-mail address is invited to complete the questionnaire on the web. Those without Internet can complete the paper questionnaire.

The enterprise has companies in many different countries. So employees speak different languages. Language problems are avoided by translating the two questionnaires into the languages spoken in the countries in which the companies are located. For individuals with literacy problems, data were collected via the collaboration of a private voluntary organization.

For the 2007 survey, a total of 22,276 employees was eligible for the survey. In countries with a large number of employees (France, Belgium, Italy, Egypt, Thailand, and Morocco) two thirds of the white-collar employees were invited to complete the questionnaire on the web and one third were asked to complete the paper questionnaire. The paper mode was used for all white-collar employees in all remaining countries. All blue-collar workers were invited to complete the paper questionnaire. An exception was France, where also an Internet questionnaire was used for blue-collar workers.

On-line data collection took place in the period from June to September. Data collection in the field was extended until October. The results were released in December. The total number of respondents was equal to 12,183. This comes down to a response rate of 54.7%. The response rate of the white-collar workers (71.5%) was much higher than the response rate of the blue-collar workers (45%). The response rate of the web respondents was higher than the response rate of the paper questionnaire respondents. Table 2.1 shows the response rates by country.

The first survey was conducted in 2007. The survey was carried out in a similar fashion in the subsequent years. The survey turned out to be very informative with respect to the attitudes of employees about their jobs and the company they belong to.

Some examples of (translated) questions in the Italcementi survey are given as follows. Figure 2.1 contains two questions that use a so-called Likert scale.

How satisfied are you with your job? Select one answer only.	
1. Very satisfied	<input type="radio"/>
2. Rather satisfied	<input type="radio"/>
3. Rather dissatisfied	<input type="radio"/>
4. Very dissatisfied	<input type="radio"/>

Thinking about your last performance assessment, how useful would you say it was? Select one answer only.	
1. Very useful	<input type="radio"/>
2. Somewhat useful	<input type="radio"/>
3. Not very useful	<input type="radio"/>
4. Not at all useful	<input type="radio"/>

FIGURE 2.1 Questions with Likert scales

For each of the following aspects of personnel management, please indicate whether you think the organizational activities are efficient or inefficient. Select one answer in each row.

	Efficient	Inefficient
1. Administrative aspects (payroll and benefits administration, holidays)	<input type="radio"/>	<input type="radio"/>
2. Identification of training needs	<input type="radio"/>	<input type="radio"/>
3. Organization of quality training courses	<input type="radio"/>	<input type="radio"/>
4. Management of career development	<input type="radio"/>	<input type="radio"/>
5. Transfer of knowledge and skills within the company	<input type="radio"/>	<input type="radio"/>
6. Anticipating staff needs	<input type="radio"/>	<input type="radio"/>
7. Ensuring employee awareness of safety measures in the workplace	<input type="radio"/>	<input type="radio"/>
8. Controlling staff costs	<input type="radio"/>	<input type="radio"/>
9. Recruitment of new staff	<input type="radio"/>	<input type="radio"/>
10. Integration of newly recruited staff	<input type="radio"/>	<input type="radio"/>
11. Management of international careers	<input type="radio"/>	<input type="radio"/>
12. Cooperation with operational managers	<input type="radio"/>	<input type="radio"/>

FIGURE 2.2 A matrix question

They both use a 4-point scale. When responding to a Likert scale question, respondents specify their level of agreement to a statement. The scale is named after its inventor, the psychologist Rensis Likert. See also chapter 6.

Figure 2.2 contains a so-called matrix question. Such a question combines several single questions all with the same type of answer.

2.4 Summary

A web survey is a relatively new data collection technique. The spread of the Internet makes the role of the web in conducting surveys more important. At first sight, it is an attractive means of data collection because it has many advantages. Among them are costs, timeliness, and the possibility of improving survey quality. They allow for simple, fast, and cheap access to large groups of potential respondents. Therefore, web surveys have quickly become very popular. However, there are also methodological challenges like selection effects and measurement errors. There are ample examples of web surveys that are not based on probability sampling. It is not always easy to distinguish good surveys from bad.

This chapter describes the various forms of on-line data collection, from simple e-mail surveys to advanced web surveys. It shows how web surveys can be used for different target populations, for cross-sectional data collection, and for longitudinal data collection (panels). It discusses the main reasons for on-line data collection, the advantages and disadvantages, areas of application, and specific related problems.

KEY TERMS

Computer-assisted personal interviewing (CAPI): A form of face-to-face interviewing in which interviewers use a laptop computer to ask the questions and to record the answers.

Computer-assisted telephone interviewing (CATI): A form of telephone interviewing in which interviewers use a telephone to ask the questions and to record the answers.

Computer-assisted web interviewing (CAWI): A form of self-interviewing in which respondents complete the questionnaires on the Internet. CAWI is a synonym for web survey.

Cross-sectional survey. A survey that observes a sample from the target population at one point in time. The objective is to describe the state of the population at that moment in time.

Internet survey: A general term for various forms of data collection via the Internet. Examples are a web survey and an e-mail survey. Included also are forms of data collection that use the Internet just to transport the questionnaire and the collected data.

Longitudinal survey: A survey that observes the same sample from the target population at several points in time. The objective is to describe the changes of the population over time.

Paradata: Data about the process by which the survey data were collected.

Qualitative interview: An in-person, in-depth interview with respondents that have completed a survey questionnaire. Such an interview aims to uncover usability problems like difficult questions or cumbersome tasks.

Self-selection survey: A survey for which the sample has been recruited by means of self-selection. It is left to the persons themselves to decide to participate in a survey.

Usability testing: Conducting an experiment to check whether respondents find it easy to complete the web survey questionnaire. Aspects tested include the speed with which the survey task is carried, the number of errors made, and familiarity with the user interface.

Web panel: A survey in which the same individuals are interviewed via the web at different points in time.

Web survey: A form of data collection via the Internet in which respondents complete the questionnaires on the World Wide Web. The questionnaire is accessed by means of a link to a web page.

EXERCISES

Exercise 2.1. Which of the following statements does not apply to web surveys?

- a. The survey can be conducted faster.
- b. The survey can be conducted cheaper.

- c. The response rate is high.
- d. Large amounts of data can be collected.

Exercise 2.2. In what respect does a web survey resemble a mail survey?

- a. They both rely on visual information transmission.
- b. They both rely on oral information transmission.
- c. They both use computer-assisted interviewing techniques.
- d. They both cost the same amount of time to conduct.

Exercise 2.3. Which of the following phenomena is not a problem of self-selection web surveys?

- a. A respondent can complete a questionnaire more than once.
- b. Persons not belonging to the target population can complete the questionnaire.
- c. The survey results show a lack of representativity.
- d. It is difficult to get a large number of respondents.

Exercise 2.4. What is the difference between a cross-sectional survey and a longitudinal survey?

- a. A cross-sectional survey measures changes over time, and a longitudinal survey measures the state of the population at one point in time.
- b. A cross-sectional survey measures the state of a population at one moment in time, and a longitudinal survey measures time changes over time.
- c. A cross-sectional survey mainly measures facts and behavior, and a longitudinal survey measures attitudes and opinions.
- d. Any mode of data collection can be used for cross-sectional surveys, whereas longitudinal surveys can only be conducted over the Internet.

Exercise 2.5. What is off-line data collection?

- a. Any form of data collection that does not use the Internet.
- b. A form of Internet data collection for which the questionnaire is not written in HTML.
- c. An survey that uses e-mail to transfer information.
- d. A survey that uses the Internet to transfer the electronic questionnaire to the respondents.

Exercise 2.6. What is the main reason national statistical institutes consider using web surveys?

- a. It shows government also uses modern ICT surveys.
- b. It reduces nonresponse in surveys.

- c. It improves the quality of the collected data.
- d. It reduces data collection costs.

REFERENCES

- Ajzen, I. (1991), The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, 50, pp. 179–211.
- Biffignandi, S. (2010a), Modeling Non-sampling Errors and Participation in Web Surveys. *Proceedings of the 45th SIS Scientific Meeting*, Padova, Italy.
- Biffignandi, S. (2010b), Internet Survey Methodology—Recent Trends and Developments. In: Lovric, M. (ed.), *International Encyclopedia of Statistical Science*, Springer, Heidelberg, Germany.
- Biffignandi, S. & Pratesi, M. (2002), Internet Surveys: The Role of Time in Italian Firms Response Behaviour. *Research in Official Statistics*, 5, pp. 19–33.
- Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001), Web Surveys. Perceptions of Burden. *Social Science Computer Review*, 19, pp. 146–162.
- Dillman, D. A., Smith, J. D., & Christian, L. M. (2009), *Internet, Mail and Mixed-Mode Surveys. The Tailored Design Methods*. John Wiley & Sons, Hoboken, NJ.
- DiSogra, C. & Callegaro, M. (2009), Computing Response Rates for Probability-Based Web Panels. *Proceedings of Section on Survey Research of the American Statistical Association*, Washington, DC.
- Lozar Manfreda, K. & Vehovar, V. (2008), *Internet Surveys*. In: De Leeuw, E., Hox, J. J., & Dillman, D. A. (eds.), *International Handbook of Survey Methodology*, Lawrence Erlbaum Associates, New York.

Sampling for Web Surveys

3.1 Introduction

A web survey is a survey in which data are collected using the World Wide Web. As for all surveys, the aim of a web survey is to investigate a well-defined population. Such populations consist of concrete objects, like persons, households, or companies. It is typical for a survey that information is collected by means of asking questions of the representatives of the objects in the population. To do this in a uniform and consistent way, a questionnaire is used.

One way to obtain information about a population is to collect data about all its elements. Such an investigation is called a *census* or *complete enumeration*. This approach has several disadvantages:

- It is very expensive. It involves a large amount of people (for example, interviewers) and other resources.
- A census is very time-consuming. Collecting and processing a large amount of data takes time. This affects the timeliness of the results. Less timely information is less useful.
- Large investigations increase the response burden of people. As many people are more frequently asked to participate, they will experience it more and more as a burden. Therefore, they will be less and less inclined to cooperate.

The last census for England and Wales took place in March 2011. It involved around 25 million households. It was possible to complete the questionnaire

on-line. For the 2010 census in the United States, the Internet was not used. The reason was that the US Census Bureau did not expect the use of the Internet would lead to lower costs or higher response rates. Moreover, there was concern about increased security risks. This decision resulted in more than 120 million forms being mailed to every house in the address list.

In many other countries, censuses were conducted in 2010 or 2011. For example, the census of Italy took place in October 2010. Questionnaires were sent to households. The households had the choice to complete either the paper form or a form on the Internet. They could also return the form directly to a municipality official. The Internet was also used for the 2010 agricultural survey. Farmers had the choice to complete a web questionnaire or to wait for a face-to-face interview. More than 66,000 questionnaires were completed on the web.

It is interesting to note that a special Facebook page was created for the Italian census. Individuals could comment on this page about the use of the Internet data collection. Most posted comments were positive: “fast”, “accurate”, “clear”, “efficient”, and “absolutely necessary”.

A *survey* collects information on only a small part of the population. This small part is obtained by taking a *sample*. The sample only provides information on the sampled elements of the population. No information will be obtained on the nonsampled elements. Still, if the sample is selected in a scientifically sound way, it is possible to make inference about the population as a whole. “Scientifically sound” means that the sample is selected using *probability sampling*. If it is clear how this selection mechanism works and it is possible to compute the probabilities of being selected in the sample, reliable and precise conclusions can be drawn about the population as a whole. The principles of probability sampling have been successfully applied in official and academic statistics since the 1940s and, to a much lesser extent, also in commercial market research. For an overview of the history of survey sampling, see chapter 1 or the earlier study by Bethlehem (2009).

Probability samples can be selected in various ways. The choice may depend on what is available with respect to sampling frames and auxiliary information. It may also depend on what the objective of the survey is. This chapter provides an overview of sampling issues that are relevant to web surveys.

3.2 Theory

3.2.1 TARGET POPULATION

The first step in setting up a web survey is to define the *target population*. This is the population to be investigated, and to which the conclusions refer. Such a population need not necessarily consist of people. Other examples could be all IT companies in a certain state, all schools in a country, or all farms in a specific region.

The definition of the target population must be unambiguous. It must always be possible to determine in practice whether a certain element does or does not

belong to the target population. Both failure to include relevant elements in the population, and to exclude irrelevant ones, may affect the survey results.

■ EXAMPLE 3.1 A survey about road pricing

There was an intensive political discussion in the Netherlands in January 2010 about the introduction of a system of road pricing. An important participant in this discussion was the Dutch Automobile Association (ANWB). This organization conducted a poll on its website. It was a self-selection survey. The poll was supposed to determine the opinion of the ANWB members about road pricing. So, the target population consisted of all members of the association. However, everyone could participate. One could participate even more than once. There was nothing preventing this. Fortunately, all respondents were asked whether they were a member. Consequently, nonmembers could be excluded from the analysis.

Within a period of a few weeks, the questionnaire was completed more than 400,000 times. About 50,000 respondents indicated they were not an ANWB member. One should take into account that people may not have answered the question about membership properly. There is always a risk of socially desirable answers.

The target population is denoted by U . It is assumed to be finite. The number of elements N is assumed to be known. Note that this is not always the case. Examples are the number of people having access to Internet or the number of foreign visitors in a country.

The elements of the target population must be *identifiable*. It means they can uniquely be assigned sequence numbers $1, 2, \dots, N$. For each element encountered in practical situations, it must be possible to determine its sequence number. In some cases, this process is straightforward. An example is a population of persons that can be identified by means of their social security number. The target population is denoted by

$$(3.1) \quad U = \{1, 2, \dots, N\}.$$

In the survey design phase, the objectives of the survey must be translated into concrete operational procedures. This involves defining the *target variables* of the survey. These variables measure the various aspects of the phenomena to be investigated.

■ EXAMPLE 3.2 Target variables

Suppose a research organization intends to carry out a general election survey. The objective of the survey is to measure voting behavior. Target variables could be whether one voted (yes/no), party one voted for (choice from a list), reasons for voting for this party (open question), etc.

A target variable is denoted by the letter Y . The values of this variable for all elements in the population U are indicated by

$$(3.2) \quad Y_1, Y_2, \dots, Y_N.$$

So, Y_k is the value of variable Y for element k , where $k = 1, 2, \dots, N$. Usually, an additional number of variables are measured in the survey. These variables are called *auxiliary variables*. These variables assist in differentiating the survey results for various subpopulations. They can also be very useful in improving estimates of population characteristics. Examples of auxiliary variables are demographic characteristics like sex, age, and marital status. An auxiliary variable is denoted by the letter X . The values of variable X for all elements in the population U are indicated by

$$(3.3) \quad X_1, X_2, \dots, X_N.$$

So, X_k is the value of X for element k , where $k = 1, 2, \dots, N$. Here, Y_k and X_k indicate single values. Both target variables and auxiliary variables can be one of three types:

- *Continuous variables*. These variables measure quantities, amounts, sizes, or values. It is possible to carry out meaningful computations on these values, like calculating totals and averages. Examples of such variables are income and age of a person, the number of cars he owns, etc.
- *Categorical variables*. These variables divide the target population into subpopulations. The values denote labels of categories. Elements with the same label belong to the same category. It is not meaningful to carry out computations on the values of a categorical variable. Examples of categorical variables are race, religion, marital status, and region of residence.
- *Indicator variables*. Such a variable measures whether an element has a certain property. It can only assume two values 0 and 1. The value is 1 if an element has the property, and the value is 0 if it does not have it. An example of an indicator variable is employment. If a person has a job, the value of the variable is 1, and otherwise, its value is 0.

The aim of a survey is to get information about the target population. This information is quantified in the form of *population parameters*. A population parameter is a function that only depends on the values in the population for one or more variables. These variables can be target variables as well as auxiliary variables.

One simple example of an often used population parameter for a continuous variable is the *population total*

$$(3.4) \quad Y_T = \sum_{k=1}^N Y_k = Y_1 + Y_2 + \dots + Y_N.$$

Suppose the target population consists of all farms in a country, and Y denotes the number of cows a farm has, then the population total is the total number of cows in the country. Related to the population total is the *population mean*

$$(3.5) \quad \bar{Y} = \frac{1}{N} \sum_{k=1}^N Y_k = \frac{Y_1 + Y_2 + \cdots + Y_N}{N} = \frac{Y_T}{N}.$$

The population mean is simply obtained by dividing the population total by the population size. If the target population consists of all inhabitants of a town, and Y denotes the age of a person, the population mean is the mean age in this town. Another important population parameter is the *adjusted population variance*. It is defined by

$$(3.6) \quad S^2 = \frac{1}{N-1} \sum_{k=1}^N (Y_k - \bar{Y})^2.$$

This quantity gives an indication of the amount of variation in the values of the target variable. If all values of Y are equal, the variance is 0. The more the values of Y are apart, the larger the variance will be. The adjusted population variance also appears in formulas for the variance of estimators.

For indicator variables, the population total denotes the number of elements in the population having a certain property. The population mean is the fraction of elements with that property. The *population percentage* is defined by

$$(3.7) \quad P = 100\bar{Y} = \frac{100}{N} \sum_{k=1}^N Y_k = 100 \frac{Y_1 + Y_2 + \cdots + Y_N}{N} = 100 \frac{Y_T}{N}.$$

Note that for indicator variables, the adjusted population variance reduces to

$$(3.8) \quad S^2 = \frac{1}{N-1} \sum_{k=1}^N (Y_k - \bar{Y})^2 = \frac{P(100-P)}{N-1}.$$

There are no specific population parameters for categorical variables. Of course, totals, fractions, or percentages of elements in categories can be estimated. In fact, this comes down to replacing the categorical variable by a set of indicator variables, one for each category. The focus in this book is on estimating population means and population percentages.

3.2.2 SAMPLING FRAMES

To select a sample from a target population in a scientifically sound way, two ingredients are required: a sampling design based on some form of probability

sampling and a sampling frame. Several sampling designs are described in a subsequent section. This section will discuss sampling frames.

A *sampling frame* is a complete listing of all elements in the target population. For every element in the list, there must be information on how to contact that element. Such contact information can comprise, for example, a name and address, a telephone number, or an e-mail address. Such lists can exist on paper (a card-index box for the members of a club or a telephone directory) or in a computer (a database containing a register of all companies). If such lists are not available, detailed geographical maps are sometimes used.

Some countries, like the Netherlands and the Scandinavian countries, have a population register. The population register of the Netherlands is a decentralized system. Each municipality maintains its own register. Demographic changes related to their inhabitants are recorded. It contains information on gender, date of birth, marital status, and nationality. Periodically, all municipal information is combined into one large register, which is used by Statistics Netherlands as a sampling frame for its surveys. Samples of persons can be selected from it.

In some countries, address lists are available. These lists contain the addresses of all houses in each country. Such address lists may have been compiled as part of a census operation. If they are maintained well, they can be used to select a sample of households.

EXAMPLE 3.3 The postal delivery points file

In the Netherlands, the Postal Delivery Points file is frequently used as a sampling frame for address. This file is maintained by TNT Post, the postal service company. This Postal Delivery Points file is a computer file containing all addresses (both of private houses and of businesses) where post can be delivered. Typically, this file can be used to draw a sample of households or companies.

The basic unit of this file is an address. Note that this is not the same as a household. It is possible that several households may live at the same address.

The ideal sampling frame for a web survey would be a list of e-mail addresses of all members of the target population. Each element is represented by an e-mail address. A random sample of elements can be selected from this list. An e-mail is sent to all selected elements. This e-mail contains a link to the website containing the on-line survey questionnaire.

It is important to control access to the questionnaire with a unique identification code for each sample element. This guarantees that only selected elements can get access to the questionnaire. The unique code can also help to prevent that someone completes a questionnaire more than once. The unique identification code should be included in the e-mail that is sent to the selected persons. Such codes are often part of the link to the survey website.

Sometimes, a sampling frame consisting of e-mail addresses is indeed available. An example is a large company where all employees have their own company e-mail address. Another example is an educational institution, where all registered students have an e-mail address. For general-population surveys, such a sampling frame is usually not available. Then a different sampling frame must be used. If a population register or address register is available, a letter can be sent to a sample of persons or addresses. This letter invites the selected persons to participate in the survey. It also contains a link to the survey questionnaire website. This approach is somewhat more cumbersome. If the link is mentioned in an e-mail, it is sufficient to click on it to start the questionnaire. If the link is mentioned in a letter, it must be typed in, which involves a risk of making typing errors.

EXAMPLE 3.4 Sample selection for the LISS panel

As part of an attempt to boost the Dutch knowledge economy, the government of the Netherlands granted funds in 2006 to establish a web panel consisting of approximately 5,000 households. This panel is called the *LISS panel*. LISS stands for Longitudinal Internet Studies for the Social Sciences. Universities are invited to submit research proposals that can be carried out free of charge in this panel. See the study by Scherpenzeel (2008) for a detailed description of this panel.

The panel has been constructed by selecting a random sample of households from the population register of The Netherlands. Selected households were recruited for this panel by means of a face-to-face interview (computer-assisted personal interviewing) or a telephone interview (computer-assisted telephone interviewing).

An initial sample of 10,150 addresses was selected from the population register. For each sample address, an attempt was made to find a corresponding landline telephone number. This was successful in 70% of the cases. These addresses were approached by telephone (CATI) for a recruitment interview. The other 30% of the cases were approached face to face (CAPI) for recruitment.

Unusable addresses were removed. These were, among others, nonexisting addresses, uninhabited addresses, business addresses, and addresses with people unable to participate (because of long-time illness or language problems). The remaining sample size was 9,944 addresses.

The recruitment interview was a 10-minute interview, in which some basic demographic questions were asked, and questions about an Internet connection at home, social integration, political interest, leisure activities, survey attitudes, loneliness, and personality. At the end of the interview, the respondent was invited to become a member of the web panel.

If the contacted household refused to participate in the short recruitment interview, it was asked to answer just three basis questions.

These three questions were followed by the invitation to participate in the panel.

TABLE 3.1 The recruitment process for the LISS panel

Phase of the recruitment process	Remaining % of addresses
Start with 9,844 addresses	100
Completed CAPI or CATI recruitment, or basic questions	75
Willing to participate in panel	64
Actually participated	48

Table 3.1 shows the results of the recruitment process. The response for the recruitment interview was reasonably high (75%). Of those respondents, 84% (corresponding to 64% of the addresses) expressed willingness to participate in the panel. Of those willing to participate, ultimately 75% did (corresponding to 48% of the starting cases).

The sampling frame should be an accurate representation of the population. There is a risk of drawing wrong conclusion from the survey if the sample has been selected from a sampling frame that differs from this population. Figure 3.1 shows what can go wrong.

The first problem is *undercoverage*. This occurs if the target population contains elements that do not have a counterpart in the sampling frame. Such elements can never be selected in the sample. An example of undercoverage is a survey where the sample is selected from a population register. Illegal immigrants are part of the population, but they will never be encountered in the sampling frame. Another example is a web survey, where respondents are selected via the Internet. Then there will be undercoverage because of people without Internet access. Undercoverage can have serious consequences. If the elements outside the

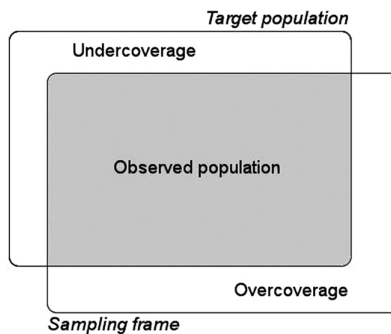


FIGURE 3.1 Target population and sampling frame

sampling frame systematically differ from the elements in the sampling frame, estimates of population parameters may be seriously biased. A complicating factor is that it is often not very easy to detect the existence of undercoverage.

The second sampling frame problem is *overcoverage*. This refers to the situation where the sampling frame contains elements that do not belong to the target population. If such elements end up in the sample and their data are used in the analysis, estimates of population parameters may be affected. It should be simple to detect overcoverage in the field. This should become clear from the answers to the questions asked.

Another example is given to describe coverage problems. Suppose a survey is carried out among the inhabitants of a town. It is decided to collect data by means of a web survey. As there is no sampling frame containing e-mail addresses of the inhabitants, it is decided to recruit people for the survey by telephone. A sample of inhabitants is selected from the telephone directory of the town. Unfortunately, this sampling frame can have serious coverage problems. Undercoverage occurs because many people have unlisted numbers, and some will have no phone at all. Moreover, there is a rapid increase in the use of mobile phones. In many countries, mobile phone numbers are not listed in directories. In the Netherlands, only two out of three people can be found in the telephone directory. A telephone directory also suffers from overcoverage because it contains the telephone numbers of shops, companies, etc. Hence, it may happen that people are contacted that do not belong to the target population. Moreover, some people may have a higher contact probability than anticipated because they can be contacted both at home and in the office.

A survey is often supposed to measure the status of a population at a specific moment in time. This is the so-called *reference date*. The sampling frame should reflect the status at this reference date. As the sample will be selected from the sampling frame before the reference date, there may be discrepancies. The sampling frame may contain elements that do not exist any more at the reference date. People may have died or companies may have ceased to exist. These are cases of overcoverage. It may also happen that new elements have come into existence after the time of sample selection and before the reference date. For example, a person moves into town or a new company is created. These are cases of undercoverage.

EXAMPLE 3.5 Overcoverage or nonresponse?

Suppose a survey is carried in a town among the people of age 18 years and older. The objective is to describe the situation at the reference date of May 1. The sample is selected in the design phase of the survey, say, at April 1. It is a large survey, so data collection cannot be completed in one day. Therefore, interviews are conducted in a period of two weeks, starting one week before the reference date and ending one week after the reference date.

Now suppose an interviewer contacts a selected person at April 29. It turns out the person has moved to another town. This is a case of overcoverage. What counts is the situation at May 1, and the person did not belong any more to the target population at the reference date. So, there is no problem. As this is a case of overcoverage, it can be ignored.

The situation is different if an interviewer attempts to contact a person at May 5, and this person turns out to have moved at May 2. This person belonged to the target population at the reference date and, therefore, should have been interviewed. This is no coverage problem but a case of nonresponse. The person should be tracked down and interviewed

Problems may also occur if the units in the sampling frame are different from those in the target population. Typical is the case the target population consists of persons and the sampling frame of addresses. This may happen if an address list (for example, a telephone directory) is used as a sampling frame. Suppose persons have to be selected with equal probabilities. A naive way to do this would be to select a sample of addresses, randomly and to draw one person at each selected address. At first sight, this is reasonable, but it ignores the fact that now not every person has the same selection probability: Members in large families have a smaller probability of being selected than members of small families.

3.2.3 BASIC CONCEPTS OF SAMPLING

To be able to obtain reliable estimates of population parameters, a random sample is selected from the population. The elements in this sample are obtained by means of a random selection procedure. This procedure assigns to every element in the target population a fixed, positive, and known probability of selection. The most straightforward way to select a random sample is giving each element the same probability of selection. Such a random sample is called a *simple random sample*.

Samples can be selected with replacement or without replacement. *Sampling with replacement* means that a selected element is returned to the population (after its characteristics have been recorded) before the next element is drawn. It is possible to select an element more than once. *Sampling without replacement* means that a selected element is not returned to the population. Therefore, a elements can only be selected at most once in a sample. Selecting an element more than once does not produce more information than selecting it once. Hence, selection without replacement is usually preferred.

It is assumed here that the sample is selected without replacement. It means that each element can appear at most once in the sample. Therefore, the sample can be represented by a set of indicators

$$(3.9) \quad a = a_1, a_2, \dots, a_N.$$

The indicator a_k assumes the value 1 if element k is selected in the sample, and otherwise it assumes the value 0, for $k = 1, 2, \dots, N$. The expected value of a_k is denoted by

$$(3.10) \quad \pi_k = E(a_k).$$

The quantity π_k is called the *first-order inclusion probability* of element k (for $k = 1, 2, \dots, N$). For deriving variance formulas, also second-order inclusion probabilities are required. The *second-order inclusion probability* of elements k and l (with $k \neq l$) is equal to

$$(3.11) \quad \pi_{kl} = E(a_k a_l),$$

and by definition $\pi_{kk} = \pi_k$. The *sample size*, i.e., the number of selected elements, is denoted by n . As the indicators a_k have the value 1 for all elements in the sample, and the value 0 for all other elements, the sample size is equal to the sum of the values of the indicators:

$$(3.12) \quad n = \sum_{k=1}^N a_k.$$

The values of the target variable are observed for the sampled elements. These values are available to estimate population characteristics. The recipe to compute such an estimate is called an *estimator*, and the result of this computation is called the *estimate*. In order to be useful, an estimator must have a number of properties.

- The estimator must be *unbiased*. This means that the average value of the estimates over all possible samples must be equal to the value of the (unknown) population parameter to be estimated. On average, the estimator will result in the correct value. It will never underestimate or overestimate the population value in a systematic way. Consequently, the expected value $E(z)$ of an estimator z must be equal to the value of the population parameter to be estimated: $E(z) = Z$.
- The estimator must be *precise*. It means that the variation in possible outcomes must be small. Consequently, the variance $V(z)$ of an estimator z of a population parameter Z must be small over all possible samples.
- *For reasons of simplicity*, linear estimators are preferred. It means an estimate is computed as a linear combination of the observed values of the target variable.

Imposing the conditions of unbiasedness and linearity leads to the estimator introduced by Horvitz and Thompson (1952). This estimator for the population mean is defined as

$$(3.13) \quad \bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^N a_k \frac{Y_k}{\pi_k}.$$

The indicators a_k see to it that only the available sample values of the target variable are used in the computation of the estimate. Note that each value Y_k is weighted with the inverse selection probability π_k . Thus, the estimator corrects for the fact that elements with a large inclusion probability are overrepresented in the sample.

The Horvitz–Thompson estimator (3.13) is an unbiased estimator of the population mean. The variance of this estimator is equal to

$$(3.14) \quad V(\bar{y}_{HT}) = \frac{1}{N^2} \sum_{k=1}^N \sum_{l=1}^N (\pi_{kl} - \pi_k \pi_l) \frac{Y_k}{\pi_k} \frac{Y_l}{\pi_l}.$$

For without replacement samples of fixed size n , the variance can be rewritten in the form

$$(3.15) \quad V(\bar{y}_{HT}) = \frac{1}{2N^2} \sum_{k=1}^N \sum_{l=1}^N (\pi_k \pi_l - \pi_{kl}) \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l} \right)^2.$$

This expression shows that the variance can be reduced by taking the first-order inclusion probabilities proportional to the values of the target variable.

To quantify the precision of an estimate, a confidence interval can be computed. The *confidence interval* is a range of possible values of the population parameter. The interval encompasses the true value of the population mean with a high probability if an estimator is unbiased. This probability is called the *confidence level*. It is denoted by $(1 - \alpha)$, where α is a small probability. Often the value $\alpha = 0.05$ is used. The confidence level is then 95%.

The 95% confidence interval of the Horvitz–Thompson estimator is equal to

$$(3.16) \quad (\bar{y}_{HT} - 1.96 \times S(\bar{y}_{HT}); \bar{y}_{HT} + 1.96 \times S(\bar{y}_{HT})),$$

where

$$(3.17) \quad S(\bar{y}_{HT}) = \sqrt{V(\bar{y}_{HT})}$$

is the *standard error* of the estimator. For a 99% confidence interval, the value 1.96 must be replaced by 2.58.

A problem is that usually the value of the variance or standard error is not known. It can only be computed if all values of the target variable in the population are available. The way out is that an unbiased estimate of the standard error can be computed just using the sample data. This leads to an estimated confidence interval

$$(3.18) \quad ((\bar{y}_{HT} - 1.96 \times s(\bar{y}_{HT}); \bar{y}_{HT} + 1.96 \times s(\bar{y}_{HT})),$$

where $s(\bar{y}_{HT})$ is the estimate for $S(\bar{y}_{HT})$.

3.2.4 SIMPLE RANDOM SAMPLING

The best known and probably most often used sampling design is a *simple random sample* without replacement. This is a sample design in which all elements have the same probability of being selected. First-order inclusion probabilities of all elements are equal. It can be shown for without replacement sampling that all first-order inclusion probabilities always sum up to n . Therefore, $\pi_k = n / N$, for $k = 1, 2, \dots, N$. Furthermore, all second-order inclusion probabilities sum up to $n(n - 1)$. Therefore, $\pi_{kl} = n(n - 1)/N(N - 1)$, for $k, l = 1, 2, \dots, N$ and $k \neq l$.

Suppose the objective of the survey is to estimate the population mean of a continuous target variable Y . Substitution of the values of the first-order inclusion probabilities in expression (3.13) results in a simple estimator, the *sample mean*

$$(3.19) \quad \bar{y} = \frac{1}{n} \sum_{k=1}^N a_k Y_k = \frac{1}{n} \sum_{i=1}^n y_i,$$

where y_1, y_2, \dots, y_n denote the n observations that have become available in the sample. This is an unbiased estimator with variance

$$(3.20) \quad V(\bar{y}) = \frac{1-f}{n} S^2,$$

where $f = n/N$ is the *sampling fraction* and S^2 is the population variance. Expression (3.20) shows that an increased sample size leads to more precise estimators. The standard error of the sample mean is equal to

$$(3.21) \quad S(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{\frac{1-f}{n}} S.$$

To compute an estimated 95% confidence interval, an unbiased estimator for S^2 is required. The sample variance

$$(3.22) \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

can be used for this.

EXAMPLE 3.6 Effect of sample size on the precision of an estimator

The effect of sample size on the precision of an estimator can be shown by means of a simulation experiment. The fictitious population of the country of Samplonia has been constructed. The population consists of 1,000 people. The working population of Samplonia consists of 341 people.

Overall, 1,000 simple random samples without replacement of size 20 have been selected from the working population. For each sample, the mean income is computed as an estimate of the mean income in the population. The distribution of these 1,000 estimates is displayed in the graph on the left in Figure 3.2.

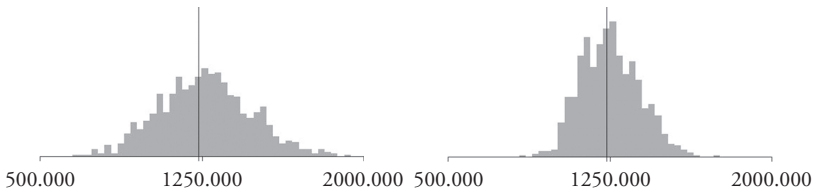


FIGURE 3.2 Random samples from the working population of Samplonia

There is a lot of variation of the estimates around the population mean indicated by the vertical line. This variation can be reduced by increasing the sample size. The graph on the right shows the distribution of 1,000 estimates based on the sample of the size 40. Indeed, doubling the sample size reduces the variance.

The objective of a survey also often is to estimate a population percentage. Typical examples are the percentage of people voting for a political party, the percentage of households having an Internet connection, and the unemployment percentage. The theory for estimating percentages does not essentially differ from the theory of estimating means. In fact, percentages are just population means multiplied by 100 where the target variable Y is an indicator variable, [i.e., it only assumes the value 1 (if the element has the specific property) or 0 (if the element does not have the property)]. Because of this restriction on the possible values, formulas become even much simpler.

If Y only assumes the values 1 and 0, the population mean \bar{Y} is equal to the proportion of elements having a specific property. The population percentage P is therefore equal to

$$(3.23) \quad P = 100\bar{Y}.$$

Estimation of a population percentage comes down to first estimating the population mean. The sample mean is an unbiased estimator for this quantity. Multiplication of the sample mean by 100 produces the sample percentage. This estimator is denoted by

$$(3.24) \quad p = 100\bar{y}.$$

As the sample mean is an unbiased estimator for the population mean, the sample percentage is an unbiased estimator of population percentage.

The variance of this estimator can be found by working out the term S^2 in variance formula (3.6) for a population in which a percentage P of the elements has a specific characteristic and a percentage $100 - P$ does not have this characteristic. This results in the simple formula

$$(3.25) \quad V(p) = \frac{1-f}{n} \frac{N}{N-1} p(100-p).$$

This variance can be estimated using the sample data. If p denotes the sample percentage, then

$$(3.26) \quad v(p) = \frac{1-f}{n-1} p(100-p)$$

is an unbiased estimator of the variance (3.25). The estimated variance is used to obtain a (estimated) confidence interval.

EXAMPLE 3.7 Estimating a percentage

An election poll was conducted in June 2010 in the Netherlands. It was a web survey. The sample size was 1,000 persons. The population consisted of 12 million voters. The results showed that in the sample 22.1% would vote for the Liberal Party and 20.5% for the Social Democrats.

The estimated variance of the estimate for the Liberal Party is

$$v(p) = \frac{1-f}{n-1} p(100-p) = \frac{1-1000/12000000}{999} 22.1 \times 77.9 = 1.723170.$$

The standard error of the estimate is therefore

$$s(p) = \sqrt{v(p)} = \sqrt{1.723170} = 1.312670.$$

The 95% confidence interval becomes

$$\begin{aligned} & (p - 1.96 \times S(p); p + 1.96 \times S(p)) \\ &= (22.1 - 1.96 \times 1.312670; 22.1 + 1.96 \times 1.312670) \\ &= (19.5; 24.7). \end{aligned}$$

The conclusion can be that with 95% confidence, the percentage of voters in the population will be between 19.5% and 24.7%.

The 95% confidence interval for the percentage of voters for the Social Democrats turns out to run between 18.0% and 23.0%. The intervals for both parties have a considerable overlap. So it cannot be concluded that in the population, the Liberals will get more votes than the Social Democrats.

3.2.5 DETERMINING THE SAMPLE SIZE

A decision to be made in the survey design phase is the size of the sample to be selected. This is an important decision. If, on the one hand, the sample is larger than really necessary, a lot of time and money may be wasted. And if, on the other hand, the sample is too small, the required precision will not be achieved, making the survey results less useful.

A relationship between the precision of an estimator and the sample size exists: The larger the sample is, the more precise the estimator will be. Therefore, the question about the sample size can only be answered if it is clear how precise the estimator must be. Once the precision has been specified, the sample size can be computed. A very high precision requires a large sample. If the costs per interview are large, this will make the survey expensive. Once a web survey questionnaire has been prepared on the Internet, and a sample of e-mail addresses is available, the costs per interview can be very low. However, it should be realized that recruitment can be expensive if it is done by means of CAPI or CATI. In practice, the sample size will always be a compromise between costs and precision.

Some formulas will be given here for the size of a simple random without replacement. The first situation to be considered is that for estimating population percentages. Then the case of estimating population means will be described.

3.2.5.1 The Sample Size for Estimating a Percentage. The starting point is that the researcher gives some indication of how large the *margin of error* at most may be. The margin is defined as the distance between the estimate and the lower or upper bound of the confidence interval. Formulas are given for the sample size that is at least required to achieve this margin of error. In the case of a 95% confidence interval, the margin of error is equal to

$$(3.27) \quad 1.96 \times S(p).$$

For a 99% confidence interval, the value of 1.96 must be replaced by 2.58. Suppose the margin of error may not exceed a value M . Rewriting this condition leads to

$$(3.28) \quad S(p) \leq \frac{M}{1.96}.$$

The variance of the estimator for a population percentage can be found in expression (3.25). Substituting this expression in inequality (3.28) leads to the condition

$$(3.29) \quad \sqrt{\frac{1-f}{n} \frac{N}{N-1} P(100-P)} \leq \frac{M}{1.96}.$$

The lower bound for the sample size can now be computed by solving n from this equality. However, there is a problem because expression contains an unknown quantity, and that is population percentage P . There are two ways to solve this problem:

- There is some indication of the value of P . Maybe there was a previous survey in which this quantity was estimated. Or maybe a subject matter expert may provide an educated guess. Such an indication can be substituted in expression (3.29), after which it can be solved.
- Nothing at all is known about the value of P . Now $P(100-P)$ is a quadratic function that assumes its minimum value 0 in the interval $[0, 100]$ for $P=0$ and $P=100$. Exactly in the middle, for $P=50$, the function assumes its maximum value. This implies that the upper bound for the variance can be computed by filling in the value $P=50$. So the worst case for the variance is obtained for this value of P . For any other value of P , the variance is smaller. If the value is determined so that the worst-case variance is not exceeded, the true variance will certainly be smaller. It should be noted that for values of P between, say, 30% and 70%, the true variance will not differ much from the maximum variance.

Solving n from inequality (3.29) leads to a lower bound of n equal to

$$(3.30) \quad n \geq \frac{1}{\frac{N-1}{N} \left(\frac{M}{1.96}\right)^2 \frac{1}{P(100-P)} + \frac{1}{N}}.$$

A simple approximation can be obtained if the population size N is very large. Then $(N-1)/N$ can be approximated by 1 and the value of $1/N$ can be ignored. This implies that expression (3.30) is reduced to

$$(3.31) \quad n \geq \left(\frac{1.96}{M}\right)^2 P(100-P).$$

■ EXAMPLE 3.8 The sample size for an opinion poll

Suppose that in an earlier opinion poll, 21% of the respondents indicated to vote for a specific party. A new poll will be conducted to measure the current support for this party. No dramatic changes are expected.

Therefore, it is not unreasonable to fill in a value of 21 for P in expression (3.31). Furthermore, the margin of error should not exceed $M = 3\%$. Substitution in expression (3.31) results in

$$n \geq \left(\frac{1.96}{3}\right)^2 21 \times 79 = 708.1.$$

So, the sample size must be at least equal to 709. The confidence level is 95%. For a confidence level of 99%, the value of 1.96 must be replaced by 2.58, leading to a minimum sample size of 1,227.

3.2.5.2 The Sample Size for Estimating a Mean. Expression (3.28) is also the starting point for the computation of the sample size if the objective of the survey is to estimate the mean of a continuous target variable. However, no simple expression for the standard error is available. Expression (3.28) can be rewritten as

$$(3.32) \quad \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)S^2} \leq \frac{M}{1.96},$$

in which S^2 is the adjusted population variance. The problem is that usually this variance is unknown. Sometimes a rough estimate is available from a previous survey. Or maybe some indication can be obtained from a test survey. In these situations, the approximate value can be substituted in expression (3.32). Rewriting the inequality leads to

$$(3.33) \quad n \geq \frac{1}{\left(\frac{M}{1.96S}\right)^2 + \frac{1}{N}}.$$

The quantity $1/N$ can be ignored for large values of N . This produces the somewhat simpler expression

$$(3.34) \quad n \geq \left(\frac{1.96S}{M}\right)^2.$$

3.2.6 SOME OTHER SAMPLING DESIGNS

An overview of four other sampling designs is given in this section. These sampling designs are stratified sampling, sampling with unequal probabilities, cluster sampling, and two-stage sampling. More on sampling designs can be found in, for example, in Cochran (1977) and Bethlehem (2009).

3.2.6.1 Stratified Sampling. To select a stratified sample, the population is divided into a number of subpopulations. These subpopulations are called *strata*. A sample is selected in each stratum. So, for each stratum, an unbiased estimate of the stratum mean or percentage can be computed. Then, these stratum estimators can be combined into an unbiased estimator of the population mean or percentages. There are various reasons to apply stratified sampling:

- If the strata are homogeneous, i.e., all elements within strata resemble each other, the variance of estimators will be small. So, estimators based on stratified sampling will be more precise than estimators based on simple random sampling.
- There may be situations in which not only estimates are required for the population as a whole, but also for specific subpopulations. By using these subpopulations as strata in stratified sampling, the researcher can see to it that a sufficient number of observations becomes available in each subpopulation.
- By applying stratified sampling with the same fraction of observations in each stratum, the sample becomes at least representative with respect to these strata.

Stratified sampling can only be implemented if a proper sampling frame is available. There must be a separate sampling frame for each subpopulation. This condition sometimes prevents application of stratified sampling. For example, it is usually not possible to stratify a sample for a general population survey by level education. The reason is there is no separate sampling frame for each level of education. However, it is possible in countries like the Netherlands to stratify a sample by region, because each municipality has its own population register.

Another example could be stratifying individuals by Internet access or not. This information is not available in a sampling frame. Therefore, it is not possible to stratify with reference to this characteristic.

To apply stratified sampling, the target population U is divided into L subpopulations (strata) U_1, U_2, \dots, U_L of sizes N_1, N_2, \dots, N_L , respectively. The strata are nonoverlapping and together cover the whole population. This implies that

$$(3.35) \quad \sum_{h=1}^L N_h = N_1 + N_2 + \dots + N_L = N.$$

The N_h values of the target variable Y in stratum h are denoted by

$$(3.36) \quad Y_1^{(h)}, Y_2^{(h)}, \dots, Y_{N_h}^{(h)}.$$

The mean in stratum h can be written as

$$(3.37) \quad \bar{Y}^{(h)} = \frac{1}{N_h} \sum_{k=1}^{N_h} Y_k^{(h)},$$

and the population mean can be written as

$$(3.38) \quad \bar{Y} = \frac{1}{N} \sum_{h=1}^L N_h \bar{Y}^{(h)}.$$

So the population mean is a weighted average of the stratum means. The (adjusted) variance in stratum h is equal to

$$(3.39) \quad S_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (Y_k^{(h)} - \bar{Y}^{(h)})^2.$$

A stratified sample of size n is selected from this population by selecting L subsamples of sizes n_1, n_2, \dots, n_L , respectively, where n_b is the sample size in stratum b , for $b = 1, 2, \dots, L$. In principle, any sampling design can be applied within the strata, but usually simple random samples without replacement are used. If the n_b observations in stratum b are denoted by

$$(3.40) \quad y_1^{(b)}, y_2^{(b)}, \dots, y_{n_b}^{(b)},$$

the sample mean

$$(3.41) \quad \bar{y}^{(b)} = \frac{1}{n_b} \sum_{i=1}^{n_b} y_i^{(b)}$$

in stratum b is an unbiased estimator of the population mean in stratum b . Now, the stratum estimators can be combined into an estimator for the population mean of Y . Using expression (3.38), it can be shown that

$$(3.42) \quad \bar{y}_S = \frac{1}{N} \sum_{b=1}^L N_b \bar{y}^{(b)}$$

is an unbiased estimator of the population mean. The variance of the sample mean in stratum b is equal to

$$(3.43) \quad V(\bar{y}^{(b)}) = \frac{1 - f_b}{n_b} S_b^2,$$

in which $f_b = n_b / N_b$. As the subsamples are selected independently, it can be shown that the variance of estimator (3.42) is equal to

$$(3.44) \quad V(\bar{y}_S) = \frac{1}{N^2} \sum_{b=1}^L N_b^2 \frac{1 - f_b}{n_b} S_b^2.$$

This variance is small if the stratum variances S_b^2 are small. This is the case if there is little variation in the values of the target variable within strata, i.e., if the strata are homogeneous with respect to the target variable.

The variance of the estimator is influenced by the sample sizes n_1, n_2, \dots, n_L in the strata, i.e., the sample *allocation*. The variance is minimal for the so-called *optimal allocation* (also called *Neyman allocation*), [see, e.g., Cochran (1977)]. This is the allocation where the n_b are proportional to $N_b \times S_b$. Applying this allocation requires stratum variances to be known. If this is not the case, and there are also no estimates available, another option is to use *proportional allocation*. This is a sample allocation where the n_b are proportional to N_b . As a result, every element in the target population has the same selection probability.

EXAMPLE 3.9 Business surveys in the Netherlands

Statistics Netherlands maintains a general business register. This is a database containing general information about all companies in the Netherlands. Variables included in the database are name and address, type economic activity (according to the so-called international NACE classification), and size class (in terms of number of employees).

For its business surveys, Statistics Netherlands draws stratified samples from this business register. Samples are stratified by type of economic activity and size class. This implies that a stratum contains companies with the same activity and of the same size. Therefore, the strata are homogeneous. This results in precise stratum estimates.

The sample allocation is often such that large companies have a larger selection probability than small companies. As the values of many target variables (for example, turnover, profit, or investments) are related to the size of the companies, this also improves the precision of estimates.

Many of these surveys employ data collection using the Internet. There are two approaches. The first approach is a web survey. Companies can complete questionnaires on the Internet directly. The other is that they receive the interview software by e-mail or CD. This software is downloaded and installed by the companies. They run the software, answer the questions, and return the answers by e-mail.

3.2.6.2 Sampling with Unequal Probabilities. It is an interesting property of the Horvitz–Thompson estimator that its variance is small if the first-order inclusion probabilities are more or less proportional to the values of the target variable (i.e., Y_k / π_k is approximately constant for all k). This is difficult to realize in practice because it requires all values of the target variable in the population to be known. If this is the case, there is no reason for carrying out a survey. However, sometimes the values in the population of an auxiliary variable X are known. So, first-order inclusion probabilities could be taken proportional to the values of this variable. If there is a strong correlation between the target variable Y

and the auxiliary variable X , then the result will be a precise estimator. An example is a shoplifting survey where shops are sampled according to their floor size, assuming there is more shoplifting in larger shops than in smaller shops.

■ **EXAMPLE 3.10** Sampling addresses for a web survey: A case of unequal probabilities

Suppose a web survey is conducted where the sampling frame is a list of addresses. A simple random sample of n addresses is selected. For each selected address, one randomly selected person is asked to complete the questionnaire. This could, for example, be implemented by selecting the person with the next birthday.

The resulting sampling design is one in which not every person has the same probability of selection. Let N denote the total number of persons in the population. This population is divided over M addresses. There live N_b persons at address b , where $N_1 + N_2 + \dots + N_M = N$. The inclusion probability of a person k at address b is now equal to

$$(3.45) \quad \pi_k^{(b)} = \frac{n}{M} \times \frac{1}{N_b}.$$

This expression is obtained by multiplying the inclusion probability of an address (n/M) by the probability of selecting a person at this address ($1/N_b$). Substitution of this expression in the Horvitz–Thompson estimator results in the estimator

$$(3.46) \quad \bar{y}_{HT} = \frac{1}{N} \sum_{b=1}^M a_b \sum_{k=1}^{N_b} b_{bk} Y_{bk} \frac{MN_b}{n},$$

where the indicator a_b indicates whether address b is selected and the indicator b_{bk} indicates whether element k at address b is selected. This estimator is not equal to the simple sample mean. It is a weighted mean with the value Y_{bk} for each respondent with the number of people N_b at the address.

3.2.6.3 Cluster Sampling. Another sampling design is *cluster sampling*. This type of sampling can be applied if there is no sampling frame for the elements in the population, but there is one for clusters of elements. In this situation, a sample of clusters can be selected, and all elements in each selected cluster can be observed. A typical example of a cluster sample is an address sample where all people at a selected address are invited to participate in the survey.

Cluster sampling does not necessarily produce precise estimators. To the contrary, the more elements within clusters resemble each other, the less efficient

the estimator will be. Another disadvantage of cluster sampling is that there is no control over the sample size. It simply depends on the numbers of elements in the selected clusters.

3.2.6.4 Two-Stage Sampling. One way to get more control is to select a *two-stage sample*. First a sample of clusters is selected, and then a sample of elements is drawn from each selected cluster. Also here, the reasons for applying this sampling design are more practical. Again, this procedure will not produce very accurate estimates, but it may be necessary to do this because of the lack of a proper sampling frame. The reason to apply a two-stage sampling design can also be reduction of costs. This only applies if interviewers are used for data collection. Interviewers have to travel less for a face-to-face survey if the addresses of selected persons are concentrated in clusters.

EXAMPLE 3.11 The Safety Monitor of Statistics Netherlands

National statistical institutes are always under pressure to reduce data collection costs. This has led to considering new ways of data collection, in which web surveys play an important role. The most far reaching change is to replace an expensive CAPI and CATI survey by a web survey. Another option is to introduce a mixed-mode survey. Statistics Netherlands has experimented with mixed-mode data collection for the Safety Monitor.

This survey measures how the Dutch feel with respect to security. Questions were asked, among others, about feelings of security, quality of life, and level of crime experienced.

The target population for this survey consisted of all persons of age 15 years and older. The sample for this survey was selected from the Dutch population register. It was a stratified sample. Strata were constructed by crossing interview regions of Statistics Netherlands by the 25 police regions in which the country is divided. Within each stratum, a two-stage sample was selected. In the first stage, municipalities were selected. In the second stage, persons were drawn in the selected municipalities.

All sample persons received a letter in which they were asked to complete the survey questionnaire on the Internet. The letter also included a postcard that could be used to request a paper questionnaire.

Two reminders were sent to those that did not respond by web or mail. If still no response was obtained, nonrespondents were approached by means of CATI, if a listed telephone number was available. If not, these nonrespondents were approached by CAPI. This four-mode survey is denoted by SM4.

To be able to compare this four-mode survey with a traditional survey, also a two-mode survey was conducted for an independent sample. Sampled persons were approached by CATI if there was a listed telephone

number, and otherwise they were approached by CAPI. The two-mode survey is denoted by SM2.

The response rate for SM4 turned out to be 59.7%. The response rate for SM2 was 63.5%. So there was not much difference. More than half of the response (58%) was obtained in the SM4 with a self-administered mode of data collection (CAWI or PAPI).

The conclusion was drawn that the four-mode survey did not increase the response. The costs of the survey were, however, much lower, because interviewers were deployed in only 42% of the cases. Focusing on just interviewer costs, and ignoring all other costs (which are much lower), Beukenhorst and Wetzels (2009) found that the costs of SM4 were only 60% of the costs if SM2.

3.2.7 ESTIMATION PROCEDURES

The precision of an estimator may be improved by using auxiliary information. A good example is sampling with inclusion probabilities proportional to the values of an auxiliary variable. If there is a strong correlation between target variable and auxiliary variable, the variance of the Horvitz–Thompson estimator will be small. Another example is stratified sampling. This comes down to using categorical auxiliary variables that divide the population in homogeneous groups.

In the examples above, auxiliary information is used in the sampling design. Auxiliary information can also be used in a different way, i.e., in the estimator itself. Some examples of improved estimation procedures are described here. It is assumed that simple random sampling without replacement is applied.

3.2.7.1 The Ratio Estimator. The *ratio estimator* assumes that a continuous auxiliary variable X is available, the values of which are more or less proportional to the values of the target variable, i.e.,

$$(3.47) \quad Y_k \approx BX_k,$$

for some constant B . The ratio estimator is defined by

$$(3.48) \quad \bar{y}_{RAT} = \frac{\bar{y}}{\bar{x}} \bar{X},$$

where \bar{x} and \bar{y} are the sample means of X and Y and \bar{X} is the population mean of X . The estimator is asymptotically unbiased, and its variance is approximately equal to

$$(3.49) \quad V(\bar{y}_{RAT}) \approx \frac{1-f}{n} \frac{1}{N-1} \sum_{k=1}^N \left(Y_k - \frac{\bar{Y}}{\bar{X}} X_k \right)^2.$$

It can be shown that this variance is smaller the better condition (3.47) is satisfied. See, for example, Cochran (1977).

3.2.7.2 The Regression Estimator. An even better estimator is the *regression estimator*. It assumes a linear relationship

$$(3.50) \quad Y_k \approx A + BX_k$$

between the values of the target variable and the auxiliary variable. A and B are constants that have to be estimated using the sample data. This can be done with ordinary least squares. The regression estimator is defined by

$$(3.51) \quad \bar{y}_{REG} = \bar{y} - b(\bar{x} - \bar{X}),$$

where

$$(3.52) \quad b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

and x_1, x_2, \dots, x_n denote the n observations for X that have become available in the sample. The estimator is asymptotically unbiased, and its variance is approximately equal to

$$(3.53) \quad V(\bar{y}_{REG}) \approx \frac{1-f}{n} S^2 (1 - R_{XY}^2),$$

where R_{XY} is the correlation between X and Y in the population. It is clear from expression (3.53) that the variance of the regression estimator is never larger than that of the simple sample mean. The stronger the correlation, the smaller the variance will be.

EXAMPLE 3.12 A Dairy farm survey

The target population of a survey consists of 200 dairy farms in the rural part of the fictitious country of Samplonia. The objective of the web survey is to estimate the average daily milk production per farm. As all farms are connected to the Internet, a probability sample can be selected. A simple random sample is drawn.

Two estimators are compared: the simple sample mean and the regression estimator. The regression estimator uses the number of cows

per farm as the auxiliary variable. This seems not unreasonable as one may expect milk production per farm to be more or less proportional to the number of cows per farm.

The selection of a sample of size 40 and the computation of the estimator has been repeated 500 times for both estimators. This gives 500 values of each estimator. Figure 3.3 contains the distribution of the values of both estimators. The histogram on the left shows the distribution of the sample mean. The distribution of the regression estimator is shown on the right.

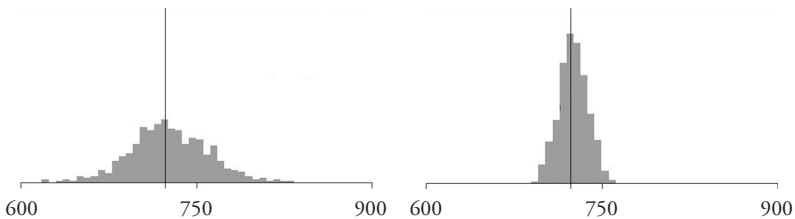


FIGURE 3.3 Comparing the sample mean with the regression estimator

The regression estimator performs better than the direct estimator. The distribution of its values concentrates much more around the true value. The standard error of the sample mean is equal to 35.6, whereas it is 12.4 for the regression estimator. So, a more precise estimator can be obtained with the same sample size if the proper auxiliary information is available.

3.2.7.3 The Poststratification Estimator. The ratio estimator and the regression estimator both use a continuous auxiliary variable. It is also possible to use a categorical auxiliary variable. A well-known example of this is the *poststratification estimator*. Suppose this auxiliary variable divides the target population U into L subpopulations (strata) U_1, U_2, \dots, U_L of sizes N_1, N_2, \dots, N_L , respectively. After a simple random sample has been selected, the sample mean can be computed in each stratum, after which these stratum estimates can be combined into an estimate

$$(3.54) \quad \bar{y}_{PS} = \frac{1}{N} \sum_{b=1}^L N_b \bar{y}^{(b)}$$

for the population mean. Note that this expression is identical to expression (3.42) for stratified sampling. However, estimator (3.54) has different statistical

properties because the underlying selection mechanism is different. It can be shown that the poststratification estimator is approximately unbiased and that its variance is equal to

$$(3.55) \quad V(\bar{y}_{PS}) = \frac{1-f}{n} \sum_{b=1}^L W_b S_b^2 + \frac{1}{n^2} \sum_{b=1}^L (1 - W_b) S_b^2,$$

in which $W_b = N_b / N$, and S_b^2 is the population variance of the target variable in stratum b . If the strata are homogeneous with respect to the values of the target variable (i.e., there is little variation within strata), this variance will be small. Note that when n becomes larger, expression (3.55) will be closer to expression (3.44) as the value of the second term in (3.55) quickly becomes smaller.

3.2.7.4 The Generalized Regression Estimator. A more general estimator can be defined of which the regression estimator and the poststratification estimator are special cases. This *generalized regression estimator* is introduced here because it is used in nonresponse correction techniques.

Suppose there are p auxiliary variables available. The p -vector of values of these variables for element k is denoted by

$$(3.56) \quad X_k = (X_{k1}, X_{k2}, \dots, X_{kp})'$$

The symbol $'$ denotes transposition of a matrix or vector. Let Y be the N -vector of all values of the target variable, and let X be the $N \times p$ -matrix of all values of the auxiliary variables. The vector of population means of the p auxiliary variables is defined by

$$(3.57) \quad \bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)'$$

If the auxiliary variables are correlated with the target variable, then for a suitably chosen vector $B = (B_1, B_2, \dots, B_p)'$ of regression coefficients for a best fit of Y on X , the residuals $E = (E_1, E_2, \dots, E_N)'$, defined by

$$(3.58) \quad E = Y - XB$$

vary less than the values of the target variable itself. Application of ordinary least squares results in

$$(3.59) \quad B = (X'X)^{-1} X'Y = \left(\sum_{k=1}^N X_k X_k' \right)^{-1} \left(\sum_{k=1}^N X_k Y_k \right).$$

For any sampling design, the vector B can be estimated by

$$(3.60) \quad b = \left(\sum_{k=1}^N a_k \frac{X_k X_k'}{\pi_k} \right)^{-1} \left(\sum_{k=1}^N a_k \frac{X_k Y_k}{\pi_k} \right).$$

The estimator b is an asymptotically design unbiased (ADU) estimator of B . It means the bias vanishes for large samples. Using expression (3.60), the generalized regression estimator is defined by

$$(3.61) \quad \bar{y}_{GR} = \bar{y}_{HT} + (\bar{X} - \bar{x}_{HT})' b,$$

where \bar{x}_{HT} and \bar{y}_{HT} are the Horvitz–Thompson estimators for the population means of X and Y , respectively. The generalized regression estimator is an ADU estimator of the population mean of the target variable. If there exists a p -vector c of fixed numbers such that $Xc = J$, where J is a vector consisting of 1's, the generalized regression estimator can also be written as

$$(3.62) \quad \bar{y}_{GR} = \bar{X}' b.$$

This situation occurs if X contains an intercept term or a set of dummy variables corresponding to all categories of a categorical variable. It can be shown that the variance of the generalized regression estimator can be approximated by

$$(3.63) \quad V(\bar{y}_{GR}) = \frac{1}{N^2} \sum_{k=1}^N \sum_{l=1}^N (\pi_{kl} - \pi_k \pi_l) \frac{E_k E_l}{\pi_k \pi_l}.$$

This is the variance of the Horvitz–Thompson estimator but with the values Y_k replaced by the residuals E_k . This variance will be small if the residual values E_k are small. Hence, the use of auxiliary variables that can explain the behavior of the target variable will result in a precise estimator.

Given simple random sampling without replacement and use of just one continuous auxiliary variable, the generalized regression estimator reduces to the regression estimator defined in equation (3.51).

Suppose a categorical auxiliary variable is available with p categories. Then this variable can be replaced by p dummy variables. Associated with each element k is a vector $X = (X_1, X_2, \dots, X_p)'$ of dummy values. The h th dummy X_{kb} assumes the value 1 if element k belongs to stratum b , and it assumes the value 0 if it belongs to another stratum. In this case, B turns out to be equal to

$$(3.64) \quad B = (\bar{Y}^{(1)}, \bar{Y}^{(2)}, \dots, \bar{Y}^{(L)})',$$

and this vector can be estimated unbiasedly by the vector

$$(3.65) \quad b = (\bar{y}_{HT}^{(1)}, \bar{y}_{HT}^{(2)}, \dots, \bar{y}_{HT}^{(L)})'$$

of Horvitz–Thompson estimators of the stratum means. The vector of population means of the L auxiliary variables turns out to be equal to

$$(3.66) \quad \bar{X} = (W_1, W_2, \dots, W_L)',$$

where $W_b = N_b / N$. If we substitute these quantities in expression (3.62), the result is

$$(3.67) \quad \bar{y}_{PS} = \frac{1}{N} \sum_{b=1}^L N_b \bar{y}_{HT}^{(b)}.$$

This is the poststratification estimator (3.54) but written down for arbitrary sampling designs. So, the poststratification estimator is indeed a special case of the generalized regression estimator.

3.3 Application

Some sampling concepts introduced in this chapter are illustrated using a fictitious target population. This population consists of 8,000 shops in a large town. The objective of a web survey is estimation of the average yearly value of theft of goods in a shop. The target variable of the survey is the value of the goods stolen in a shop in a specific year. There is also an auxiliary variable, and this is the floor space of the shop.

The population has been generated such that there is an approximate linear relationship between the value of the stolen goods and the floor space of a shop. Figure 3.4 shows the scatter plot of these two variables. There is a clear relationship. Note that the variation of the value of the stolen goods increases as the size of the shop increases. Also note that the distribution of floor space is very skew. There are many small shops and only a few big shops.

A business register is available containing (among other address information) the e-mail address of a contact person of each business. Therefore, a web survey can be conducted for which the sample is drawn by means of a probability sample.

Suppose a simple random of size $n = 100$ is selected without replacement from this population of shops, and the sample mean of the value of the stolen goods is used as an estimator of the population mean of the value of stolen goods. Insight into the distribution of the estimator can be obtained by repeating the selection of the sample and the computation of the estimate a large number of times. Figure 3.5 shows the results of this experiment for 1,000 repetitions. The dotted line represents the population mean to be estimated (223.61). As proven by Bowley (1906, 1926), the sample mean has approximately a normal distribution. The distribution is symmetric around the population value, implying that the estimator is unbiased.

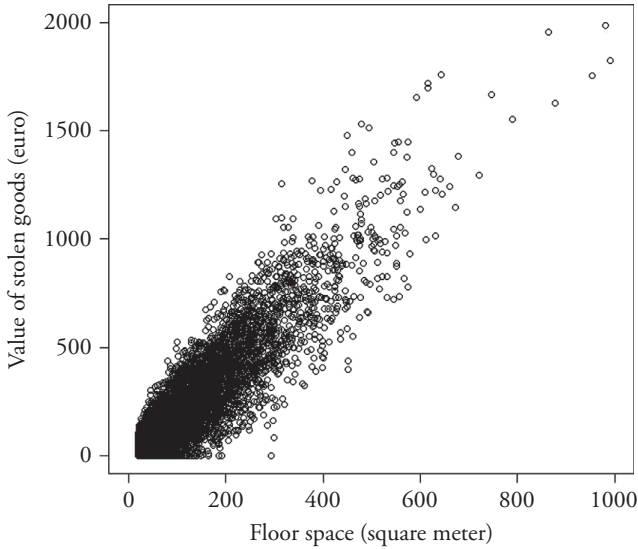


FIGURE 3.4 The relationship between the value of stolen goods and floor space

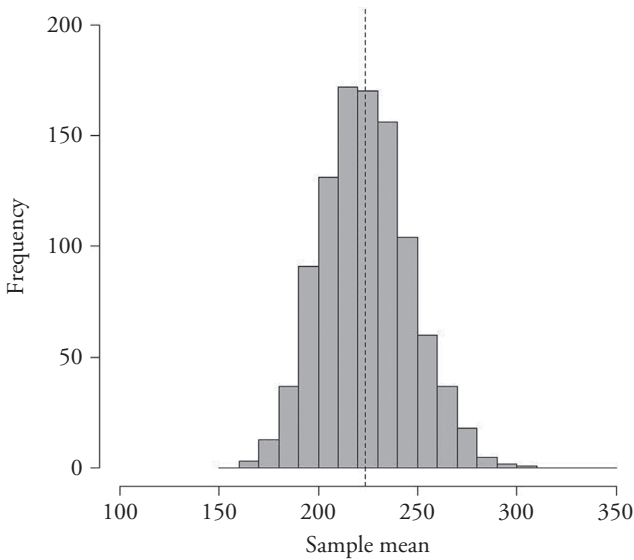


FIGURE 3.5 The distribution of the sample mean of the value of stolen goods ($n=100$)

The standard error of the sample mean is equal to 23.00. Hence, the margin of error of the 95% confidence interval = $1.96 \times 23.00 = 45.08$. Estimates will therefore not differ more than 45.08 from the true population mean (in 95% of the cases).

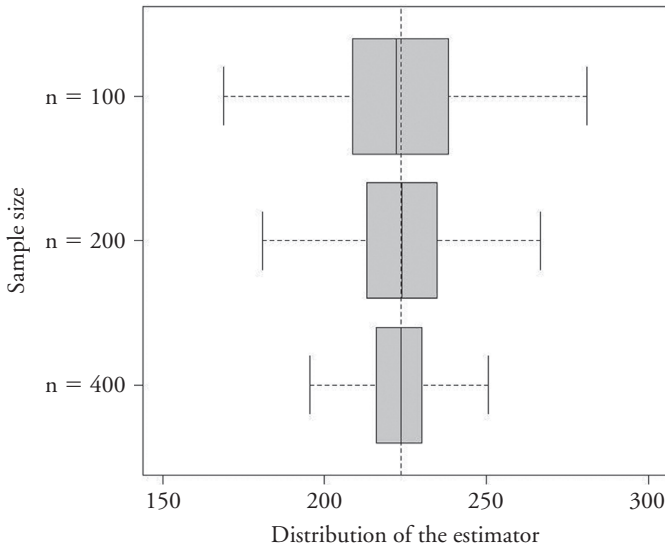


FIGURE 3.6 Distribution of the estimator of the value of stolen goods for simple random sampling with various sample sizes

The precision of estimates is for a large part determined by the sample size: The precision will be larger as the sample size increases. This is shown in Figure 3.6. The distribution of three different estimators is displayed by means of box plots. The distribution of each estimator was obtained by drawing 1,000 samples. In all three cases, the sample design was simple random sampling without replacement, but the sample sizes were 100, 200, and 400. All three distributions are symmetric around the population mean 223.61. Therefore, the estimators are unbiased.

It is clear that the variation of the estimates decreases as the sample size increases. The margin of error for a sample of size 100 is 45.08, and for a sample of size $n = 400$, it is 21.91. It is a rule of thumb that the precision is proportional to the square root of the sample size.

Business registers often contain additional information on companies, like type of economic activity and size of the company (number of employees). It is assumed in this application that floor size is included in this register. This makes it possible to draw a stratified sample. The shops are divided into four (floor) size groups. The characteristics of the four strata are summarized in Table 3.2.

The variance of the shoplifting value is in the first three strata much smaller than in the population as a whole. Apparently these three strata are fairly homogeneous. This makes stratified sampling a promising approach.

The last two columns contain the allocation of a sample of size 100 for optimal and proportional allocation. Note that in the case of optimal allocation, most elements (44) are drawn from the third stratum. The reason is that this

TABLE 3.2 Characteristics of the four size strata of the shops

Stratum	Number of elements	Mean	Variance	Optimal allocation	Proportional allocation
Less than 50	2160	51.785	51.044	12	27
50–100	2315	127.963	83.697	21	29
100–250	2733	301.444	150.686	44	34
250 and more	792	703.215	277.652	24	10
Total	8000	223.610	231.459	101	100

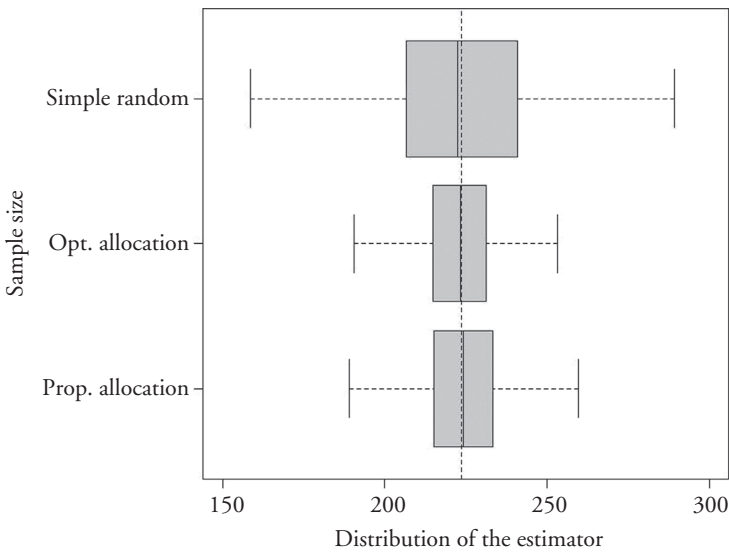


FIGURE 3.7 Distribution of the estimator of the value of stolen goods for simple random sampling and different types of stratified sampling

stratum is large and not very homogenous. Only 12 elements are selected from the first stratum because it is so homogeneous. Proportional allocation leads to a different distribution of the sample elements over the strata as the homogeneity of the strata is not taken into account. Stratified sampling is compared with simple random sampling in Figure 3.7. All distributions are based on 1,000 samples of size $n = 100$.

The distribution of the estimators for stratified sampling is shown for both optimal allocation and proportional allocation. They can be compared with the distribution of the sample mean. The precision of the stratification estimators is much higher than that of the sample mean. The margin of error of the sample mean is 45.08. In the case of stratification with optimal allocation, the margin of

error is only 22.96. This is the highest precision that can be obtained with a stratified sampling. In the case of proportional allocation, the margin of error is only slightly larger (26.27).

If the individual values of an auxiliary variable X are available for each element in the population, and all these values are positive, sampling with unequal probabilities can be considered. This only improves the precision of an estimator if there is a correlation between the target variable and the auxiliary variable. In the case of the shoplifting survey, the correlation coefficient of the value of the stolen goods in the shop and the floor space of the shop is 0.892, which is a strong correlation. The margin of error is only 19.97 for a sample of size 100. Hence, it pays (in terms of precision) to use the floor space as an auxiliary variable in the sampling design.

The precision of an estimator may be improved by using auxiliary information. Some examples of improved estimators were described in Section 3.2.7. Three estimators are applied to the fictitious shoplifting example: the ratio estimator, the regression estimator, and the poststratification estimator. All three estimator require an auxiliary variable for which the population distribution is known.

The ratio estimator is most effective if the values of the target variable and the auxiliary are proportional. The regression estimator is the estimator of choice if there is a general linear relationship between the values of the target variable and the auxiliary variable. The poststratification estimator can be used if the auxiliary variable is categorical and it divides the population into homogeneous strata.

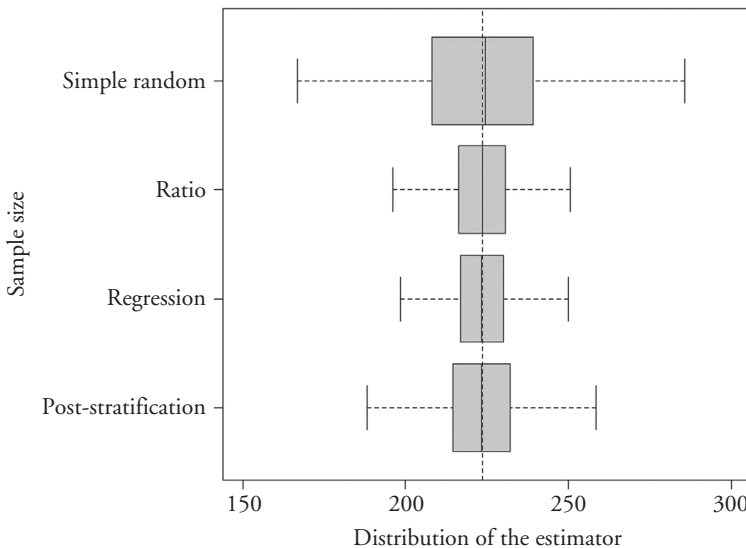


FIGURE 3.8 Distribution of the ratio, regression, and poststratification estimator

Figure 3.8 shows the distribution of these three estimators for the shoplifting example. Floor space is used as the auxiliary variable for the ratio estimator and the regression estimator. This auxiliary variable is transformed into a categorical variable with four categories (see Table 3.2) for the poststratification estimator. For reasons of comparison, also the distribution of the simple sample mean is shown.

All three estimators perform better than the simple sample mean. The regression estimator has the highest precision, closely followed by the ratio estimator. This is not surprising as there is a linear relationship between the value of stolen goods and the floor space of the shop. The ratio estimator assumes the constant term of the regression line to be 0. This is not the case here. Therefore, the ratio estimator does not perform as well as the regression estimator.

The poststratification estimator has approximately the same behavior as the estimator for stratified sampling with proportional allocation. The explanation that poststratification results (on average) in numbers of observations per stratum that are close to proportional allocations.

The conclusion of this application is that it pays to use auxiliary information, either in the sampling design or in the estimation procedure.

3.4 Summary

A web survey is an instrument to collect information about a specific population. Typically, not all elements are investigated in a survey but just a sample. This sample must have been selected by means of a probability sample where every element in the population has a positive probability of selection.

There are various ways to do this. Common sampling designs are simple random sampling, stratified sampling, sampling with unequal probabilities, cluster sampling, and two-stage sampling.

If every element in the population has a known and positive probability of selection, it is always possible to define a unbiased estimator. This is the so-called Horvitz–Thompson estimator. This estimator can be improved by taking advantage of available auxiliary variables. Examples of such estimators are the ratio estimator, the regression estimator, and the poststratification estimator.

KEY TERMS

Allocation: The distribution of the sample over the strata in stratified sampling. Optimal (or Neyman) allocation results in the highest precision of the estimator. It requires knowledge of the variances in the strata. If this is not the case, proportion allocation can be used.

Cluster sampling: A sampling design in which the population has been divided into a number of clusters. A sample of clusters is drawn, and all elements in the selected clusters are included in the sample.

Horvitz–Thompson estimator: An unbiased estimator that can be computed if the selection probabilities of all elements are known and positive.

Poststratification estimator: An estimator that can be computed if the sizes of all strata in a stratified population are available. It is a precise estimator if the strata are homogeneous.

Ratio estimator: An estimator that can be computed if the population mean of an auxiliary variable is available. It is a precise estimator if the target variable and auxiliary variable are (approximately) proportional.

Regression estimator: An estimator that can be computed if the population mean of an auxiliary variable is available. It is a precise estimator if there is a (approximate) linear relationship between target variable and auxiliary variable.

Simple random sampling: A sampling design in which elements are selected with equal probabilities.

Stratified sampling: A sampling design in which the population is divided into a number of strata and where a random sample is drawn from each stratum.

Two-stage sampling: A sampling design in which the population is divided into a number of clusters. A sample of clusters is drawn. From each selected cluster, a sample of elements is drawn.

Unequal probability sampling: A sampling design in which elements are selected with probabilities that are proportional to the values of some auxiliary variable.

EXERCISES

Exercise 3.1. Suppose a simple random sample of size 100 is selected from a population of size 1,000. What is the value of the second-order inclusion probability for every pair of elements?

- a. 0.1.
- b. 0.09.
- c. 0.01.
- d. 0.0099.

Exercise 3.2. Suppose a simple random sample of size 1,000 is selected from a population of size 16,000,000. What would happen to the variance of the sample mean if the sample size was doubled to 2,000?

- a. It would be twice as small.
- b. It would be approximately twice as small.
- c. It would be twice as big.
- d. It would be approximately twice as big.

Exercise 3.3. Under which condition will stratified sampling lead to more precise estimates than the simple sample mean?

- a. The values of the target variable vary little within strata.
- b. The stratum means vary little.
- c. The strata are all of approximately the same size.
- d. The subsamples in all strata are of the same size.

Exercise 3.4. Which of the statements below does not apply to cluster sampling?

- a. It is not clear beforehand how large the sample size will be.
- b. It can reduce the travel costs of interviewers.
- c. No sampling frame at all is necessary.
- d. Generally, it will lead to more precise estimators than simple random sampling.

Exercise 3.5. Under which condition is the variance of the regression estimator smaller than the variance of the simple sample mean?

- a. This is always the case.
- b. Only if the correlation between target variable and auxiliary variable is greater than 0.
- c. Only if the correlation between target variable and auxiliary variable is equal to 1.
- d. Only if the squared correlation between target variable and auxiliary variable is greater than 0.

Exercise 3.6. What is the effect of an increasing sample size on the value of the confidence level of the confidence interval?

- a. It remains 0.95.
- b. It increases in size and approaches 1.00.
- c. It decreases in size and approaches 0.00.
- d. It remains 0.05.

REFERENCES

- Bethlehem, J. G. (2009), *Applied Survey Methods, a Statistical Perspective*. John Wiley & Sons, Hoboken, NJ.
- Beukenhorst, D. and Wetzels, W. (2009), *A comparison of Two Mixed-Mode Designs of the Dutch Safety Monitor : Mode Effects, Costs, Logistics*. Technical paper DMH 206546, Statistics Netherlands, Methodology Department, Heerlen, the Netherlands.

- Bowley, A. L. (1906), Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society*, 69, pp. 548–557.
- Bowley, A. L. (1926), Measurement of the Precision Attained in Sampling. *Bulletin of the International Statistical Institute* XII, Book 1, pp. 6–62.
- Cochran, W. G. (1977), *Sampling Techniques*, 3rd ed. John Wiley & Sons, New York.
- Horvitz, D. G. & Thompson, D. J. (1952), A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, pp. 663–685.
- Scherpenzeel, A. (2008), An Online Panel as a Platform for Multi-disciplinary Research. In: Stoop, I. & Wittenberg, M. (eds.), *Access Panels and Online Research, Panacea or Pitfall?* Aksant, Amsterdam, pp. 101–106.

Errors in Web Surveys

4.1 Introduction

Survey researchers have control over many different aspects of a survey. With the proper choice of a sampling frame, a sampling design, and an estimation procedure, they can obtain precise estimators of population characteristics. Unfortunately not everything is under control. Survey researchers may be confronted with various phenomena that may have a negative impact on the quality and, therefore, the reliability of the survey outcomes. Some of these disturbances are almost impossible to prevent. Efforts will then have to be aimed at reducing their impact as much as possible. Nevertheless, notwithstanding all efforts to eliminate or reduce problems, final estimates of population parameters may be distorted. Estimates differ from the true value to be estimated. This difference is called the *total error* of the estimate.

Errors may occur in surveys whatever the mode of data collection, but some errors are more likely to occur in some types of surveys. For an example, it makes a difference whether interviewers conduct interviews or the respondents complete the questionnaires themselves. Focus in this chapter is on errors in web surveys. Sometimes the impact of a specific type of error in a web survey is compared with other types of surveys.

Sources of error will, if present, increase the uncertainty with respect to the correctness of estimates. This uncertainty can manifest itself in the distribution of an estimator in two ways: (1) it can lead to a systematic deviation (bias) from the true population value, or (2) it can increase the variation around the true value of the population parameter.

Let \bar{y}_E be an estimator for the population mean \bar{Y} . Chapter 3 discusses the properties of a good estimator. One is that an estimator must be *unbiased*. This

means its average value over all possible outcomes must be equal to the population mean to be estimated:

$$(4.1) \quad E(\bar{y}_E) = \bar{Y}.$$

An estimator may be biased as a result of survey errors. Suppose one objective of a survey is to estimate the average amount of time per day people spend on the Internet. If a web survey is conducted for this, people without Internet access will not be in the sample. As these people do not spend time on the Internet, the estimate will be too high. The estimator has an upward bias. This bias of the estimator \bar{y}_E is denoted by

$$(4.2) \quad B(\bar{y}_E) = E(\bar{y}_E) - \bar{Y}.$$

Another desirable property of an estimator is that its variance is as small as possible. This means that

$$(4.3) \quad V(\bar{y}_E) = E[\bar{y}_E - E(\bar{y}_E)]^2$$

must be small. An estimator with a small variance is called *precise*. An estimator can be made more precise by increasing the sample size or by using auxiliary information. This is discussed in more detail in chapter 3.

A precise estimator may still be biased. Therefore, just the value of the variance itself is not a good indicator of how close estimates are to the true value. A better indicator is the *mean square error*. This quantity is defined by

$$(4.4) \quad M(\bar{y}_E) = E(\bar{y}_E - \bar{Y})^2.$$

It is the expected value of the squared difference of the estimator from the value to be estimated. Writing out this definition leads to a different expression for the mean square error:

$$(4.5) \quad M(\bar{y}_E) = V(\bar{y}_E) + B^2(\bar{y}_E).$$

The mean square error contains both sources of uncertainty: a variance component and a bias component. The mean square error of an estimator is equal to its variance if it is unbiased. A small mean square error can only be achieved if both the variance and the bias are small. Figure 4.1 distinguishes four different situations that may be encountered in practice.

The vertical line in each graph represents the population mean. The distribution in the upper left corner shows the ideal situation for an estimator: It is precise and unbiased. All possible outcomes are close to the true value, and there is no systematic overestimation or underestimation. The situation in the lower left corner is less attractive. The estimator is still unbiased, but it has a substantial variance. Confidence intervals will be wider. Reliability is not affected. The confidence level of a 95% confidence interval remains 95%. The only difference is that these intervals are wider. Correct conclusions can still be drawn, but they are less precise.

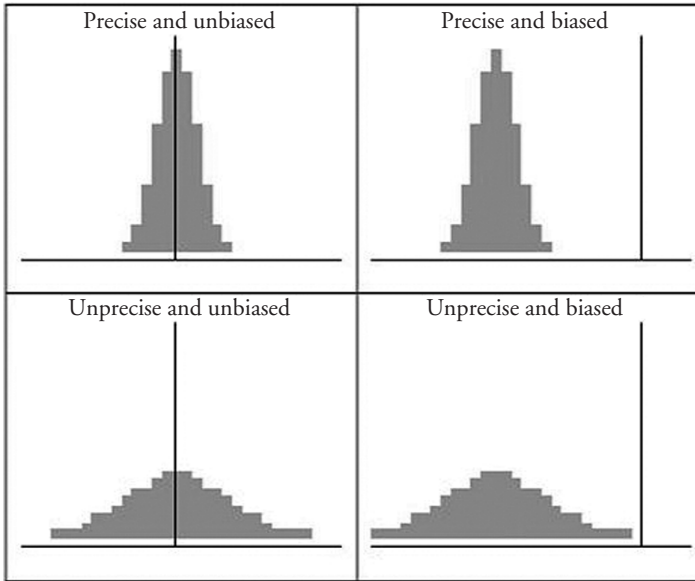


FIGURE 4.1 The bias and precision of an estimator

The situation is completely different for the graph in the upper right corner. The estimator is precise but has a substantial bias. The confidence interval computed using the survey data will almost certainly not contain the true value. The confidence level is seriously affected. Estimates are unreliable. Most likely, wrong conclusions are drawn. The graph in the lower right corner offers the highest level of uncertainty. The estimator is biased, and moreover, it is also not precise. This is the situation in which the mean square error has its largest value.

Survey estimates are never exactly equal to the population characteristics they intend to estimate. There will always be some error. This error can have many causes. Bethlehem (2009) describes a classification of possible causes. It is shown in Figure 4.2. This classification is an extended version of the one described by Kish (1967).

The ultimate result of all these errors is a discrepancy between the survey estimate and the population parameter to be estimated. Two broad categories of phenomena can be distinguished contributing to this total error: sampling errors and nonsampling errors.

Sampling errors are introduced by the sampling design. They occur because estimates are based on a sample from the population and not on a complete enumeration of the population. Sampling errors vanish if the complete population is observed. As only a sample is available for computing population characteristics, and not the complete data set, one has to rely on estimates. The sampling error can be split into an estimation error and a specification error.

The *estimation error* denotes the effect caused by using a probability sample. Every new selection of a sample will result in a different set of elements and, thus,

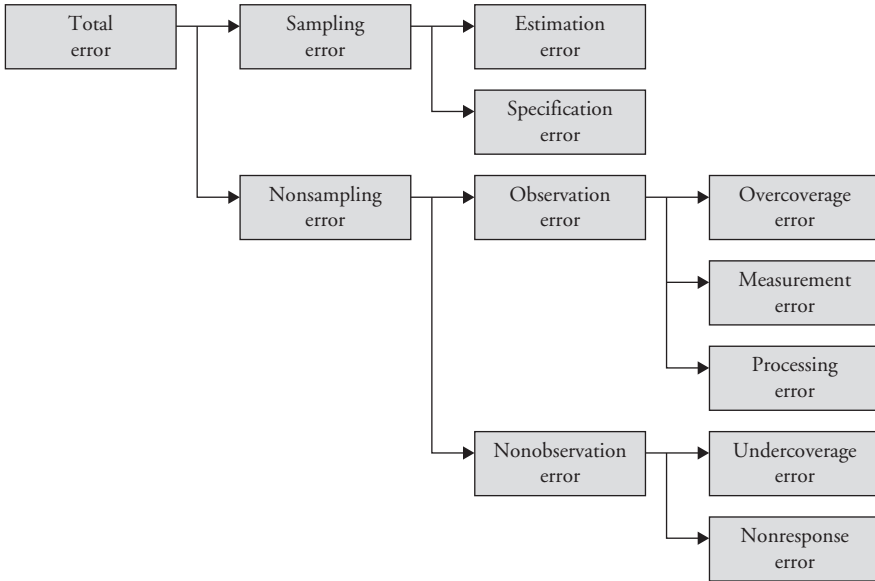


FIGURE 4.2 A classification of survey errors

in a different value of the estimator. The estimation error can be controlled through the sampling design. For example, the estimation error can be reduced by increasing the sample size, or by taking selection probabilities proportional to the values of some well-chosen auxiliary variable. Sampling errors are unrelated to the mode of data collection.

A *specification error* occurs when wrong selection probabilities are used in the computation of an estimator. It is shown in Chapter 3 that it is always possible to construct an unbiased estimator, provided the correct selection probabilities are known and used in the estimator. This is the so-called Horvitz–Thompson estimator. If wrong selection probabilities are used, this estimator will be biased. The differences between anticipated and true selection probabilities may be caused by problems in the sampling frame.

EXAMPLE 4.1 Specification errors in an address sample

A researcher intends to select a simple random sample of persons. He does this by drawing a simple random sample of addresses and by picking one person at random at each selected address. If he assumes selection probabilities to be equal and consequently uses the sample mean as an estimator for the population mean, there will be a specification error.

The true selection probabilities are not equal. They depend on the number of persons living at an address. For example, persons in large households have a smaller selection probability than single persons.

Members of large households will be underrepresented in the sample. If the target variable of the survey is related to the size of the household, the sample mean will be biased. For example, if the target variable would be to number the hours the members of the household spend on the Internet, the value of the sample mean will have a substantial downward bias.

Self-selection web surveys are affected by a special type of specification error. As respondents are recruited by self-selection, the true selection probabilities are unknown. Usually the researcher assumes all selection probabilities are equal, so that the sample mean can be used as an estimator of the population mean. Unfortunately, self-selection probabilities tend to depend on the characteristics of the respondents, and therefore, they may vary substantially. So, true selection probabilities differ from anticipated probabilities, resulting in a specification error.

True selection probabilities may differ from anticipated selection probabilities because of problems in the sampling frame. For example, if elements have multiple occurrences in the sampling frame, their selection probabilities will be larger. This type of selection error can only be detected and avoided by thorough investigation of the sampling frame.

Nonsampling errors are caused by problems that even can occur if the whole population is investigated (instead of a sample). They denote errors made during the process of obtaining answers to questions asked. Nonsampling errors can be divided into observation errors and nonobservation errors.

Observation errors are one cause of nonsampling errors. They refer to errors made during the process of obtaining answers from respondents and of recording and further processing these answers. Three types of observation errors are distinguished here: overcoverage errors, measurement errors, and processing errors.

An *overcoverage error* is caused by elements that are included in the survey and that do not belong to the target population. Such elements should not be included in the survey. They should be ignored. If they are included, they “pollute” the collected data and, hence, may lead to errors in estimators.

EXAMPLE 4.2 A survey about road pricing

There was an intensive political discussion in the Netherlands in January 2010 about the introduction of a system of road pricing. An important participant in this discussion was the Dutch Automobile Association (ANWB). This organization conducted a poll on its website. It was a self-selection survey. The poll was supposed to determine the opinion of the ANWB members about road pricing. So, the target population consisted of all members of the association. However, everyone with Internet could participate. Children could complete the questionnaire as well as people in

other countries. There was no restriction. Everyone could complete the questionnaire even more than once. There was nothing preventing this.

Fortunately, all respondents were asked whether they were a member of the ANWB. Consequently, the overcoverage (nonmembers) could be excluded from the analysis.

Within a period of a few weeks, the questionnaire was completed more than 400,000 times. Approximately 50,000 respondents indicated they were not an ANWB member. So there was substantial overcoverage.

Note that people may not have answered the question about membership properly. There is always a risk of socially desirable answers. Hence, there could still be hidden overcoverage.

A *measurement error* occurs if the answers given by respondents differ from the true answer. They may not understand a question, do not know the true answer, or do not want to give the true answer. This is caused, for example, by interview effects, question wording effects, and memory effects belong to this group of errors. Measurement errors are an important source of errors in web surveys. Therefore, a separate section (Section 4.2.1) is devoted to this problem.

A *processing error* occurs if errors are made during the phase of recording and processing the collected data. In the case of surveys with paper questionnaire forms, respondents or interviewers can make errors in writing down the answer. Such problems cannot occur in web surveys. However, mistakes can also be made if the questionnaire is on the Internet. It is easy to click on the wrong answer.

It is not uncommon in a survey that new variables are derived from those that have been measured in the survey. For example, one key variable of the Dutch Labor Force Survey is employment status in three categories: “employed”, “unemployed”, and “not in labor force”. People are considered unemployed if they are working less than 12 hours a week and are actively seeking and available for one or more jobs with a total of more than 12 hours a week. Putting people in the right category requires the answers to a set of questions. Deriving new variables from existing ones is usually done in computer programs. The algorithms for this can sometimes be complex, particularly if they must cope properly with every possible practical situation. Programming errors can be the cause of processing errors.

Nonobservation errors are the second cause of nonsampling errors. These errors occur because intended measurements cannot be carried out. Two types of nonobservation errors are distinguished: undercoverage errors and nonresponse errors.

Undercoverage occurs if elements in the target population do not have a corresponding entry in the sampling frame. These elements can and will never be selected for the survey. Undercoverage can be a problem in web surveys if the target population is wider than just those with access to Internet. This typically happens if general-population surveys are conducted by means of the Internet. The undercoverage errors can be substantial. Therefore, a separate chapter is devoted to it: Chapter 8.

Another type of nonobservation error is caused by *nonresponse*. Nonresponse is the phenomenon that elements selected in the sample do not provide the required information, or that the collected information is useless. Nonresponse occurs in almost every survey, whatever the mode of data collection. Some specific groups respond better than other groups. Therefore, some groups are overrepresented and other groups are underrepresented. This leads to biased estimates. Section 4.2.2 describes the nonresponse problem in more detail.

The classification above makes clear that many things can go wrong during the process of collecting survey data, and usually it does. Some errors can be avoided by taking preventive measures at the design stage. However, some errors will remain. Therefore, it is important to check collected data for errors and, where possible, to correct these errors. This activity is called *data editing*. Data editing procedures are not able to handle every type of survey error. They are most suitable for detecting and correcting measurement errors, processing errors, and possibly overcoverage. Phenomena like selection errors, undercoverage, and nonresponse require a different approach.

4.2 Theory

4.2.1 MEASUREMENT ERRORS

Measurement error is a general concept that is also used outside the field of survey research. In general terms, it can be defined as the difference between the actual value of a quantity and the value obtained by a measurement. The measuring instrument can have a *random error* if repeated measurements produce values that vary around the true value. This may be caused by the limited precision of the instrument. It can also have a *systematic error* if repeated measurements produce values that are systematically too high or too low. This may, for example, be caused by incorrect calibration of the instrument.

The measuring instrument of a survey is the questionnaire. This is not a perfect instrument. A measuring staff can be used for measuring someone's length, and the weight of a person can be determined by a weighing scale. These physical measuring devices are generally very accurate. The situation is different for a questionnaire. It only indirectly measures someone's behavior or attitude. Schwarz et al. (2008) describe the tasks involved in answering a survey question:

- Step 1: The respondents need to understand the question. They have to determine the information they are asked to provide. If they do not understand the question, they may decide to rephrase the questions and answer this question.
- Step 2: They need to retrieve the relevant information from their memory. In the case of a nonfactual question (for example, an opinion question), they will not have this information readily available. Instead, they have to form an opinion on the spot using whatever information comes to mind. In the case

of a factual question (for example, a question about behavior), they have to retrieve from their memory information about events in the proper time period.

- Step 3: They have to translate the relevant information in a format fit for answering the questions. Although they may have an answer in mind, it is often not in the proper format. For example, they may have to reformat their answer so that it fits one of the answer options of a closed question.
- Step 4: Respondents may hesitate to give their answer. If the question is about a sensitive topic, they may decide to refuse to give an answer. And if an answer is socially undesirable, they may change their answer.

The process of answering questions is a complex one. Several things can go wrong in this process, leading to measurement errors. Problems with the survey questions will affect the quality of the collected data and, consequently, the survey results. It is of the utmost importance to design and test the survey questionnaire carefully. It is sometimes said that questionnaire design is an art and not a skill. Nevertheless, long years of experience have led to several useful principles. Several issues are described in this section.

Kalton and Schuman (1982) distinguish factual and nonfactual questions. *Factual questions* are asked to obtain information about facts and behavior. There is always an individual true value. This true value could also be determined, at least in theory, by some other means than asking a question of the respondent. Examples of factual questions are as follows: “What is your regular hourly rate of pay on this job”, “do you own or rent your place of residence,” and “do you have an Internet connection in your home?”

The fact to be measured by a factual question must be precisely defined. It has been shown that even a small difference in the question text may lead to a substantially different answer. As an example, a question about the number of rooms in the household can cause substantial problems if it is not clear what constitutes a room and what does not. Should a kitchen, bathroom, hall, and landing be included?

Nonfactual questions ask about attitudes and opinions. An *opinion* usually reflects views on a specific topic, like voting behavior in the next general elections. An *attitude* is a more general concept, reflecting views about a wider, often more complex issue. With opinions and attitudes, there is no such thing as a true value. They measure a subjective state of the respondent that cannot be observed by another means. The attitude only exists in the mind of the respondent.

Various theories explain how respondents determine their answer to an opinion question. One such theory is the *online processing model* described by Lodge, Steenbergen, and Brau (1995). According to this theory, people maintain an overall impression of ideas, events, and persons. Every time they are confronted with new information, this summary view is updated spontaneously. When they have to answer an opinion question, their response is determined by this overall impression. The online processing model should typically be applicable to opinions about politicians and political parties.

There are situations in which people do not have formed an opinion about a specific issue. They only start to think about it when confronted with the question. According to the *memory-based model* of Zaller (1992), people collect all kinds of information from the media and in contacts with other people. Much of this information is stored in memory without paying attention to it. If asked to answer an opinion question, respondents may recall some of the relevant information stored in memory. Because of the limitations of the human memory, only part of the information is used. This is the information that immediately comes to mind when the question is asked. This is often information that only recently has been stored in memory. Therefore, the memory-based model can explain why people seem to be unstable in their opinions. The answer may easily be determined by the way the issue was recently covered in the media.

The remainder of this section is devoted to a description of several effects that can lead to measurement error. Where possible, it is indicated whether web surveys in particular are vulnerable to these effects.

4.2.1.1 Satisficing. If persons participate in a survey, they often will have to answer many questions. To do this properly requires a substantial cognitive effort. Although they may initially be motivated to do so, they are likely to become fatigued during the course of the interview. Interest in answering the questions will fade away. If the interview takes long to finish, they become impatient and distracted. As a result, they will devote less energy to answering the questions. As Krosnick (1991) describes it: Respondents are less thoughtful about the meaning of the questions, they search their memories less thorough, they integrate information more carelessly, and they may select an answer option more haphazardly. The first more or less acceptable answer that comes into mind is given. This phenomenon is called *satisficing*.

Holbrook, Green, and Krosnick (2003) argue that satisficing occurs more in telephone survey than in face-to-face surveys. Respondents of telephone survey may be more distracted because they might be engaged in different activities while they answer questions. This can be seen as a form of multitasking. Heerwegh and Loosveldt (2008) suggest that satisficing is even more a problem in web surveys. While respondents are answering questions they can also be involved in other activities on their computer, like answering email or visiting other websites. They also note that the cognitive demands of answering web survey questions are higher than those of an interviewer-assisted surveys.

Krosnick (1991) distinguished two forms of satisficing: weak satisficing and strong satisficing. *Weak satisficing* denotes the situation in which respondents still go through the four steps of answering a question, but they do it less thoroughly. They put less effort in attempting to understand the meaning of the question, they search their memory less well for relevant information, they may integrate the retrieved information more carelessly, and they will pick more easily an arbitrary answer option. *Strong satisficing* denotes the situation in which respondents simplify the answer process even more. They interpret each question only superficially by skipping steps 2 and 3 (retrieval and processing of information). They just pick an answer that seems reasonable.

Satisficing can come in many forms. Two forms of weak satisficing (response order effects and acquiescence) and four forms of strong satisficing (status quo endorsement, nondifferentiation, answering “don’t know,” and arbitrary answers) are described here.

4.2.1.2 Response Order Effects. *Response order effects* can occur if respondents are answering closed questions. They have to pick the proper answer from a (sometimes long) list of possible answer options. Instead of thinking carefully about which option is appropriate, the first reasonable option is chosen. In case of an interviewer-assisted survey (face-to-face or by telephone), the interviewer reads out loud the answer options. It is difficult for respondents to remember all options. Since only the last few options are still in their short-term memory, they restrict their judgment to these options. As a result, there is a preference for options near the end of the list. This is called a *recency effect*.

Self-administered surveys (web, mail) suffer, by contrast, from a *primacy effect*. This is the tendency to pick an answer early in the list of options. Reading to a list of possible options and considering each option, requires a considerable effort. Therefore respondents may stop at the first reasonable option.

Response order effects are described in more detail by Krosnick and Alwin (1987). According to the underlying theory, web surveys suffer from primacy effects. This was indeed the case in an experiment with the Dutch Safety Monitor that was described by Kraan et al. (2010). The effect was also shown by Schwarz, Hippler, Horst and Noelle-Neumann (1992); Sudman, Bradburn, and Schwarz (1996); and Couper et al. (2004).

Not only the order of the response options in a web survey matters, but also the format in which the response options are presented to the respondents. For a closed question, the HTML language offers various ways of displaying the possible answer options, and selecting one of them. Examples are shown in Figures 4.3–4.6. Figure 4.3 shows the use of *radio buttons*. An option is selected by clicking on the corresponding radio button. The advantage of this technique is that always only one answer is selected. Selecting a new answer deselects the

The image shows a screenshot of a web survey question. The question text is "In the last seven days, what type of music did you listen to most?". Below the question is a list of ten music genres, each preceded by a radio button. The "Country" option is selected, indicated by a filled-in radio button. The other options are unselected, indicated by empty radio buttons.

In the last seven days, what type of music did you listen to most?

- Chart / Top 40
- Dance
- Rock
- R & B
- Hip-hop
- Country
- Folk
- Easy listening
- Jazz
- Classical
- New age
- Other music

FIGURE 4.3 A closed question with radio buttons

currently selected answer. Once an answer is given, it is not possible to erase it and let the question remain unanswered.

Attention should be paid to displaying radio buttons when the list of possible answers is long. All question information should be visible on the screen, and not require any scrolling. Therefore, it is better to split a long list of radio buttons over a number of columns. Figure 4.4 gives an example. Here the answers are distributed over two columns. To give a visual clue that these two lists belong together, Dillman, Tortora, and Bowker (1998) suggest putting them in a kind of box. This is realized in Figure 4.4 by means of a gray background.

The HTML language also offers a different technique to select an item from a list. It is called a *drop-down box*. Figure 4.5 shows an example. The initial state of the box is shown on the left. Only one option is visible: the first one or the selected one. To select an answer, the respondent must click on the drop-down box, after which the list of possible answers becomes visible. This is shown on the right in Figure 4.5. If this list is very long, it only becomes partially visible. It depends on the browser used how long this list is. For example, 20 items are shown in Firefox 3.6 and 30 items in Internet Explorer 8. If the list is longer, scroll bars are provided to make other items visible. The respondent selects an answer by clicking on it in the list.

In the last seven days, what type of music did you listen to most?

<input type="radio"/> Chart / Top 40	<input type="radio"/> Folk
<input type="radio"/> Dance	<input type="radio"/> Easy listening
<input type="radio"/> Rock	<input type="radio"/> Jazz
<input type="radio"/> R & B	<input type="radio"/> Classical
<input type="radio"/> Hip-hop	<input type="radio"/> New age
<input checked="" type="radio"/> Country	<input type="radio"/> Other music

FIGURE 4.4 A closed question with two columns of radio buttons

In the last seven days, what type of music did you listen to most?

Click here to answer ▾

Click here to answer
 Chart / Top 40
 Dance
 Rock
 R & B
 Hip-hop
 Country
 Folk
 Easy listening
 Jazz
 Classical
 New age
 Other music

FIGURE 4.5 A closed question with a drop-down box

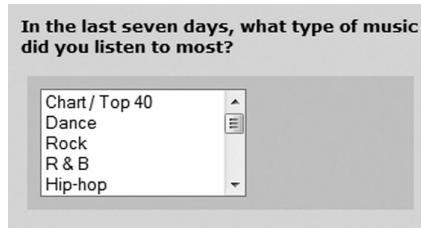


FIGURE 4.6 A drop-down box with a fixed number of items

Drop-down boxes have several disadvantages. In the first place, respondents have to do more work to select an answer (as compared with radio buttons). They have to perform three actions: clicking the box, scrolling to the right answer, and clicking this answer. In the second place, there can be serious primacy effects if only part of the list is displayed. And in the third place, it is unclear how much space the question requires on the screen.

It is possible to modify the behavior of the drop-down box, so that it always shows a fixed number of items in the list. Figure 4.6 shows an example in which this number is set to 5. The amount of space needed for such a question is now fixed and small. However, it suffers from a serious primacy effect. Couper et al. (2004) have shown that this effect is particularly large for the format in Figure 4.6. Therefore, where possible, radio buttons should be preferred.

The advantages and disadvantages of various answer formats of closed questions are also discussed in the studies by Couper (1999), Heerwegh and Loosveldt (2002), and Dillman (2007).

4.2.1.3 Acquiescence. A second form of weak satisficing is *acquiescence*. This is where respondents tend to agree with statements in questions, regardless of their content. They simply answer “yes.” Holbrook et al. (2003) and Krosnick (1999) suggest that acquiescence occurs partly because respondents only superficially think about the statement offered in the question. This will result in a confirmatory answer.

Acquiescence is typically a problem for agree/disagree, true/false, or yes/no questions. Respondents tend to answer agree, true, or yes irrespective of the topic of the question. Krosnick (1999) estimates the bias from acquiescence to be in the order of 10%. The literature suggests that acquiescence is more common among respondents with a lower socioeconomic status.

EXAMPLE 4.3 Bias caused by acquiescence

Schuman and Presser (1981) describe an experiment showing the effect of acquiescence. Respondents were randomly divided into two groups. The first group was asked to respond to the statement “*Individuals are more to blame than social conditions for crime and lawlessness in this country.*” There

were two possible answers: agree or disagree. The statement for the second group was reversed: "*Social conditions are more to blame than individuals for crime and lawlessness in this country.*" Table 4.1 shows the results of this experiment,

TABLE 4.1 Are individuals or social conditions to blame?

Statement	Agree	Disagree
Individuals are more to blame	59.6%	40.4%
Social conditions are more to blame	56.8%	43.2%

The percentages of respondents agreeing with the statement do not differ much. The differences are within the margin of error. These percentages are always higher than the percentages of people disagreeing, whatever the statement. Reversing the statement does not seem to have an effect on the percentages.

According to De Leeuw (1992), there is less acquiescence in self-administered surveys than in interviewer-assisted surveys. Respondents tend to agree more with statements made in questions if interviewers are present. Without interviewers, respondents may feel more anonymous and, therefore, will be more inclined to answer sensitive questions honestly.

This suggests that acquiescence will be less of a problem in web surveys and more in face-to-face and telephone surveys.

4.2.1.4 Endorsing the Status Quo. Surveys sometimes ask respondents to give their opinion about changes. A typical example is a question whether government should change its policy with respect to a specific issue. Here are some examples of such questions:

- Should the defense budget of the United States be increased or decreased?
- Should gun control laws in the United States become more strict or less strict?
- Should the monarchy in the Netherlands be abandoned?
- Should same-sex marriages be recognized legally, or should they be prohibited?
- Should new nuclear power plants be built in the country?

The easiest way to answer such a question without thinking is to select the option to keep everything the same. If the option of no change is not explicitly offered, not many respondents will insist on giving this answer. However, if this option is explicitly mentioned, the number of people selecting it will substantially increase. According to Krosnick (1991), the percentage selecting "no change" will increase by 10% to 40%.

The tendency to endorse the status quo raises the question of whether to include a middle category (representing “no change”) in a closed question. On the one hand, some researchers think it should be included. If it is not there, respondents with a neutral view are forced to give an answer that does not correspond to their attitude or opinion. On the other hand, there are researchers who think the middle option should be excluded. If it is there, it will be too easy for respondents to avoid giving a clear opinion.

■ EXAMPLE 4.4 Including or excluding a middle option

Kalton, Roberts, and Holt (1980) describe an experiment where the effect of including a middle option is determined. There were two random samples of approximately 800 persons each. One sample was offered a question with a middle category:

Do you think that drinking alcohol in moderation is good for your health or bad for your health, or do you think it makes no difference to your health?

The other sample had to answer the question without the middle category:

Do you think that drinking alcohol in moderation is good for your health or bad for your health?

The result is shown in Table 4.2. Note that even if there were no middle option, 19.6% of the people gave this answer. This was coded by the interviewers as such under the option “Other.” Offering the middle option increased the percentage of respondents in this category from 19.6% to 56.0%.

TABLE 4.2 The effect of offering a middle option

Response	With middle option	Without middle option
Good for health	26.4%	51.9%
Bad for health	12.1%	16.0%
Other	5.4%	12.5%
Makes no difference	56.0%	19.6%
Total	99.9%	100.0%

It should be noted that the middle category was the last option offered. Therefore, a recency effect also may contribute somewhat.

The good/bad ratio for the question with the middle option is 2.2, whereas without the middle option, it is 3.2. Apparently this ratio is affected by excluding a middle option: Respondents do not spread proportionally over the categories Good and Bad.

Similar experiments as described in Example 4.4 have been conducted by other researchers. Bishop (1987) showed that just mentioning the middle option in the question text (and not offering it explicitly as an answer option) also increases selection of the middle option. Furthermore, he showed that it makes a difference in a telephone survey whether the middle option is really placed in the middle of the set of answer option or at the end of it. In the latter case, the recency effect will cause even more respondents to choose for this option.

Heerwegh and Loosveldt (2008) compared a face-to-face survey with a web survey. They found that in the web survey (among students) more respondents selected the middle response option. They concluded the data collected by means of the web survey was of lower quality as a result of satisficing.

Note that there will be no recency effect in a web survey. Therefore, putting the middle response option at the end of the list of answer options will not increase selection of this option even more. Tourangeau, Couper, and Conrad (2004) conducted experiments showing a preference for the visual middle of the set of answer options although this option did not correspond to the conceptual middle of the options.

4.2.1.5 Nondifferentiation. Nondifferentiation is a form of satisficing that typically occurs if respondents have to answer a series of questions with the same response options. The original idea was that this would make it easier for respondents to answer the questions. Changing the response options from question to question would increase the cognitive burden of respondents.

Over time it has become clear that this approach is problematic because satisficing respondents tend to select the same answer for all these questions irrespective of the question content. For example, Heerwegh and Loosveldt (2008) compared a face-to-face survey with a web survey. They showed that respondents in the web surveys used less different scale values. So, there was more nondifferentiation.

A series of questions with the same set of answer options can be combined into a *grid question* or a *matrix question*. Each row of a matrix question represents a single question, and each column corresponds to an answer option. An example is shown in Figure 4.7.

At first sight, grid questions seem to have some advantages. A grid question takes less space on the questionnaire form than a set of single questions, and it provides respondents with more oversight. Therefore it can reduce the time it takes to answer questions. Couper, Traugott, and Lamias (2001) indeed

	Excellent	Very good	Good	Fair	Poor
1. How would you rate the overall quality of the radio station?	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. How would you rate the quality of the news programs?	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. How would you rate the quality of the sport programs?	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. How would you rate the quality of the music programs?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

FIGURE 4.7 An example of a matrix question

found that a matrix question takes less time to answer than a set of single questions.

According to Dillman, Smyth, and Christian (2009) answering a matrix question is a complex cognitive task. It is not always easy for respondents to link a single question in a row to the proper answer in the column. Moreover, respondents can navigate through the matrix in several ways, row-wise, column-wise, or a mixture of the two. This increases the risk of missing answers to questions, resulting in a higher item nonresponse.

Dillman et al. (2009) advise to limit the use of matrix questions as much as possible. If they are used, they should not be too wide or too long. Preferably, the whole matrix should fit on a single screen. This is not so easy to realize as different respondents may have set different screen resolutions on their computer screens. If respondents have to scroll, either horizontally other vertically, they may easily get confused, leading to wrong or missed answers.

Fricker et al. (2005) investigated differences between a web survey and a computer-assisted telephone interviewing (CATI) survey. They showed that the respondents in the web survey gave less differentiated answers to attitude questions in matrix format.

Several authors, see for example Krosnick (1991) and Tourangeau et al. (2004), express concern about a phenomenon that is sometimes called *straightlining*. Respondents give the same answer to all questions in the matrix. The simply check all radio buttons in the same column. Often this is the column corresponding to the middle response option. Figure 4.8 shows an example.

Straightlining the middle response option can be seen as a form of endorsing the status quo. It can also be seen as a form of nondifferentiation. It is a means of quickly answering a series of questions without thinking. It manifests itself in very short response times. So, short response times for matrix questions (when compared with a series of single questions) are not always a positive message. It can mean that there are measurement errors caused by satisficing.

If a matrix question is used, much attention should paid to its visual layout. For example, a type of shading as in Figure 4.8 reduces confusion and, therefore, reduces item nonresponse. *Dynamic shading* may even help more. Kaczmirek (2010) distinguishes preselection shading from postselection shading. *Preselection shading* comes down to changing the background color of a cell or row of the matrix question if the cursor is moved over it by the respondent. Preselection shading helps the respondent to locate the proper answer to the proper question. It is active before the answer is clicked. Postselection shading means shading of a

	Excellent	Very good	Good	Fair	Poor
1. How would you rate the overall quality of the radio station?	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. How would you rate the quality of the news programs?	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. How would you rate the quality of the sport programs?	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. How would you rate the quality of the music programs?	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

FIGURE 4.8 An example of a matrix question with straightlining

cell or row in the matrix after the answer has been selected. This feedback informs the respondent which answer to which question was selected. Kaczmirek (2010) concludes that particularly preselection and postselection shading of complete rows improves the quality of the answers. However, preselection shading of just the cell reduced the quality of the answers. There was more nondifferentiation.

Galesic et al. (2007) also experimented with postselection shading. The font color or the background color was changed immediately after respondents answered a question in the matrix. This helped respondents to navigate and, therefore, improved the quality of the data.

4.2.1.6 Don't Know. Questions are asked in a survey to collect information about respondents. However, respondents are sometimes not able to provide this information. They simply do not know the answer. If an opinion question is asked, it is usually assumed that respondents have an opinion with respect to the specific issue. This need not be the case. They simply may not have an opinion. Also, if a factual question is asked, respondents may lack the knowledge to answer it.

The question is how to ask a question in such a way that always a true answer is given: “don't know” if the respondent really does not know the answer and a “real” answer if the respondent has one. There are various approaches, and the most effective approach may depend on the type of survey.

On one end of the spectrum, one could offer “don't know” just as one of the answer options. This may lead to a form of satisficing where respondents choose this option to avoid having to think about a real answer. On the other end of the spectrum, one can decide to not offer “don't know” as an answer at all. So respondents are forced to give a real answer, even if they do not have one. It is clear that both approaches may lead to measurement errors.

Besides the two extremes mentioned there are more ways to handle “don't know.” Several approaches are presented here.

- *Offer “don't know” explicitly*

Offering “don't know” explicitly as one of the answer options has the advantage that respondents not knowing the answer can answer so. This approach accepts the existence of a group of persons that cannot answer the questions, and thus, “don't know” is considered a substantive answer.

This approach may suffer from satisficing. People not wanting to think about an answer or not wanting to give an answer have an escape by answering “don't know.” Several authors have shown that explicitly offering “don't know” substantially increases the percentage of respondents choosing this option. See, for example, Sudman and Bradburn (1982),

- *Offer “don't know” explicitly but less obviously*

To make it less easy for respondents to choose the option “don't know,” one could decide to offer this option but at the same time attempt to make it less obvious. The option could be placed elsewhere on the screen or shown in a smaller or less bright font.

Tourangeau et al. (2004) experimented with questions where the “don’t know” option was visually separated from the substantive options by a dividing line. This was counterproductive as it caused more respondents to select the “don’t know” options, as more attention was drawn to this option.

DeRouvray and Couper (2002) experimented with questions where the “don’t know” option was displayed in a smaller and lighter font so that its visual prominence was reduced. This did not affect the number of respondents selecting this option.

Vis-Visschers et al. (2008) offered “don’t know” as a special button at the bottom of the screen. It turned out that many respondents overlooked this option and complained they could not answer “don’t know”.

- *Offer “don’t know” implicitly*

To make “don’t know” an even less obvious option, one could decide not to offer it explicitly on the screen. This is common in CAPI/CATI-software. Only a list of substantive answer options is shown on the screen. If respondents insist they do not know the answer, the interviewer can record this by using a special key combination. For example, the option “don’t know” is always by default available in the Blaise system for computer-assisted interviewing by pressing <Ctrl-K>; see Statistics Netherlands (2002).

A special key combination may work for experienced interviewers but not for inexperienced respondents in web surveys. There are, however different ways to offer “don’t know” implicitly. Vis-Visschers et al. (2008) investigated an approach whereby respondents were offered questions with only substantive answer options. If they did not select an option and attempted to skip the question, the question was offered again, but then “don’t know” was included in the list. It turned out that some respondents did not understand this mechanism, as they complained that they could not answer “don’t know.” The other respondents less frequently selected “don’t know.”

DeRouvray and Couper (2002) experimented with a similar approach. The answer option “don’t know” was not offered for the question. If respondents attempted to skip the question without answering it, a new screen appeared offering two choices: (1) Go back and answer the question, and (2) record the answer as “don’t know” and proceed to the next question. This approach resulted in the lowest “don’t know” rates. It also resembles the CAPI/CATI approach.

- *Do not offer “don’t know”*

To avoid satisficing, one may decide not to offer the option “don’t know.” This implies that respondents always have to provide a substantive answer, even if they do not know the answer. According to Couper (2008) this violates the norm of voluntary participation. Respondents should have the option not to answer a question. Forcing respondents to answer may frustrate respondents resulting in a breakoff. Also Dillman (2007) strongly

recommends not forcing respondents to answer. He warns about detrimental effects to respondent motivation, data quality, and the risk of breakoff.

The treatment of “don’t know” and the effects this can have on the collected data depends on the type of survey. Heerwegh and Loosveldt (2008) compared the use of “don’t know” in a face-to-face survey with a web survey. There was a set of opinion questions about estimating parent’s views of immigrants. “Don’t know” was visually offered as the last response option. The respondents of the face-to-face survey were informed that it was possible to answer “don’t know.” The respondents of the web survey could see “don’t know” as one of the options on the screen. The option “don’t know” was selected much more frequently in the web survey. A possible explanation is that people without an opinion do not want to admit their ignorance to interviewers. They may feel foolish. Therefore, they feel forced to give a substantive answer although they lack relevant information for formulating a relevant judgment.

If there is a risk that “don’t know” is avoided to prevent embarrassment, one may consider using a filter question. See Krosnick (1991) and Schuman and Presser (1981). This filter question asks whether respondents have an opinion about a specific issue. If they say they have, then a next question asks what their opinion really is.

■ EXAMPLE 4.5 Using a filter question for “don’t know”

The Dutch Parliament discussed in 2007 a possible change in the Electricity Law. The main purpose of the change was to make the law consistent with European directives. This was not a controversial or otherwise interesting issue. Therefore, it was not taken up by the media, and in fact, the general public were not aware of the change in law.

Tiemijer (2008) conducted an experiment in which he measured the effect of a filter question for “don’t know.” The question asked is shown in Figure 4.9.

The parliament will discuss a proposal to change the 1998 Electricity law shortly. To what extent do you agree or disagree with the changes in the 1998 Electricity law??

- Strongly agree
- Somewhat agree
- Neutral
- Somewhat disagree
- Strongly disagree
- Country
- Don't know / No opinion

FIGURE 4.9 Question about the 1998 Electricity Law

The sample size was 395. The sample was divided randomly into three groups. The first group just answered the question in Figure 4.9. The second group also answered this question, but it was preceded by a weak filter question (*Do you have an opinion about the changes in the 1998 Electricity Law, or not?*). For the third group, there also was a filter question, but it was a stronger question (*Have you heard or read enough about this proposal to be able to form an opinion about it?*). Respondents answering *No* to the filter question were classified as “Don’t know/No opinion.”

TABLE 4.3 The effect of offering a middle option

Response	No filter question	Weak filter question	Strong filter question
Don’t know/no opinion	55%	79%	86%
Somewhat/strongly agree	7%	3%	3%
Neutral	23%	3%	2%
Somewhat/strongly disagree	15%	14%	9%
Total	100%	99%	100%

Table 4.3 contains the results. If no filter question were asked, the percentage of “don’t know” was 55%. This is a high percentage but lower than expected because the topic of the question was completely unknown to the respondents. This percentage increases to 79% for the weak filter question and even to 86% for the strong filter question. Apparently, a filter question makes it less embarrassing for respondents to admit they do not know the answer.

4.2.1.7 Arbitrary Answer. Selecting “don’t know” as an answer is one way for respondents to avoid having to think about a proper answer. If giving this answer is considered undesirable, respondents may also decide to just pick an arbitrary answer. Converse (1964) already describes the problem of random answers. Krosnick (1991) calls this behavior “metal coin flipping.”

This type of satisficing can also occur for a special type of question called the check-all-that-apply question. This is described by Dillman et al. (1998). An example of such a question is shown in Figure 4.10. It is a closed question for which more than one answer can be selected. It is common practice for web surveys to use square *check boxes* (instead of round *radio buttons*) for check-all-that-apply questions.

A check-all-that-apply question asks respondents to check all appropriate items from a (sometimes long) list of answer options. This can be a lot of work. Instead of checking all relevant answers, they may just check some arbitrary answers and stop when they think they have check enough answers. Moreover,

In the last seven days, what type of music did you listen to most?

- Chart / Top 40
- Dance
- Rock
- R & B
- Hip-hop
- Country
- Folk
- Easy listening
- Jazz
- Classical
- New age
- Other music

FIGURE 4.10 A check-all-that-apply question

satisficing respondents tend to read only the first part of the list, not the complete list. This causes a bias toward answers in the first part of the list.

It is common practice not to use check-all-that-apply questions in telephone surveys. Instead, respondents have to answer either yes (applies) or no (does not apply) for each item in the list. This raises the question of whether in web surveys check-all-that-apply questions should be replaced by sets of these forced choice questions too. This would mean changing the question format as in Figure 4.10 by a format as in Figure 4.11. Smyth et al. (2006) have shown the format in Figure 4.11 leads to more selected options, and respondents take more time to answer the questions. This is an indication that the format as in Figure 4.10 may cause satisficing.

In the last seven days, what type of music did you listen to most?

Yes No

- Chart / Top 40
- Dance
- Rock
- R & B
- Hip-hop
- Country
- Folk
- Easy listening
- Jazz
- Classical
- New age
- Other music

FIGURE 4.11 A check-all-that-apply question with forced choice

It should be noted that completing the answer to the forced choice questions requires more work, and this may frustrate respondents. It may lead to straightlining, a form of satisficing described in Section 4.2.1.5.

4.2.1.8 Socially Desirable Answers. With respect to data collection, there is a substantial difference between interviewer-assisted surveys (e.g. CAPI and CATI) on the one hand and self-administered surveys (web, mail) on the other. Interviewers carry out the fieldwork in a CAPI or CATI survey. There are no interviewers, however, in a web survey. It is a self-administered survey. Therefore, the quality of the collected data may be lower because of higher nonresponse rates and more errors in the answers to the questions.

De Leeuw (2008) and Dillman et al. (2009) discuss the differences between various modes of data collection. They observe that a positive effect of the presence of interviewers is that they are in control of the interview. They lead the respondent through the interview. They see to it that the right question is asked at the right moment. If necessary, they can explain the meaning of a question. They can assist respondents in getting the right answers to the question. Interviewers can motivate respondents, answer questions for clarification, provide additional information, and remove causes for misunderstanding. All this will increase the quality of the collected data.

The presence of interviewers can also have a negative effect. It will lead to more socially desirable answers for questions about potentially sensitive topics. Giving *socially desirable answers* is the tendency that respondents give answers that will be viewed as more favorable by others. This particularly happens for sensitive questions about topics like sexual behavior and use of drugs. If a true answer would not make the respondents look good, they will refuse to answer or give a different answer. A meta-analysis by De Leeuw (1992) shows that the effects of socially desirable answers are stronger in interviewer-assisted surveys. Respondents tend to give more truthful answers in self-administered surveys.

Holbrook and Krosnick (2010) describe an example of socially desirable answers in an election survey. Voting is seen as admirable and valued civic behavior. Therefore, nonvoters may be reluctant to admit they did not vote. This effect will occur particularly if the respondent is asked to report something embarrassing directly and explicitly. If the respondent can answer more anonymously and confidentially, a more truthful answer will be given. Holbrook and Krosnick (2010) find that a socially desirable answer leads to a bias in telephone surveys but not in web surveys.

Kreuter, Presser, and Tourangeau (2008) conducted an experiment in which they compared the effects of socially desirable answers in a web survey, a CATI survey, and an interactive voice recognition (IVR) survey. It is a telephone survey without interviewers. Questions are asked by software, and respondents answer by the telephone keypad or by giving a verbal answer. IVR can be placed somewhere between CATI and web in the spectrum from interviewer-assisted to self-administered. Kreuter et al. (2008) had administrative information available that enabled them to compare the given answers with the true answers. They

showed that the amount of correctly reported answers to sensitive questions in web surveys is higher than in the CATI and IVR surveys,

Kraan et al. (2010) describe an experiment in the Netherlands where two different data collection modes for the Safety Monitor are compared. They find that respondents in a web survey are more critical about the performance of the police than respondents in a CAPI or CATI survey. Apparently, respondents in the interviewer-assisted surveys are less inclined to make critical remarks about the police.

4.2.1.9 Some Web Survey Design Issues. Web survey questionnaires are offered in two different ways: the question-based approach and the form-based approach. Figure 4.12 shows an example of the *question-based approach*. There is always only one question on the screen. After the answer to the question has been entered, the respondent clicks on a button to go to the next question. Alternatively, the respondent can go back to previous question to change an answer.

The question-based approach is used if the questionnaire contains routing instructions. If the next question to be asked depends on the answers to one or more previous questions, the answers to these previous questions must first be processed by the web server before the next question can be shown on the screen.

The question-based approach is also used if consistency checks are carried out immediately after an answer has been entered. This might result in showing a warning or error message, after which the question reappears on the screen in order to correct the answer.

The question-based approach is typically used for a large and complex questionnaire where respondents remain on-line while answering questions. This approach is more or less similar to the approach used for CAPI or CATI.

Figure 4.13 shows an example of the *form-based approach*. It more or less mimics the paper questionnaire on the screen. This approach does not allow routing instructions in the questionnaire. It is also not possible to have checks before the end of the page is reached.

Respondents need not be on-line while they answer the questions. It is possible to download the form, go off-line, answer the questions, go on-line again, and upload the complete questionnaire. The form-based approach is generally recommended for small and simple questionnaires.

It is also possible to apply some kind of hybrid approach whereby a questionnaire is composed of several blocks of questions. Each block is presented as a form on the screen, like in the form-based approach. So the questionnaire

The screenshot displays a web survey interface. At the top left is a small image of a radio. To its right is the title 'Local Radio Listening Survey 2010'. On the top right, there is a 'Progress:' label followed by a horizontal progress bar. Below the title bar, the question '4. Did you listen to your local radio during the last seven days?' is presented. Underneath the question are two radio button options: 'Yes' and 'No'. At the bottom of the interface, there are two buttons: 'Previous' on the left and 'Next' on the right.

FIGURE 4.12 The question-based approach

Local Radio Listening Survey 2010

1. What is your gender?

Male
 Female

2. What is your age?

years

3. What is your marital status?

Never married
 Married
 Separated
 Divorced
 Widowed

4. Did you listen to your local radio during the last seven days?

Yes
 No

5. Which type of programs do you listen to most on your local radio station?

Music
 News and current affairs
 Sports
 Culture
 Other programs

FIGURE 4.13 The form-based approach

consists of a number of forms. There are no routing instructions or checks within blocks, but there might be between blocks. More about the advantages and disadvantages of both approaches can be found in the study by Couper (2008).

One advantage of the form-based approach is that respondents see how long the questionnaire is. This may encourage them to finish the questionnaire (but also discourage them if it is very long). The question-based approach lacks this overview as respondents only see one question at a time. To help respondents, it is sometimes advised to include some kind of progress indicator. Such an indicator informs respondents about where they are in the questionnaire. The example in Figure 4.12 contains a progress indicator. Such indicators can take many different forms, and research results are mixed about their usefulness. Progress indicators seem to help if they show quick progress, but they may have the opposite effect if progress is slow. It may be wise not to use progress indicators if the questionnaire has a complex routing structure. Then the number of questions to be answered may be different for different respondents. It may even depend on the answers to questions that have not yet been asked. Therefore, it is impossible to indicate progress. See Couper (2008), Heerwegh (2004), or Kaczmirek et al. (2004) for more on progress indicators.

The advantage of the question-based approach is that a single question always fits on the computer screen. This is not the case for the form-based

1. What month and year did you begin your studies at the university?

2. What month and year did you begin your studies at the university?

 Month Year

3. What month and year did you begin your studies at the university? Please provide your answer using two digits for the month and four digits for the year.

 Month Year

FIGURE 4.14 Formatting a date question

approach. If the form contains many questions, it may only be partially visible. Consequently, the respondent has to scroll to see other parts of the questionnaire. Dillman et al. (2009) warn against scrolling as it may increase the risk that respondents miss questions.

It is not uncommon in CAPI and CATI surveys to have consistency checks in the questionnaire. These checks are carried out immediately after the relevant questions have been asked. If an inconsistency is detected, an error message is displayed on the screen. The interviewer discusses the problem with the respondent, which may result in correcting the answers to one or more questions. Research shows that these checks improve the quality of the collected data.

Checks may also be included in a web survey, but one should be careful as there are no interviewers who can explain the problem to the respondents. Couper (2008) suggests that error messages should at least be polite, illuminating, and helpful. They should not contain threatening words and should not blame the respondent. Also Dillman et al. (2009) warn that unfriendly, unspecific, or unclear error messages may cause respondents to break off the interview.

The lack of interviewer assistance in a web survey makes it even more important to make clear to the respondents what is expected from them when they answer a question. The visual design of the question is of vital importance. Two examples illustrate this. The first example relates to asking for dates. Dates can be formatted in many different ways. If a researcher wants dates to be entered in a specific format, he should give as much guidance as possible to the respondents.

Figure 4.15 shows three different ways in which one could ask for the month and year in which a student started his/her university studies. The first question is formatted as an open question. Any text can be answered. The computer attempts to extract a date from the input. If this is not possible, an error message will appear. The second question has two input fields, one for the month and one for the year. It is still not clear whether the date has to be entered as words or as numbers. The third question also has two input fields. It is now indicated both in

FIGURE 4.15 Formatting an open question

the text of the question and below the input fields that digits are expected. Christian, Dillman, and Smyth (2007) conducted an experiment in which they compared, among others, the format of the second question with the format of the third question. Only 44% of the respondents entered their answer properly for the second format. The percentage of right answers was 94% for the third format.

A second example relates to the format of open questions. The size of the input box should reflect the amount of information that is expected. Example 4.14 shows two open questions. The input field of the first question suggests that only one line of text is sufficient. The input field of the second question suggests that respondents can enter several lines of text. The scrollbar even suggest that the text can be longer than the size of the box. Research shows that indeed the second format leads to longer answers than the first format.

4.2.1.10 Other Measurement Errors. Section 4.2.1.9 discusses the types of measurement errors that may be specific to web surveys. Other types of measurement errors also may occur irrespective of the mode of data collection. Some are mentioned in this section.

Questions requiring respondents to recall events that have happened in the past are a source of errors. The reason is that people make *memory errors*. They tend to forget events, particularly when they happened a long time ago or when they are not too salient. Important events, more interesting events, and more frequently happening events will be remembered better than other events. The effects of recall errors are more severe as the length of the reference period is longer.

The question in Figure 4.16 is a simple question to ask, but for many people, it is difficult to answer. Recall errors may even occur for shorter time periods. In the

FIGURE 4.16 A recall question

1981 Health Survey of Statistics Netherlands, respondents had to report contacts with their family doctor over the last three months. Memory effects were investigated by Sikkel (1983). It turned out that the percentage of not-reported contacts increased linearly in time. The longer ago an event took place, the more likely it is that it was forgotten. The percentage of unreported events for this question increased on average with almost 4% per week. Over the total period of three months, about one quarter of the contacts with the family doctor was not reported.

Recall questions may also suffer from *telescoping*. This occurs if respondents report events as having occurred either earlier or later than they actually did. As a result, an event is incorrectly reported within the reference period, or incorrectly excluded from the reference period. Bradburn, Sudman, and Wansink (2004) note that telescoping leads more often to overstating than to understating several events. Particularly, for short reference periods, telescoping may lead to substantial errors in estimates.

Question order can affect the results in two ways. One is that mentioning something (an idea, an issue, or a brand) in one question can make people think of it while they answer a later question, when they might not have thought of it if it had not been previously mentioned. In some cases, this problem may be reduced by randomizing the order of related questions. Separating related questions by unrelated ones might also reduce this problem, although neither technique will completely eliminate it.

Tiemeijer (2008) mentions an example where the answers to a specific questions were affected by a previous question. The Eurobarometer (www.europa.eu/public_opinion) is an opinion survey in all member states of the European Union (EU) that has been conducted since 1973. The European Commission uses this survey to monitor the evolution of public opinion in the member states. This may help in making policy decisions. The 2007 Eurobarometer contained the question shown in Figure 4.17.

It turned out that 69% of the Dutch respondents were of the opinion that the country had benefited from the EU. A similar question was included at the same time in a Dutch opinion poll (*Peil.nl*). However, the question was preceded by another question that asked respondents to select the most important disadvantages of being a member of the EU. Among the items in the list were the too fast extension of the EU, the possibility of Turkey becoming a member state, the introduction of the euro, the waste of money by the European Commission, the loss of identity of the member states, the lack of democratic rights of citizens, the veto rights of member states, and the possible interference of the European Commission with national issues. As a result, only 43% of the respondents considered membership in the EU beneficial.

8. Taking everything into consideration, would you say that the country has on balance benefited or not from being a member of the European Union?

- Yes
- No

FIGURE 4.17 A context-sensitive question

1. **An increase of the powers of the European Parliament will be at the expense of the national parliament.**
Do you think the powers of the European Parliament should be increased?

- Yes
 No

2. **Many problems cross national borders. For example, 50% of the acid rain in The Netherlands comes from other countries.**
Do you think the powers of the European Parliament should be increased?

- Yes
 No

FIGURE 4.18 Questions containing additional information

Opinion questions may address topics about which respondents may not yet have made up their mind. They may even lack sufficient information for a balanced judgment. Questionnaire designers may sometimes provide additional information in the question text. Such information should be objective and neutral and should not influence respondents in a specific direction. Saris (1997) performed an experiment to show the dangers of changes in the question text. He measured the opinion of the Dutch about increasing the power of the European Parliament. Respondents were randomly assigned one of the two questions in Figure 4.18.

In case respondents were offered the first question, 33% answered “yes” and 42% answered “no”. In case respondents were offered the second question, 53% answered “yes” and only 23% answered “no”. These substantial differences are not surprising, as the explanatory text in the first question stresses a negative aspect and the text in the second question stresses a positive aspect.

4.2.2 NONRESPONSE

4.2.2.1 The Nonresponse Problem. Nonresponse occurs when elements (persons, or companies) in the selected sample, which also are eligible for the sample, do not provide the requested information or the provided information is not usable. The problem of nonresponse is that the researcher does not have control any more over the sample selection mechanism. Therefore, it becomes impossible to compute unbiased estimates of population characteristics. The validity of inference about the population is at stake.

There are two types of nonresponse: unit nonresponse and item nonresponse. *Unit nonresponse* occurs when a selected element does not provide any information at all; i.e., the questionnaire form remains empty. *Item nonresponse* occurs when some questions have been answered but no answer is obtained for some other, possibly sensitive, questions. So, the questionnaire form has been partially completed. This section focuses on unit nonresponse.

It is a consequence of unit nonresponse that the realized sample size is smaller than planned. This decreases the precision of estimates. However, if there are no other effects, valid estimates can still be obtained because computed confidence

intervals still have the proper confidence level. To avoid realized samples that are too small, the initial sample size can be taken larger. For example, if a sample of 1,000 elements is required, and the expected response rate is in the order of 60%, the initial sample size should be approximately $1,000/0.6=1,667$.

The main problem of nonresponse is that estimates of population characteristics may be biased. This situation occurs if, as a result of nonresponse, some groups in the population are over- or underrepresented in the sample, and these groups behave differently with respect to the characteristics to be investigated. Then nonresponse is said to be *selective*.

It is likely that survey estimates are biased unless very convincing evidence to the contrary is provided. Bethlehem (2009) gives examples of several Dutch surveys where nonresponse is selective. A follow-up study of the Dutch Victimization Survey showed that people, who are afraid to be home alone at night, are less inclined to participate in the survey. In the Dutch Housing Demand Survey, it turned out that people who refused to participate, have lesser housing demands than people who responded. And for the Survey of Mobility of the Dutch Population, it was obvious that the more mobile people were underrepresented among the respondents.

Nonresponse is a problem in almost every survey, whatever the mode of data collection. The magnitude and effect of nonresponse may differ from mode to mode; see Bethlehem, Cobben, and Schouten (2011). This section gives a short overview of the nonresponse problem and describes aspects that are specific to web surveys.

It will be shown that the amount of nonresponse is one of the factors determining the magnitude of the bias of estimates. The higher the nonresponse rate, the higher larger the bias will be.

EXAMPLE 4.6 Nonresponse in the Dutch housing demand survey

The effect of nonresponse is shown using data the Dutch Housing Demand Survey. Statistics Netherlands carried out this survey in 1981. The initial sample size was 82,849. The number of respondents was 58,972, which comes down to a response rate of 71.2%.

To obtain more insight into the nonresponse, a follow-up survey was carried out among the nonrespondents. One of the questions asked was whether they intended to move within two years. Table 4.4 shows the results.

TABLE 4.4 Nonresponse in the Dutch housing demand survey 1981

Do you intend to move within 2 years?	Response	Nonresponse	Total
Yes	17,515	3,056	20,571
No	41,457	20,821	62,278
Total	58,972	23,877	82,849

The percentage of people with the intention to move within two years is $100 \times 17,515/58,972 = 29.7\%$ if just the response data are used. This percentage is much lower for the complete sample (response and nonresponse): $100 \times 20,571/82,849 = 24.8\%$. The reason is clear: There is a substantial difference between respondents and nonrespondents with respect to the intention to move within two years. For nonrespondents, this percentage is only $100 \times 3,056/23,877 = 12.8\%$.

4.2.2.2 Causes of Nonresponse. Nonresponse can have many causes. It is important to distinguish these causes. To be able to reduce nonresponse in the field, one must know what caused it. Moreover, different types of nonresponse can have different effects on estimates and, therefore, may require different treatment. Main causes of nonresponse are noncontact, refusal, and not-able.

Nonresponse from *noncontact* occurs if it is impossible to get into contact with the respondent. This can happen in a face-to-face survey if the respondent is not at home. Noncontact occurs in a telephone survey if the telephone is not answered. Various forms of noncontact are possible in web surveys. It depends on the way in which sample persons are selected. If the sampling frame is a list with e-mail addresses, noncontact occurs if the e-mail with the invitation to participate in the survey does not reach a selected person. The e-mail address may be wrong, or the e-mail may be blocked by a spam filter. If the sampling frame is list of postal addresses and letters with an Internet address are sent to selected persons, noncontact may be caused by persons not receiving the letter. If recruitment for a web survey takes place by means of a face-to-face or telephone survey, noncontact can be from respondents being not at home or not answering the telephone.

Nonresponse from *refusal* can occur after contact has been established with a sample person. Refusal to cooperate can have many reasons: People may not be interested, they may consider it an intrusion of their privacy, they may have no time, and so on. Sometimes a refusal can be temporary. In this case, it may be attempted to make an appointment for another day and/or time. But often a refusal is permanent.

If sample persons for a web survey are contacted by an e-mail or a letter, they may postpone and forget to complete the questionnaire form. This can be considered a weak form of refusal. Sending a reminder helps to reduce this form of refusal.

Nonresponse from *not-able* is a type of nonresponse where respondents may be willing to respond but are not able to do so. Reasons for this type of nonresponse can be, for example, illness, hearing problems, or language problems. If a letter with an Internet address of a web questionnaire is sent to a sample person, this person receives the letter, and he/she wants to participate in the web survey but does not have access to the Internet, this can also be considered a form of nonresponse as a result of not-able.

It should be noted that lack of Internet access should sometimes be qualified as undercoverage instead of nonresponse. If the target population of a survey is

wider than just those with Internet and the sample is selected using the Internet, people without Internet have a zero selection probability. They will never be selected in the surveys. This is undercoverage. Nonresponse from not-able occurs if people have been selected in the survey but are not able to complete the questionnaire form (on the Internet).

4.2.2.3 Response Rate. Because the negative impact nonresponse may have on the quality of survey results, the *response rate* is considered to be an important indicator of the quality of a survey. Response rates are frequently used to compare the quality of surveys and to explore the quality of a survey that is repeated over time. Bethlehem (2009) defines the response rate as

$$(4.6) \quad \text{Response rate} = \frac{n_R}{n_E} = \frac{n_R}{n_R + n_{NR}},$$

where n_R is the number of (eligible) respondents, n_E is the total number eligible persons in the sample, and n_{NR} is the number of (eligible) nonrespondents. Eligible persons are persons that belong to the target population and have been selected in the sample. In practice it may difficult to compute n_E as it is not always possible to determine whether noncontacted persons are eligible.

Another complication concerns web and mail surveys. These are self-administered surveys. If there are no interviewers for recruiting respondents, it is not possible to establish eligibility or the cause of nonresponse. There are only two possibilities: The questionnaire form is completed or not. This may hinder analysis of the effects of nonresponse on the survey results.

EXAMPLE 4.7 Response rates in the LISS panel

The Longitudinal Internet Studies for the Social Sciences (LISS) panel is a web panel consisting of approximately of 5,000 households. This panel was set up in 2006 by CentERdata, a research institute in the Netherlands. The objective of the panel is to provide a laboratory for the development and testing of new, innovative research techniques.

The panel is based on a true probability sample of households drawn from the population register by Statistics Netherlands. The initial sample consisted of 10,150 households. Telephone numbers were added to the selected names and addresses. This was only possible for registered numbers. Households with a registered telephone were contacted by means of CATI. Addresses without a registered number and those who could not be contacted by telephone were visited by the interviewers (CAPI).

During the recruitment phase, persons in sampled households were first asked to participate in a short interview. General background questions about the respondent were asked. At the end of the interview,

respondents were told about the panel and asked if they would like to participate. Households without access to Internet, or who were worried that an Internet survey might be too complicated for them, were told about the simple-to-operate computer with Internet access that could be installed in their homes for free for the duration of the panel. To demonstrate the use of this computer, they were shown a demonstration video. The response process could be split into several steps:

1. Contact: Those who could be contacted for the recruitment interview.
2. Primary response: Those who participated in the recruitment interview.
3. Secondary response: Those who agreed to participate in the panel.
4. Tertiary response: Those who actually did participate in the panel.

The response rates are shown in Table 4.5 The column “Response rate” shows the response percentage in each separate step of the data collection process. The column “Cumulative” shows the cumulative effect of the current and previous steps.

TABLE 4.5 Response rates in the recruitment phase of the LISS panel

Step	Response rate	Cumulative
Contact	85.7%	85.7%
Primary response	80.9%	69.3%
Secondary response	74.1%	51.4%
Tertiary response	89.4%	45.9%

The column “Response rate” in Table 4.5 shows the response rate in each of the four steps. The secondary response is lowest. Only three out of four participants in the recruitment interview also agreed to participate in the panel. The two main reasons for this refusal were (1) that these persons considered the burden of participating in a panel too high, and (2) that they did not have Internet access and could not be persuaded to use it by offering a simple-to-use computer.

The tertiary response is highest: 9 out of 10 of those who agreed to participate in the panel also do so. Apparently, people kept their promise.

The cumulative column shows that ultimately only 45.9% of the sample persons became an active member of the LISS panel.

Note that the noncontact rate is high. In 14% of the cases, it was not possible (by CAPI or CATI) to establish contact with the sampled households.

More about the response to the LISS panel can be found in the study by Scherpenzeel (2009).

Like any other survey, web surveys suffer from nonresponse. A web survey is a self-administered survey. Therefore, web surveys have a potential for high nonresponse rates. An additional source of nonresponse includes technical problems that may be encountered by respondents having to interact with the Internet; see Couper (2000), Dillman and Bowker (2001), Fricker and Schonlau (2002), and Heerwegh and Loosveldt (2002). Slow modem speeds, unreliable connections, high connection costs, low-end browsers, and unclear navigation instructions may frustrate respondents. This often results in respondents breaking off completion of the questionnaire. To keep the survey response up to an acceptable level, every measure must be taken to avoid these problems. This requires a careful design of web survey questionnaire instruments.

4.2.2.4 The Effect of Nonresponse. Under the *random response model*, it is possible to investigate the possible impact of nonresponse on estimators of population characteristics. This model assumes every element k in the population to have an (unknown) response probability ρ_k . If element k is selected in the sample, a random mechanism is activated that results with probability ρ_k in response and with probability $1 - \rho_k$ in nonresponse. Under this model, a set of response indicators

$$(4.7) \quad R_1, R_2, \dots, R_N$$

is introduced, where $R_k = 1$ if the corresponding element k responds, and where $R_k = 0$ otherwise. So, $P(R_k = 1) = \rho_k$, and $P(R_k = 0) = 1 - \rho_k$.

Now suppose a simple random sample without replacement of size n is selected from this population. This sample is denoted by the set of indicators a_1, a_2, \dots, a_N , where $a_k = 1$ means that element k is selected in the sample, and otherwise $a_k = 0$. The response only consists of those elements k for which $a_k = 1$ and $R_k = 1$. Hence, the number of available cases is equal to

$$(4.8) \quad n_R = \sum_{k=1}^N a_k R_k.$$

Note that this realized sample size is a random variable. The number of nonrespondents is equal to

$$(4.9) \quad n_{NR} = \sum_{k=1}^N a_k (1 - R_k),$$

where $n = n_R + n_{NR}$.

The values of the target variable only become available for the n_R responding elements. The mean of these values is denoted by

$$(4.10) \quad \bar{y}_R = \frac{1}{n_R} \sum_{k=1}^N a_k R_k Y_k.$$

It can be shown, see Bethlehem (2009), that the expected value of the response mean is approximately equal to

$$(4.11) \quad E(\bar{y}_R) \approx \tilde{Y},$$

where

$$(4.12) \quad \tilde{Y} = \frac{1}{N} \sum_{k=1}^N \frac{\rho_k}{\bar{\rho}} Y_k$$

and

$$(4.13) \quad \bar{\rho} = \frac{1}{N} \sum_{k=1}^N \rho_k$$

is the mean of all response probabilities in the population. From (4.11), it is clear that, generally, the expected value of the response mean is unequal to the population mean to be estimated. Therefore, this estimator is biased. This bias is approximately equal to

$$(4.14) \quad B(\bar{y}_R) = \tilde{Y} - \bar{Y} = \frac{S_{\rho Y}}{\bar{\rho}} = \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}},$$

where $S_{\rho Y}$ is the covariance between the values of the target variable and the response probabilities, $R_{\rho Y}$ is the corresponding correlation coefficient, S_Y is the standard deviation of the variable Y , and S_{ρ} is the standard deviation of the response probabilities. From this expression of the bias, several conclusions can be drawn:

- The bias vanishes if there is no relationship between the target variable and the response behavior. This implies $R_{\rho Y} = 0$. The stronger the relationship between the target variable and the response behavior, the larger the bias will be.
- The bias vanishes if all response probabilities are equal. Then $S_{\rho} = 0$. Indeed, in this situation, the nonresponse is not selective. It just reduces the sample size.
- The magnitude of the bias increases as the mean of the response probabilities decreases. Translated in practical terms, this means that lower response rates will lead to larger biases.

Currently, the response rates for general-population surveys that are carried out as web surveys have a lower response rate than comparable CAPI or CATI surveys. Beukenhorst and Giesen (2010) report the response rates for some web surveys of Statistics Netherlands: 21% for the Safety Monitor, 26% for the Mobility Survey, and 35% for the Health Interview Survey. Holmberg (2010) describes an experiment where respondents could choose between mail and web.

Of the sample persons, only 11.8% selected the web questionnaire and 58.1% the mail questionnaire. The nonresponse was 30.1%. Those selecting the mail questionnaire did this because it was immediately available. They did not have to go to their computer.

4.2.2.5 Analysis and Correction of Nonresponse. It is important to carry out a nonresponse analysis on the data that have been collected in a survey. Such an analysis should make clear whether response is selective and, if so, which technique should be applied to correct for a possible bias. Unfortunately, the available data with respect to the target variables will not be of much use. There are only data for the respondents and not for the nonrespondents. So it is not possible to establish whether respondents and nonrespondents differ with respect to these variables. The way out of this problem is to use auxiliary variables; see Figure 4.19.

An auxiliary variable is in this context a variable that has been measured in the survey, and for which the distribution in the population (or in the complete sample) is available. So it is possible to establish whether there is a relationship between this variable and the response behavior.

Three different response mechanisms are distinguished. The first one is *Missing Completely at Random* (MCAR). The occurrence of nonresponse (R) is completely independent of both the target variable (Y) and the auxiliary variable (X). The response is not selective. Estimates are not biased. There is no problem.

In the case of MCAR, the response behavior (R) and any auxiliary variable (X) are unrelated. If it is also known that there is a strong relationship between the target variable (Y) and the auxiliary variable (X), this is an indication there is no strong relation between target variable (Y) and response behavior (R), and thus, estimators do not have a severe bias.

It should be noted that if there is no strong relationship between the auxiliary variable (X) and the target variable (Y), analysis of the relationship between the auxiliary variable (X) and the response behavior will provide no information about a possible bias of estimates.

The second response mechanism is *Missing at Random* (MAR). This situation occurs when there is no direct relation between the target variable (Y) and

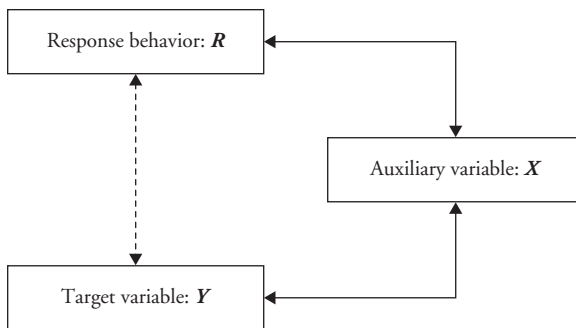


FIGURE 4.19 Relationship among target variable, response behavior, and auxiliary variable

the response behavior (R), but there is a relation between the auxiliary variable (X) and the response behavior (R). The response will be selective, but this can be cured by applying a weighting technique using the auxiliary variable. Chapter 10 is devoted to such weighting techniques.

In the case of MAR, the response behavior (R) and the corresponding auxiliary variable (X) will turn out to be related. If it is also known that there is a strong relationship between the target variable (Y) and the auxiliary variable (X), this is an indication there is a (indirect) relation between target variable (Y) and response behavior (R), and thus, the estimators may be biased.

The third response mechanism is *Not Missing at Random* (NMAR). There is a direct relationship between the target variable (Y) and the response behavior (R), and this relationship cannot be accounted for by an auxiliary variable. The estimators are biased. Correction techniques based on use of auxiliary variables will be able to reduce such a bias.

All this indicates that the relationship between auxiliary variables and response behavior should be analyzed. If such a relationship exists, and it is known there is also a relationship between the target variables and auxiliary variables, there is a serious risk of biased estimates. So application of nonresponse correction techniques should be considered.

EXAMPLE 4.8 Nonresponse in the LISS panel

The LISS panel is a web panel consisting of approximately of 5,000 households. More details can be found in Example 4.7.

The panel is based on a true probability sample of households drawn from the population register by Statistics Netherlands. The initial sample consisted of approximately 10,000 households. The response process could be split into several steps:

1. Contact: Those who could be contacted for the recruitment interview.
2. Primary response: Those who participated in the recruitment interview.
3. Secondary response: Those who agreed to participate in the panel.
4. Tertiary response: Those who actually did participate in the panel.

Because the sample was selected from the population register of the Netherlands, the age distribution of persons in the response can be compared with the age distribution in the complete sample. The mosaic plot in Figure 4.20 does this in a graphical way.

The vertical bars represent the age categories. The width of these bars is proportional to the number of persons in the corresponding age groups. The shades of gray represent the steps in the response process: NC = noncontact, NR1 = primary nonresponse, NR2 = secondary nonresponse, NR3 = tertiary nonresponse, and R = ultimate response.

The graph shows a clear relationship between age and response behavior. What is striking is the low response from people of age 70 years and older. This is mainly a tertiary nonresponse. These elderly people participate in the recruitment survey but do not want to become a member of the panel. The response is also low for young people (18–30 years). The main problem here is noncontact.

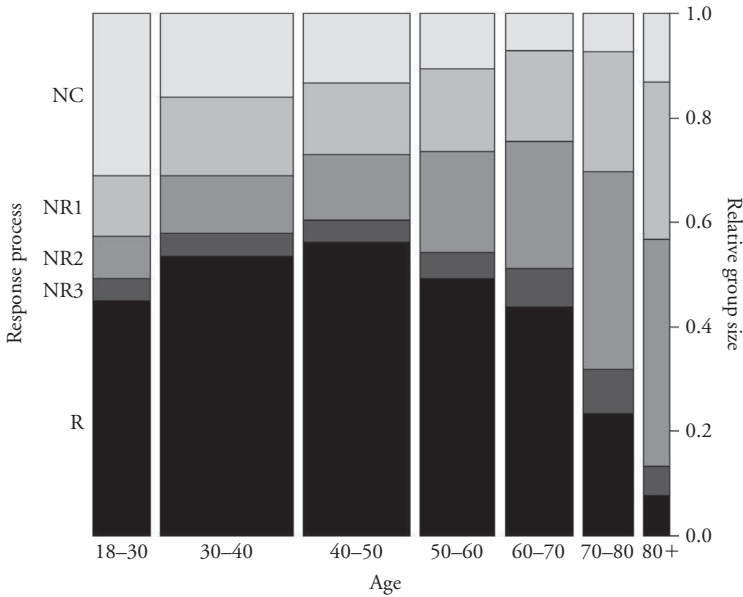


FIGURE 4.20 Nonresponse in the recruitment phase of the LISS panel

There is ample evidence that nonresponse often causes population estimates to be biased. This means that something has to be done to prevent wrong conclusions from being drawn from the survey data. Several correction approaches are possible. The most frequently used one is *adjustment weighting*. It assigns weights to observed elements. These weights are computed in such a way that overrepresented groups get a smaller weight than underrepresented groups. Adjustment weighting has many aspects. Chapter 10 is completely dedicated to this approach.

4.3 Application

4.3.1 THE SAFETY MONITOR

Many national statistical institutes face a challenge in reducing data collection costs. This has led to considering new ways of data collecting, in which web

surveys play an important role. The most far-reaching change would be to replace traditional, expensive, interviewer-assisted CAPI and CATI surveys by self-administered, and therefore cheaper, web surveys. Another option is to introduce mixed-mode surveys.

Statistics Netherlands has conducted experiments to determine whether a mixed-mode survey can replace a CAPI or CATI survey without affecting the quality of the results. One such experiment with the Dutch Safety Monitor is described in this section. A more detailed account of these experiments is given by Beukenhorst and Wetzels (2009) and Kraan et al. (2010).

The Dutch Safety Monitor is an annual survey of Statistics Netherlands. It measures the actual and perceived safety of the people in the country. Respondents are asked questions about feelings of safety, quality of life, and level of crime experienced. The sample for this survey is selected from the Dutch population register. Only persons age 15 years and older are selected.

Until 2008 the sample size was approximately 20,000. Local authorities also collected data on the same topics at the regional level. These surveys were conducted parallel to the Safety Monitor. This resulted in inconsistent estimates for safety feelings and crime victimization. Therefore, it was decided to integrate the national and the regional surveys into a new Integrated Safety Monitor. This new survey had a sequential mixed-mode design with the web as one mode. To assess the effects of the change in survey design, the old Safety Monitor was carried out in parallel with the new Integrated Safety Monitor in 2008.

The old Safety Monitor applied two modes of data collection. If sampled persons had a known telephone number, they were approached by CATI. If this was not the case, they were approached by CAPI. The sample size was in 2008 approximately equal to 6,000 persons. This survey will be denoted by SM2.

The new Integrated Safety Monitor had four modes of data collection. All sample persons received a letter in which they were asked to complete the survey questionnaire on the Internet. The letter also included a postcard that could be used to request a paper questionnaire. Two reminders were sent to those that did not respond by web or mail. If still no response was obtained, nonrespondents were approached by means of CATI if a listed telephone number was available. If not, these nonrespondents were approached by CAPI. This four-mode survey will be denoted by SM4.

4.3.2 MEASUREMENT ERRORS

The objective of the Safety Monitor is to measure several indicators. Each indicator is based on numerous underlying questions. All indicators assume a value on a scale from 0 to 10. Three indicators are considered here:

- Harassment in the neighborhood (0 = No harassment, 10 = Harassment occurs frequently)
- Police performance (0 = Negative, 10 = Positive)
- Degradation of the neighborhood (0 = No degradation, 10 = Degradation occurs frequently)

TABLE 4.6 Average value of indicators in SM4 and SM2

Indicator	SM4	SM2	Difference
Harassment in neighborhood	1.65	1.34	+ 0.31
Police performance	5.50	5.88	-0.38
Degradation of neighborhood	3.64	2.97	+ 0.67

TABLE 4.7 Response distribution of “perceived graffiti” by mode in SM4

Answer	CAPI	CATI	Mail	Web
Occurs frequently	12%	13%	12%	11%
Occurs sometimes	23%	23%	31%	33%
Occurs (almost) never	61%	65%	45%	48%
Refuses to answer	0%	0%	1%	0%
Don’t know	4%	1%	12%	7%
Total	100%	102%	101%	99%

Table 4.6 contains the average scores for these three indicators in both surveys. All differences are significant. Apparently, respondents in SM4 have a more negative attitude toward these aspects than respondents in SM2. This conclusion gives rise to questions as to what the cause of these differences is.

The next step in the analysis is establishing whether the differences can be attributed to mode effects. As an example, Table 4.7 contains the response distribution by mode of a question in SM4 that asks about the perceived occurrence of graffiti in the neighborhood.

There is a striking difference for the interviewer-assisted modes on the one hand and the self-administered modes on the other. For the mail and web modes, there is a substantial shift of respondents from to “Occurs (almost) never” to “Occurs sometimes”. There are several possible explanations for this shift:

- It is case of nondifferentiation. Respondents in the self-administered modes simply choose the middle category.
- There is a recency effect in the interviewer-assisted modes. This would result in respondents selecting more the last answer option read (“Occurs (almost) never”).

Figure 4.21 shows the analysis for an attitude question. It asks respondents for their judgment of the statement “I feel at home with the people here in the neighborhood.”

There are two striking phenomena in this graph. The first one is that much more people agree with this statement in the CATI and CAPI mode. This may be caused by acquiescence. It is the tendency to agree with the statements made by the interviewer. Apparently people avoid contradicting the interviewer. CATI

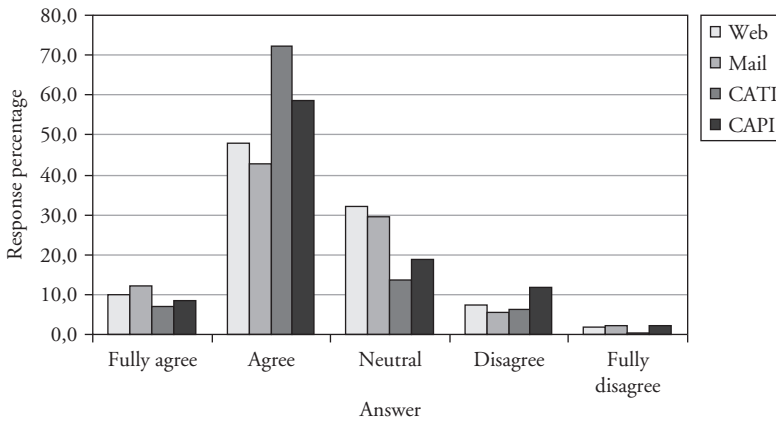


FIGURE 4.21 Response in SM4 to the question “I feel at home with the people here in the neighborhood”

and CAPI are both interviewer-assisted modes of data collection. Therefore, one can expect that acquiescence occurs only for these modes, and not for the web and mail modes.

Another striking phenomenon is that much less people choose the neutral category in CAPI and CATI. This may be caused by the tendency of people to give more socially desirable answers when interviewers are present. Not having a clear opinion is not considered socially desirable. Therefore, they may express an opinion they may not have.

It should be noted that the differences between SM4 and SM2 may be caused by other effects than mode effects. There also can be *selection effects*. This is the phenomenon that different subpopulations respond in different modes. For example, SM4 has a web mode that was not available in SM2. Because Internet coverage is high among young people, this may cause more young people to be in the SM4 survey. Therefore, it is important to check the composition of the response in the different modes. Kraan et al. (2010) did this for the variable age. It turned out there were no large differences in the age distributions by mode.

4.3.3 NONRESPONSE

The response rate of SM4 turned out to be 59.7%. The response rate for SM2 was 63.5%. So there is not much difference. Table 4.8 shows the composition of the response for both surveys.

More than half of the response (58%) was obtained in the SM4 with a self-administered mode of data collection (web or mail). The conclusion can be drawn that the four-mode survey did not increase the response. The costs of the survey were, however, much lower because interviewers were deployed in only 42% of the cases. Focusing on just interviewer costs, and ignoring all other costs (which are much lower), Beukenhorst and Wetzels (2009) found that the costs of SM4 were only 60% of the costs of SM2.

TABLE 4.8 Composition of the response of both Safety Monitors

Data collection mode	SM4	SM2
Web	41.8%	
Mail	16.2%	
CATI	30.5%	71.6%
CAPI	11.5%	28.4%
Total	100.0%	100.0%

TABLE 4.9 Ranges of the response rates (%) in the categories of auxiliary variables

Auxiliary variable	Categories	Range in SM4	Range in SM2
Household size	2	52%–64%	54%–67%
Gender	2	60%–62%	63%–65%
Age	7	55%–66%	52%–70%
Ethnicity	4	41%–64%	44%–67%
Degree of urbanization	5	52%–66%	53%–71%
Average		60%	64%

The quality of the response is not only determined by the response rate. The composition of the response is more important. If the response composition differs substantially from the sample composition, the estimates may be biased. To get insight into possible composition differences, the response rates in the categories of auxiliary variables can be compared. Table 4.9 shows the result.

The table shows there is little difference in the response rates. Only for *Age* and *Degree of urbanization* the range seems to be smaller for SM4. This is an indication that the composition may have improved.

A phenomenon that repeats itself in many general-population surveys is the low response rate of specific groups. Well-known examples are people in highly urbanized areas and ethnic minority groups. Beukenhorst and Wetzels (2009) investigated whether the four-mode survey reduces this problem. This turned out not to be the case for the urbanized areas because response was also very low here for the web and mail modes. They drew the same conclusion for ethnic groups.

The results of this experiment show that, in this case, the mixed-mode approach (with four modes) was not very successful with respect to the quality of the outcomes. The response rate did not increase substantially, and the composition of the response did not improve much. The hope that introducing new modes would increase the response rate of low-response groups was not fulfilled.

The experiment showed that a mixed-mode survey can be very successful with respect to survey costs. The four-mode survey was substantially less expensive than the two-mode survey.

4.4 Summary

Survey researchers have control over many different aspects of the survey process. With the proper choice of a sampling frame, a sampling design, and an estimation procedure, they can obtain the precise estimators of population characteristics. Unfortunately not everything is under control. Survey researchers may be confronted with various phenomena that may have a negative impact on the quality, and therefore on the reliability, of the survey outcomes. Some of these disturbances are almost impossible to prevent. After giving an overview of what can go wrong, this chapter concentrates on two types of problems: measurement errors and nonresponse errors.

Measurement errors occur when the answers given by respondents differ from the true answers. The reason can be that the respondent does not understand a question, does not know the true answer, or does not want to give the true answer. The nature and magnitude of measurement errors may be different for different modes of data collection. This chapter focuses on measurement errors in web surveys. An important aspect is the absence of interviewers in web surveys. This may have positive effects (for example, less socially desirable answers) but also negative effects (for example, more satisficing). This chapter stresses the importance of paying careful attention to the design of the web survey questionnaire because small errors in the design may lead to large errors in the answers.

Another important issue is nonresponse. Particularly high nonresponse rates may cause large biases in estimates. It is important to distinguish among the various causes of nonresponse (noncontact, refusal, not-able) because they may have a different impact on estimates. The cause of nonresponse may be difficult to establish for web surveys.

It should always be attempted to correct for a possible nonresponse bias. Auxiliary variables are required for this. These variables must have been measured in the survey, and moreover, their distribution in the population or complete sample must be available.

KEY TERMS

Acquiescence: The tendency that respondents tend to agree with statements in questions, regardless of their content. They simply answer “yes.”

Data editing: The activity of checking collected data for errors and, where possible, of correcting these errors.

Eligible: A sample element is eligible for a survey if it belongs to the target population of the survey.

Estimation error: The deviation of the estimate from the true value caused by investigating only a sample instead of the complete population.

Grid question: See the definition of “matrix question.”

Item nonresponse: A type of nonresponse occurring when some questions have been answered but no answer is obtained for some other, possibly sensitive, questions. So, the questionnaire form has been partially completed.

Missing at Random (MAR): Nonresponse depends on auxiliary variables only. Estimators will be biased, but a correction is possible if some technique is used that takes advantage of this auxiliary information.

Missing Completely at Random (MCAR): Nonresponse happens completely independent of all survey variables. The estimators will not be biased.

Matrix question: A series of questions with the same set of answer options combined in a matrix. Each row of a matrix question represents a single question, and each column corresponds to an answer option.

Measurement error: An error occurring if the answer given by a respondent differs from the true answer. The reason can be that the respondent does not understand a question, does not know the true answer, or does not want to give the true answer.

Memory error: The error caused by respondents who forget to report events, particularly when they happened a long time ago or when they are not too salient.

Nondifferentiation: A form of satisficing that typically occurs if respondents have to answer a series of questions each with the same set of response options. They tend to select the same answer for all these questions irrespective of the question content.

Nonobservation error: A type of nonsampling error that occurs when intended measurements cannot be carried out. Two types of nonobservation errors can be distinguished: undercoverage errors and nonresponse errors.

Nonresponse: The phenomenon that elements in the selected sample, which are also eligible for the survey, do not provide the requested information or that the provided information is not usable.

Not Missing at Random (NMAR): Nonresponse depends directly on the target variables of the survey. The estimators will not be biased, and the correction techniques will not be successful.

Nonsampling error: The error caused by problems that even can occur if the whole population is investigated (instead of a sample). Nonsampling errors can be divided into observation errors and nonobservation errors.

Observation error: An error made during the process of obtaining answers from respondents, and recording and further processing these answers. Three types of observation errors are distinguished here: overcoverage errors, measurement errors, and processing errors.

Overcoverage error: An error caused by elements that are included in the survey that do not belong to the target population.

Primacy effect: The tendency of respondents to pick an answer early in the list of answer options of a closed question. Primacy effects typically occur in interviewer-assisted surveys.

Processing error: An error made during the phase of recording and processing the collected data.

Random response model: A model for nonresponse that assumes every element in the population to have an (unknown) response probability.

Recency effect: The tendency of respondents to pick an answer at or near the end of the list of answer options of a closed question. Recency effects typically occur in self-administered surveys.

Response order effect: The tendency that the answer selected by the respondent depends on its location in the list of answer options. Primacy and recency effects are special cases.

Response rate: The number of responding eligible elements in the sample divided by the number of eligible elements in the sample.

Sampling error: The error in the estimate introduced by the sampling design. It is caused by the fact that estimates are based on a sample from the population and not on a complete enumeration of the population.

Satisficing: The phenomenon that respondents do not do all they can to provide a correct answer. Instead they attempt to give a satisfactory answer with minimal effort.

Selection effect: The phenomenon that different subpopulations favor different data collection modes of a mixed-mode survey.

Socially desirable answer: The tendency that respondents give answers that will be viewed as more favorable by others. This particularly happens for sensitive questions.

Specification error: An error occurring when the selection probabilities used for the computation of an estimate differ from the true selection probabilities.

Straightlining: The tendency that respondents give the same answer to all single questions in a matrix question. They simply check all answer options in the same column.

Total error: The combined effect of all phenomena that contribute to an estimated value that deviates from the true value of a population characteristic.

Undercoverage: The sampling frame does not cover completely the target population of the survey. There are persons in the population who do not appear in the sampling frame. They will never be selected in the sample.

Unit nonresponse: This type of nonresponse occurs when a selected element does not provide any information at all; i.e., the questionnaire form remains completely empty.

EXERCISES

Exercise 4.1. Which of the following sources of error does not belong to the category of observation errors?

- a. Measurement error
- b. Overcoverage
- c. Undercoverage
- d. Processing error

Exercise 4.2. Memory effects occur if respondents forget to report certain events or when they make errors about the date of occurrence of events. To which source of errors do these memory effects belong?

- a. Estimation error
- b. Undercoverage
- c. Measurement error
- d. Nonobservation error

Exercise 4.3. Which of the following effects may occur in a web survey?

- a. Both primacy and recency effects
- b. Only primacy effects
- c. Only recency effects
- d. Nonobservation error

Exercise 4.4. Which of the following phenomena is not a form of satisficing?

- a. Response order effects
- b. Acquiescence
- c. Socially desirable answers
- d. Nondifferentiation

Exercise 4.5. What is the best way to format a closed question?

- a. With a set of check boxes
- b. With a drop-down list
- c. With a set of radio buttons
- d. With a text input field

Exercise 4.6. In what situation are progress indicators effective?

- a. In the case of complex questionnaire with a lot of routing
- b. In the case of long questionnaires without routing
- c. In the case of short questionnaires without routing
- d. In the case of questionnaires with a lot of checks

Exercise 4.7. Which phenomenon makes it difficult, if not impossible, to compute the response rate of a web survey?

- a. Overcoverage
- b. Undercoverage
- c. Noncontact
- d. Refusal

Exercise 4.8. A survey is usually carried out to measure the state of a target population at a specific point in time (the reference date). The survey outcomes are supposed to describe the population at this date. Ideally, the fieldwork of the survey should take place at that date. This is not possible in practice. Interviewing usually takes place in a period of several days or weeks around the reference date.

Suppose a web survey is carried among employees of a company. Each employee has a company e-mail address. So there is a sampling frame. A sample of employees is selected from this sampling frame two weeks before the reference date. Selected employees are asked to complete the questionnaire in a period of four weeks: the two weeks between sample selection and reference date and the two weeks after the reference date.

Explain for each situation described below what kind of problem there is: nonresponse, undercoverage, overcoverage, or another sampling frame error:

- a. A selected employee died before the sample selection date.
- b. A selected employee died between the sample selection date and the reference date.
- c. A selected employee died between the reference date and the sample selection date.

Exercise 4.9. A town council wants to do something about the traffic problems in the town center. There is a plan to turn it into a pedestrian area, so cars cannot go into the center any more. The town council wants to know what companies think of this plan. A simple random sample of 1,000 companies is selected. Each selected company is invited to participate in a web survey. The company is asked whether it is in favor of the plan. Furthermore, the location of the company is recorded (town center or suburb). The results of the survey are summarized in the following table:

	Suburbs	Town center
In favor	120	80
Not in favor	40	240

- a. Determine the response percentage.
- b. Determine the percentage respondents in favor of the plan.
- c. Compute a lower bound and an upper bound of the percentage of companies in the complete sample in favor of the plan.

Exercise 4.10. The local authorities of a town want to know how satisfied citizens are with public transport facilities in town. They conduct a web survey. The target population is defined as all citizens that used public transport at least once in the last year. A sample is selected from the population register of the

town. Selected persons are sent a letter with the Internet address of the survey. The results of the survey are summarized in the following table:

Result	Frequency
Overcoverage	320
Refusal	240
Noncontact	80
Not-able	40
Response	440
Total	1,120

Compute the response rate of the survey. Make it clear how the response rate was computed and which assumptions were made.

REFERENCES

- Bethlehem, J. G. (2009), *Applied Survey Methods, a Statistical Perspective*. John Wiley & Sons, Hoboken, NJ.
- Bethlehem, J. G., Cobben, F., & Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*. Wiley Handbooks in Survey Methodology, Wiley, Hoboken NJ.
- Beukenhorst, D. & Giesen, D. (2010), *Internet Use for Data Collection at Statistics Netherlands*. Paper presented at the 2nd International Workshop on Internet Survey Methods, Daejeon, South Korea.
- Beukenhorst, D. & Wetzels, W. (2009), *A Comparison of Two Mixed-mode Designs of the Dutch Safety Monitor: Mode Effects, Costs, Logistics*. Technical Paper DMH 206546, Statistics Netherlands, Methodology Department, Heerlen, the Netherlands.
- Bishop, G. F. (1987), Experiments with the Middle Response Alternative in Survey Questions. *Public Opinion Quarterly*, 51, pp. 220–232.
- Bradburn, N. M., Sudman, S., & Wansink, B. (2004), *Asking Questions, The Definitive Guide to Questionnaire Design—For Market Research, Political Polls, and Social and Health Questionnaires*. Jossey-Bass, San Francisco.
- Christian, L. M., Dillman, D., & Smyth, J. (2007), Helping Respondents Get It Right the First Time: The Influence of Words, Symbols and Graphics in Web Surveys. *Public Opinion Quarterly*, 71, pp. 113–125.
- Converse, P. E. (1964), The Nature of Belief Systems in Mass Public. In: Apter, D. (ed.), *Ideology and Discontent*. Free Press, New York, pp. 206–261.
- Couper, M. P. (1999), Usability Evaluation of Computer Assisted Survey Instruments. *Proceedings of the Third ASC International Conference*. ASC, Edinburgh, U.K. pp. 1–14.
- Couper, M. P. (2000), Web Surveys, A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, pp. 464–494.
- Couper, M. P. (2008), *Designing Effective Web Surveys*. Cambridge University Press, New York.

- Couper, M. P., Tourangeau, R., Conrad, F. G., & Crawford, S. D. (2004), What They See Is What They Get: Response Options for Web Surveys. *Social Science Computer Review*, 22, pp. 111–127.
- Couper, M. P., Traugott, M., & Lamias, M. (2001), Web Survey Design and Administration. *Public Opinion Quarterly*, 65, pp. 230–253.
- De Leeuw, E. D. (1992), *Data Quality in Mail, Telephone, and Face-to-Face Surveys*. TT-Publications, Amsterdam, the Netherlands.
- De Leeuw, E. D. (2008), Choosing the Mode of Data Collection. In: De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (eds.), *International Handbook of Survey Methodology*. Lawrence Erlbaum Associates, New York/London, pp. 113–135.
- DeRouvray, C. & Couper, M. P. (2002), Designing a Strategy for Reducing “No Opinion” Responses in Web-Based Surveys. *Social Science Computer Review*, 20, pp. 3–9.
- Dillman, D. A. (2007), *Mail and Internet Surveys. The Tailored Design Method*. John Wiley & Sons, New York.
- Dillman, D. A. & Bowker, D. (2001). The Web Questionnaire Challenge to Survey Methodologists. In: Reips, U. D. & Bosnjak, M. (eds.), *Dimensions of Internet Science*. Pabst Science Publishers, Lengerich, Germany, pp. 159–178.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009), *Internet, Mail, and Mixed-mode Surveys, The Tailored Design Method*. John Wiley & Sons, New York.
- Dillman, D. A., Tortora, R. D., & Bowker, D. (1998), *Principles for Constructing Web Surveys*. SERC Technical Report 98–50, Pullman, Washington, DC.
- Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005), An Experimental Comparison of web and telephone surveys. *Public Opinion Quarterly*, 69, pp. 370–392.
- Fricker, R. & Schonlau, M. (2002), Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Field Method*, 15, pp. 347–367.
- Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007), *Using Change to Improve Navigation in Grid Questions*. Paper presented at the General Online Research Conference (GOR '07), Leipzig, Germany.
- Heerwegh, D. (2004), *Using Progress Indicators in Web Surveys*. Paper presented at the 59th AAPOR Conference, Phoenix, Arizona.
- Heerwegh, D. & Loosveldt, G. (2002), An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys. Paper presented at the International Conference on Improving Surveys, Copenhagen, Denmark.
- Heerwegh, D. & Loosveldt, G. (2008), Face-to-face Versus Web Surveying in a High-internet Coverage Population. Differences in Response Quality. *Public Opinion Quarterly*, 72, pp. 836–846.
- Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003), Telephone Versus Face-to-face Interviewing of National Probability Samples with Long Questionnaires. *Public Opinion Quarterly*, 67, pp. 79–125.
- Holbrook, A. L. & Krosnick, J. A. (2010), Social Desirability Bias in Voter Turnout Reports. Tests Using the Item Count Technique. *Public Opinion Quarterly*, 74, pp. 37–67.
- Holmberg, A. (2010), *Using the Internet in Individual and Household Surveys, Summarizing Some Experiences at Statistics Sweden*. Paper presented at the 2nd International Workshop on Internet Survey Methods, Daejeon, South Korea.

- Kaczmirek, L. (2010), Attention and Usability in Internet Surveys: Effects of Visual Feedback in Grid Questions. In: Das, M., Ester, P., & Kaczmirek, L. (eds.), *Social and Behavioral Research and the Internet, Advances in Applied Methods and Research Strategies*. Routledge, New York/London, pp. 191–214.
- Kaczmirek, L., Neubarth, W., Bosnjak, M., & Bandilla, W. (2004), Progress Indicators in Filter-based Surveys: Computing Methods and Their Impact on Drop Out. Paper presented at the RC33, the 6th International Conference on Social Science Methodology, Amsterdam, the Netherlands.
- Kalton, G., Roberts, J., & Holt, D. (1980), The Effect of Offering a Middle Response Option with Opinion Questions. *Journal of the Royal Statistical Society, Series D*, 29, pp. 65–78.
- Kalton, G. & Schuman, H. (1982), The Effect of the Question on Survey Responses: A Review. *Journal of the Royal Statistical Society*, 145, pp. 42–57.
- Kish, L. (1967), *Survey Sampling*. John Wiley & Sons, New York.
- Kraan, T., Van den Brakel, Buelens, B., & Huys, H. (2010), *Social Desirability Bias, Response Order Effect and Selection Effects in the New Dutch Safety Monitor*. Discussion Paper 10004, Statistics Netherlands, The Hague/Heerlen, the Netherlands.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008), Social Desirability Bias in CATI, IVR and Web Surveys. *Public Opinion Quarterly*, 72, pp. 847–865.
- Krosnick, J. A. (1991), Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5, pp. 213–236.
- Krosnick, J. A. (1999), Survey Research. *Annual Review of Psychology*, 50, pp. 537–567.
- Krosnick, J. A. & Alwin, D. F. (1987), An Evaluation of Cognitive Theory of Response-order Effects in Survey Measurement. *Public Opinion Quarterly*, 51, pp. 201–219.
- Lodge, M., Steenbergen, M. R., & Brau, S. (1995). The Responsive Voter: Campaign Information and the Dynamics of Candidate Evaluation. *American Political Science Review*, 89, pp. 309–326.
- Saris, W. E. (1997), The Public Opinion About the EU Can Easily Be Swayed in Different Directions. *Acta Politica*, 32, pp. 406–436.
- Scherpenzeel, A. (2009), *Start of the LISS Panel: Sample and Recruitment of a Probability-based Internet Panel*. CentERdata, Tilburg, the Netherlands.
- Schuman, H. & Presser, S. (1981), *Questions and Answers in Attitude Surveys*. Academic Press, New York.
- Schwarz, N., Hippler, H. J., & Noelle-Neumann, E. (1992), A Cognitive Model of Response Order Effects in Survey Measurement. In: Schwarz, N. & Sudman, S. (eds.), *Context Effects in Social and Psychological Research*, Springer-Verlag, New York, pp. 187–201.
- Schwarz, N., Knäuper, B., Oyserman, D., & Stich, C. (2008), The Psychology of Asking Questions. In: De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (eds.), *International Handbook of Survey Methodology*. Lawrence Erlbaum Associates. New York/London, pp. 18–34.
- Sikkel, D. (1983), Geheugeneffecten bij het Rapporteren van Huisartsencontacten. *Statistisch Magazine* 3, nr. 4. Netherlands Central Bureau of Statistics, pp. 61–64.
- Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006), Comparing Check-all and Forced-choice Formats in Web Surveys. *Public Opinion Quarterly*, 70, pp. 66–77.

- Statistics Netherlands. (2002), *Blaise Developer's Guide*. Statistics Netherlands, Heerlen, the Netherlands.
- Sudman, S. & Bradburn, N. M. (1982), *Asking Questions: A Practical Guide to Questionnaire Design*. Jossey-Bass, San Francisco.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996), *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco.
- Tiemeijer, W. L. (2008), *Wat 93,7 Procent van de Nederlanders Moet Weten over Opiniepeilingen*. Aksant, Amsterdam, the Netherlands.
- Tourangeau, R., Couper, M. P., & Conrad, F. (2004), Spacing, Position and Order, Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, 68, pp. 368–393.
- Vis-Visschers, R., Arends-Tóth, J., Giesen, D., & Meertens, V. (2008), *Het Aanbieden van 'Weet niet' en Toelichtingen in een Vragenlijst*. Report DMH-2008-02-21-RVCS, Statistics Netherlands, Heerlen, the Netherlands.
- Zaller, J. R. (1992), *The Nature and Origins of Mass Opinion*. Cambridge University Press, Cambridge, U. K.

Web Surveys and Other Modes of Data Collection

5.1 Introduction

5.1.1 MODES OF DATA COLLECTION

Survey data collection has evolved over the years. The period before the 1970s was the era of traditional interviewing using paper forms. Basically, there were three ways to do this: face-to-face interviewing, telephone interviewing, and mail interviewing.

Face-to-face interviewing is a mode of data collection that was already used for the first censuses. So, it is not surprising that it was also used for the first surveys. Face-to-face interviewing means that interviewers visit the persons selected in the sample. They ask the questions and record the answers on the questionnaire form.

Telephone interviewing is a mode of data collection where interviewers call selected persons by telephone. If contact is made with the proper person, and this person agrees to participate, the interview is started and conducted over the telephone. The interviewers ask the questions and record the answers on the questionnaire form.

Mail interviewing is a mode of data collection that requires no interviewers. Respondents complete the questionnaire themselves. They read the questions and write down the answers. The completed questionnaire is returned to the survey organization.

The rapid developments of information technology since the 1970s have made it possible to use microcomputers for data collection. Thus, *computer-assisted interviewing* (CAI) emerged. The paper questionnaire was replaced by a computer program containing the questions to be asked. The computer took control of the interviewing process, and it also checked answers to questions on the spot. Computer-assisted interviewing has three important advantages over traditional forms of interviewing:

- It relieves the interviewers of the task of choosing the correct route through the questionnaire. This task is taken care of by the interview software. Therefore, interviewers can concentrate on asking questions and on assisting respondents in getting the right answers.
- It can improve the quality of the collected data. Answers can be checked and corrected during the interview. This is more effective than having to do it afterward in the survey agency.
- Data are entered in the computer already during the interview. This results in a “clean” record. No more subsequent data entry and data editing is necessary. This considerably reduces the time needed to process the survey data, and thus it improves the timeliness of the survey results.

Computer-assisted interviewing has different modes of data collection. They are the electronic analogues of the traditional modes of data collection.

Computer-assisted personal interviewing (CAPI) is a form of face-to-face interviewing where interviewers take their laptop computers to the homes of the respondents. There they start the interview program and enter the answers to the questions appearing on the screen. Interviewers send the collected data to the survey organization by means of the Internet. In return they may receive new names and addresses of persons to interview.

Computer-assisted telephone interviewing (CATI) is the electronic form of telephone interviewing. Interviewers call respondents from the call center of the survey organization. The interview software decides when and who to contact. If contact is established by telephone, and the person agrees to participate in the survey, the interviewer starts the interview program. The first question appears on the screen. If this is answered, and no error is detected, the software proceeds to the next question on the route through the questionnaire.

Computer-assisted self-interviewing (CASI), or sometimes also *computer-assisted self-administered questionnaires* (CSAQ), is the electronic analogue of mail interviewing. The electronic questionnaire is sent to the respondents. They run the software on their own computer, answer the questions, and send the answers back to the survey agency. Nowadays it is common practice to download the software from the Internet. The answers are returned electronically in the same fashion.

The rapid emergence and diffusion of the Internet in the 1990s has led to a new mode of data collection: the *web survey*, sometimes also called *computer-assisted web interviewing* (CAWI). The respondents are invited to go to a specific

page on the Internet. There they complete the questionnaire themselves. No interviewers are involved. In fact, a web survey is a special type of CASI survey.

5.1.2 THE CHOICE OF THE MODES OF DATA COLLECTION

The choice of the mode of data collection is not always an easy one. There are two important, but often conflicting, factors: costs and quality.

Conducting a survey can be expensive and time-consuming, particularly if the sample size is large and the questionnaire is long and complex. Many people may be involved in setting up and carrying out a survey. Depending on the survey, there may be researchers, questionnaire designers, interviewers, supervisors, data entry typists, analysts, and so on. Staff costs may well be the largest component of the total survey costs. There may be other costs, like the costs of hardware and software. Equipping a large group of interviewers with laptops for a CAPI survey requires substantial investments. Setting up a smooth-running, large-scale web survey is not possible without installing several web servers. Hardware and software requirements are much less for mail surveys, but then there are printing and mailing costs.

To reduce the problems that may be caused by nonresponse, one may decide to use incentives. Such incentives are most effective if they are given at or before the time of the interview attempt, and not promised as a reward after a completed interview. This means that the costs of incentives are proportional to the sample size.

Large surveys cost more than small surveys. Unfortunately, the precision of the survey results is also related to the sample size: the larger the sample, the more precise the estimates. Therefore, the higher the demands for precision are, the more expensive the survey will be. In many practical situations, the sample size will be a compromise between costs and precision.

The costs of a survey will be determined by many different factors. It will be clear that the mode of data collection is certainly one of them. But cost is only one aspect that plays a role in choosing the type of survey to conduct. Quality is another aspect. Quality can have many dimensions. For example, Eurostat, the statistical office of the European Union, distinguishes the following five dimensions for the quality of statistics in its European Statistics Code of Practice:

- *Relevance*: European statistics must meet the needs of users.
- *Accuracy and reliability*: European statistics must accurately and reliably portray reality.
- *Timeliness and punctuality*: European statistics must be disseminated in a timely and punctual manner.
- *Coherence and comparability*: European statistics should be consistent internally, over time, and comparable between regions and countries; it should be possible to combine and make joint use of related data from different sources.

- *Accessibility and clarity*: European statistics should be presented in a clear and understandable form, disseminated in a suitable and convenient manner, available, and accessible on an impartial basis with supporting metadata and guidance.

Some of these dimensions of quality may also be relevant for choosing the proper mode of data collection. This section focuses on the quality dimension *accuracy and reliability*.

A survey estimator is called *reliable* if repeating the survey would result in (approximately) the same estimate. Reliability does not imply *validity*. An estimator is valid if it estimates what it intends to estimate. A reliable estimator can produce consistently wrong estimates, for example, if the estimator has a fixed bias. The reliability of an estimator can be quantified by means of its variance or standard error. An estimator is called *precise* if it has a small variance.

An estimator is *accurate* if repeatedly conducting the survey would result in estimates that are all close to the true value. The accuracy of an estimator can be quantified by means of its mean square error. This quantity contains both a variance and a bias component.

EXAMPLE 5.1 Accuracy and reliability of a bathroom scale

Suppose a survey is conducted about obesity. The weight of all respondents is determined by the interviewers using a bathroom scale. If someone who weighs 90 kilograms steps on the scale repeatedly and gets readings of 30, 150, 70, 120, 50, and so on, the scale is not reliable. If all readings are close to 120, the scale is reliable but not accurate. If all readings are close to 90, the scale is reliable and accurate.

Several quality aspects may have a different impact for a different mode of data collection. So they could play a role in deciding the mode of data collection. The following aspects will be taken into account here:

- *Coverage*. Some data collection modes may suffer more from coverage problems. Problems occur if the mode of data collection requires use of a sampling frame that does not coincide with the target population of the survey. For example, a telephone survey may suffer from serious undercoverage if the sampling frame is a list of registered telephone numbers. Another example is a web survey, where those without Internet access will never be selected in the sample.
- *Nonresponse*. The occurrence of nonresponse may cause estimators of population characteristics to be biased. Nonresponse problems seem to be more severe if no interviewers are involved in the survey. An extensive overview of the relationship between nonresponse and the mode of data collection is given in the study by Jelke G. Bethlehem Fannie Cobben, and Barry Schouten (2011).

- *Measurement errors.* The impact of measurement errors on the answers to the survey questions was already described in Chapter 4. Measurement errors can have several causes. One main cause is *satisficing*. This is the phenomenon that respondents do not do all they can to provide a correct answer. Instead they attempt to give a satisfactory answer with minimal effort. Satisficing comes in different forms:
 - *Response order effects.* This is the tendency that the answer selected by the respondent depends on its location in the list of answer options. Primacy effects occur in self-administered surveys (respondents typically pick an answer early in the list of answer options of a closed question), and recency effects occur in interviewer-assisted surveys (respondents typically pick an answer at or near the end of the list of answer options of a closed question).
 - *Acquiescence.* This is the tendency that respondents tend to agree with statements in questions, regardless of their content. They simply answer “yes.” There is less acquiescence in self-administered surveys than in interviewer-assisted surveys.
 - *Status quo endorsement.* Surveys sometimes ask respondents to give their opinion about changes. A typical example is a question of whether government should change its policy with respect to a specific issue. The easiest way to answer such a question without thinking is to select the option to keep everything the same. There seems to be less status quo endorsement in interviewer-assisted surveys.
 - *Nondifferentiation.* This is a form of satisficing that typically occurs if respondents have to answer a series of questions with the same set of response options. Satisficing respondents tend to select the same answer for all these questions irrespective of the question content. The literature suggests there is more nondifferentiation in self-administered surveys.
 - *Answering “don’t know.”* This is a form of satisficing where respondents choose this answer to avoid having to think about a real answer. Not making it possible to answer “don’t know” may also cause measurement errors as respondents not knowing the answer are forced to give one. The way “don’t know” is treated in a survey may depend on the mode of data collection. For example it is generally advised to offer “don’t know” in CAPI and CATI surveys only implicitly. Some research suggests not offering “don’t know” explicitly in a web survey. If respondents skip the question because they do not know they answer, the question reappears, but now with “don’t know” as one of the possible answers.
 - *Arbitrary answer.* Respondents may decide to just pick an arbitrary answer in order to avoid having to think about a proper answer. They may also give an arbitrary answer if giving the proper answer is considered undesirable. This behavior is sometimes also called “metal coin flipping.” This phenomenon typically occurs in web surveys for check-all-that-apply questions.
- *Socially desirable answers.* This is the tendency that respondents give answers that will be viewed as more favorable by others. This particularly happens for

sensitive questions. If a true answer would not make the respondents look good, they will refuse to answer or give a different answer. The literature shows that the effects of socially desirable answers are stronger in interviewer-assisted surveys. Respondents tend to give more truthful answers in self-administered surveys.

- *Questionnaire design effects.* Questionnaires can be presented to interviewers in different ways, depending on the mode of data collection. Questionnaires can be printed on paper or displayed on a computer screen. Electronic questionnaires can consist of a set of pages/screens where each page contains a separate question (the *question-based approach*), or it can consist of one form containing all questions (the *form-based approach*). Particularly for self-administered questionnaires, the design of the questionnaire is very important. An in-depth treatment of many of issues is given by Dillman, Smyth and Christian (2009).
- *Checking.* Errors in the answers to the questions may be detected by carrying out extensive checking during the interview. Most electronic questionnaires contain *domain checks*. For example, for a question about the size of the household of the respondent, it is checked whether the answer is a number within a specific range. Some computer-assisted interviewing software systems also have the possibility of conducting *consistency checks*. For example, if the age of respondent is not older than 12 years, he/she cannot be married. If a problem is detected, it is reported to the interviewer or respondent, so that they can correct the error. Checks generally improve the quality of the collected survey data. Interactive checking is not possible for paper questionnaires.
- *Routing.* Electronic questionnaires may contain routing instructions. These instructions see to it that respondents only answer questions that are relevant to them and that they skip irrelevant questions. Usually routing is forced for electronic questionnaires. For paper questionnaires, routing cannot be forced. Respondents may fail to follow instructions so that they answer the wrong questions or get lost in the questionnaire.
- *Timeliness.* The less time it requires to conduct a survey and to process the collected data, the more relevant the results will be. Computer-assisted modes of data collection will typically take less time than their paper analogues.

The various data collection modes will be considered in more detail in the next sections. They will be scored using the criteria listed in this section.

5.2 Theory

5.2.1 FACE-TO-FACE SURVEYS

A face-to-face survey is a mode of data collection where interviewers visit the homes of the respondents, or another location convenient for the respondent. Together, the interviewer and the respondent complete the questionnaire.

An advantage of face-to-face interviewing is that a visit of an interviewer can be longer than a telephone call. It is not uncommon for respondents to allow interviewers to be in their homes for an hour or more. It is unlikely that a telephone call can take more than 30 minutes. Self-administered survey questionnaires should even be shorter. If more time is available for an interview, more questions can be asked and the questionnaire can be more complex.

Face-to-face surveys are expensive. The main reason is that interviewers do the data collection. They go from one address to the other. Only a part of their time can be spent on interviewing. Travel also costs time. This limits the number of interviews they can do in one day. If no contact can be made with persons at the selected address, callbacks will be made. This will increase costs even more.

Data collected by means of paper questionnaires have to be entered in the computer for further processing. This may require data entry personnel, particularly for large surveys. To clean the data, often a data editing procedure is carried out. The collected data are checked, and the detected errors are corrected. Data editing of large population surveys requires substantial resources.

In case of a CAPI survey, interviewers are equipped with laptop computers. This is a major investment. Because data are already being entered into the computer during the interview, no subsequent separate data entry is required. If also the entered data are checked during the interview, a separate data editing procedure is not needed any more. On the one hand, use of laptops increases the costs of a CAPI survey as compared with a traditional face-to-face survey, and on the other hand, integrating data entry and data editing in the interview decreases the costs.

Samples for face-to-face surveys are usually selected from sampling frames consisting of names and addresses. For example, Statistics Netherlands uses the population register for this purpose. This register covers the complete population, with the exception of illegal immigrants. The postal service in the Netherlands maintains a list of all addresses where mail can be delivered. This list also covers the complete population. So there are no coverage problems. In countries where no population register or address list is available, area sampling may be applied. This will not be the cause of serious coverage problems. Undercoverage may occur if incomplete sampling frames are used. An example is a telephone directory. Such a directory only contains registered telephone numbers. There will usually be a substantial amount of nonlisted numbers. For example, only between 60% and 70% of the Dutch population is listed in the telephone directory.

An important aspect of face-to-face interviewing is that the interviewer and the selected person are together at the same location. This makes it easier for the interviewer to convince a person to participate. This is much more difficult to achieve by telephone or by sending written text. As a consequence, response rates are higher for face-to-face surveys than for other types of surveys.

Goyder (1987) was one of the first to compare the response rates of different modes of data collection. He analyzed a large number of surveys conducted in the United States and Canada in the period from 1930 to 1980. The average response rate of face-to-face surveys was around 67%. The rate was lower for telephone surveys: 60%. The response was lowest for mail surveys: 58%. Similar

conclusions were drawn by Hox and De Leeuw (1994). They compared response rates of a number of surveys in Europe, the United States, and Canada. They found that face-to-face surveys had the highest response rates (on average 70%), followed by telephone surveys (67%) and mail surveys (61%).

The main difference between face-to-face and telephone surveys on the one hand and mail surveys on the other is the presence or absence of interviewers. The presence of interviewers can have advantages. They can provide additional information about the surveys and the questions. Thus, they can assist the respondents in answering the questions. This may help respondents correctly interpreting questions. This reduces the risks of item nonresponse and, therefore, has a positive impact on the quality of the collected data.

Sending an advance letter has proved to increase the response rates of face-to-face surveys. Such letters announce the visit of the interviewer. They also contain background information about the survey and explain why it is important to participate. They take away the surprise of an unexpected visit and provide legitimacy for the survey. Biemer and Lyberg (2003) note that an advance letter may also have a negative effect, as they give selected persons more time to think of an excuse not to participate. However, the literature shows that the prevailing effect is that of higher response rates.

Physical impediments may lead to nonresponse. Doormen, gatekeepers, or locked gates may make it difficult or impossible for an interviewer to make contact with a person. Locked communal entrances or intercom systems prevent making face-to-face contact. As a result, the interviewer cannot show an identification card or a copy of the advance letter. It may even happen that the person is not living any more at that address without the interviewing knowing this.

Response order effects may occur in a face-to-face survey. The interviewer reads out loud the answer options for closed questions, making it difficult for respondents to remember all options. Since only the last few options are still in their short-term memory, they restrict their judgment to these options. As a result, there is a preference for options near the end of the list (recency effect). This effect may be reduced by using show cards. A *show card* contains the list of all possible answers to a question. It allows respondents to read through the list at their own pace and to select the answer that reflects their situation or opinion best.

Acquiescence is the tendency that respondents tend to agree with statements in questions, regardless of their content. They simply answer “yes.” The reason is that respondents only superficially think about the statement offered in the question. This will result in a confirmatory answer. Acquiescence is typically a problem for agree/disagree, true/false, or yes/no questions. The literature suggests that acquiescence is more common among respondents with a lower socioeconomic status. There is less acquiescence in self-administered surveys than in interviewer-assisted surveys. Respondents tend to agree more with statements made in questions if interviewers are present. Without interviewers, respondents may feel more anonymous and, therefore, will be more inclined to answer sensitive questions honestly. All this suggests that acquiescence may be a problem in face-to-face surveys.

Survey questions may ask respondents to express their opinion about changes. A typical example is a question that asks whether government should change its policy with respect to a specific issue. The easiest way to answer such a

question without thinking is to select the option to keep everything the same. If the option of no change is not explicitly offered, not many respondents will insist on giving this answer. However, if this option is available, the number of people selecting it will increase substantially. *Endorsing the status quo* occurs more in self-administered surveys than in interviewer-assisted surveys. Therefore, it is expected to be less of a problem in face-to-face surveys.

Nondifferentiation may occur if respondents have to answer a series of questions with the same response options. The original idea was that this would make it easier for respondents to answer the questions. Changing the response options from question to question would increase the cognitive burden of respondents. However, keeping response options the same is also problematic because respondents tend to select the same answer for all these questions irrespective of the question content. Nondifferentiation can be even more a problem if a series of questions with the same set of answer options is combined into a matrix question. See Figure 5.2 for an example. Respondents simply select all answers in the same column. This is often the column corresponding to the middle response option. This phenomenon is called *straightlining*. Nondifferentiation occurs more in self-administered surveys than in interviewer-assisted surveys. Therefore, it is not a serious problem in face-to-face surveys.

Respondents are sometimes not able to provide the required information because they simply do not know the answer. It therefore seems reasonable to have “don’t know” as one of the answer options. However, this may lead to a form of *satisficing* where respondents choose this option to avoid having to think about a real answer. It is common in CAPI/CATI- software to offer “don’t know” implicitly. Only a list of substantive answer options is shown on the screen and read out to the respondents. See Figure 5.4 for an example. If respondents insist they do not know the answer, the interviewer can record this by using a special key combination. A more or less similar approach can be followed for traditional face-to-face surveys. Initially the respondents can choose from the list of substantive answers. If they insist they do not know the answer, it is recorded as “don’t know” on the form by the interviewer. An alternative approach is to start with a filter question. The respondents are first asked whether they have an opinion about a specific issue. Only if they do, their opinion is asked about it. With these approaches, the treatment of “don’t know” does not seem to be a problem in face-to-face surveys.

Respondents who want to avoid having to think about the proper answer may decide to just pick an *arbitrary answer*. This type of *satisficing* particularly occurs for a special type of question called the *check-all-that-apply question*. It is a closed question for which more than one answer can be selected. It is common practice for computer-assisted surveys to use square check boxes for such questions. The respondents are asked to check all appropriate items in the list of answer options. This can be a lot of work. Instead of checking all relevant answers, they may just check some arbitrary answers and stop when they think they have checked enough answers. This problem typically occurs in mail and web surveys. It is not a problem for interviewer-assisted surveys where interviewers ask for each item in the list separately whether it applies to them. In fact, the check-all-that-apply question is replaced by a set of yes/no questions.

Interviewer-assisted surveys perform less well with respect to answering questions about sensitive topics. The presence of an interviewer may prevent a respondent to give an honest answer to a question about a potentially embarrassing topic. Self-administered surveys (mail, web) perform better. Note that a sensitive question in CAPI may lead not only to item nonresponse but also to a *socially desirable answer*. See De Leeuw (2008) for more information.

The way the questionnaire is designed may have an effect on the way the questions are answered. This is particularly important for self-administered questionnaires. Respondents are usually not familiar with questionnaire forms. If it is not clear if and how questions must be answered, this may be a source of errors. Therefore, it is crucial to pay attention to questionnaire design. Design aspects are less important in interviewer-assisted surveys. The interviewers are in charge of navigating through the questionnaire, asking the questions and recording the answers. They usually have received training to do their job. So, questionnaire design will not be much of an issue in face-to-face surveys.

It is not uncommon in computer-assisted surveys to have *consistency checks* in the questionnaire. These checks are carried out immediately after the relevant questions have been answered. If an inconsistency is detected, an error message is displayed on the screen. The interviewer discusses the problem with the respondent, which may result in correcting the answers to one or more questions. Research shows that these checks improve the quality of the collected data. Therefore it is good to have checks in the CAPI software. It is not possible to implement extensive checking in paper questionnaire forms. This means that such a quality improvement cannot be obtained for face-to-face surveys with traditional paper questionnaire forms.

One important advantage of computer-assisted interviewing is that *automatic routing* instructions can be included in the questionnaire. These instructions see to it that respondents only answer relevant questions, whereas irrelevant questions are skipped. This reduces the number of questions that have to be answered. Moreover, it also avoids respondents to become irritated because they have to answer inapplicable questions. Automatic routing also reduces the workload of the interviewers. They are relieved of the task of making sure the correct route is followed through the questionnaire. Automatic routing can be implemented in CAPI surveys. It is impossible to do this in face-to-face surveys with paper questionnaires. Of course, printed routing instructions can be added, but there are no guarantees that because of errors or confusion respondents end up in the wrong part of the questionnaire.

The fieldwork for face-to-face surveys is time consuming. Interviewers have to travel from one address to the other. This limits the number of interviews they can do on a day. There are substantial differences between computer-assisted data collection and data collection with paper forms in terms of time needed for subsequent processing. Traditional face-to-face data collection is in fact a sequential three-step process. First, there is data collection in the field, resulting in completed paper forms. Second, the data on the forms must be entered into the computer. Particularly for large surveys (in terms of number of questions and number of forms) this may take time. Third, the entered data have to be checked, and detected errors must be corrected. This also takes a considerable amount of time. As a result, the data are not yet ready for analysis straight after completion

of the fieldwork. In contrast, a CAPI survey combines the three steps combined into one. The answers to the questions are immediately entered in the computer and checked during the interview. After completion of the fieldwork, almost no subsequent data processing is required. The data are ready for analysis. So timeliness is less of a problem for CAPI surveys.

Table 5.1 contains a summary of the effects of various phenomena in face-to-face surveys. A plus (+) indicates a positive effect or no negative effect and a minus (–) a negative effect.

TABLE 5.1 Cost and quality aspects of face-to-face surveys

Costs	–	Involving interviewers, who also have to travel, makes a face-to-face survey expensive. Use of laptops increases the costs even more.
Coverage	+	Sampling frames for face-to-face interviewing generally do not exclude persons from the target population.
Nonresponse	+	Face-to-face surveys have higher response rates than other modes of data collection.
Response order effects	–	If the possible answers to a closed question are read out loud, there will be recency effects.
Acquiescence	–	Respondents tend to agree with statements in questions, regardless of their content.
Status quo endorsement	+	If interviewers are present, respondents are less inclined to take the easy way out (select no change).
Nondifferentiation	+	If interviewers are present, respondents are less inclined to select the same answer for a set of questions.
Answering “don’t know”	+	Don’t know is not explicitly offered, but it is accepted as an answer if respondents insist they do not know the answer.
Arbitrary answer	+	Because of the presence of interviewers, respondents are less inclined to give an arbitrary answer.
Socially desirable answers	–	Because of the presence of interviewers, respondents are more inclined to give socially desirable answers. This happens particularly for sensitive questions.
Questionnaire design effects	+	Because the interviewers are in charge of completing the questionnaires, there will generally be no problems.
Checking	+/-	A CAPI questionnaire with checks will improve the quality of the data. For face-to-face interviewing with paper questionnaires, the data may contain errors.
Routing	+/-	A CAPI questionnaire with routing will improve the quality of data. There are no missing data because of answering the wrong questions. Routing errors may occur in paper questionnaires.
Timeliness	–	The fieldwork of a face-to-face survey is time consuming. Data processing after the fieldwork is quicker for a CAPI survey than for a traditional survey.

5.2.2 TELEPHONE SURVEYS

A telephone survey is a mode of data collection where interviewers call selected persons by telephone. If contact is made with the proper person, and this person agrees to cooperate, the interview is started and conducted over the telephone. Telephone interviewing has in common with face-to-face interviewing that there are interviewers asking the questions. This has advantages and disadvantages.

One disadvantage of telephone interviewing is that an interview cannot last as long as a face-to-face interview. To avoid problems, it is generally advised to limit the interview to at most 30 minutes.

CATI is the computerized form of telephone interviewing. It was one of the first modes of data collection to be computerized. From the point of view of the respondent, there is no difference between a traditional telephone survey and a CATI survey. They just answer questions by telephone without being aware of what is on the other end of the line: The interviewers can have a paper questionnaire form or a computer (or both).

An important component of a CATI survey is the call management system. The objective of such a system is to manage and schedule call attempts. It typically may involve the following tasks:

- Offering the interviewer the right telephone number at the right moment, taking into account possible appointments that have been made and the quota the interviewers can handle.
- Handling busy signals. Apparently, someone is at home. Therefore, the number will be offered again after a short while (a few minutes).
- Handling no-answers or answering machine calls. Typically, the number will be offered again at a different day and/or a different time of the day. For example, if there is no answer in the afternoon, the next call could be made in the evening.
- Managing appointments. If an interviewer makes an appointment with a respondent to call back at a specific date and time, the system must offer the call to this interviewer at the appropriate date and time.
- Producing summaries of the progress of the fieldwork.

Telephone surveys are expensive but not as expensive as face-to-face surveys. The interviewers are a major cost component. Because they do not have to travel, but do their work in a call center, they can do more interviews per day than face-to-face survey interviewers. This implies that fewer interviewers are needed for a telephone survey. Moreover, there are also no travel costs.

Data collected by means of paper questionnaires have to be entered in the computer for further processing. This may require data entry personnel, particularly for large surveys. To clean the data, often a data editing procedure is carried out. The collected data are checked, and the detected errors are corrected. Data editing of large population surveys requires substantial resources.

There are two general approaches to sampling for telephone surveys. The first one is to use a list of telephone numbers. An example is a telephone directory. This sampling frame may suffer from serious undercoverage as many people have unlisted numbers. Particularly, the numbers of mobile telephones may be missing. The rapidly increasing popularity of mobile phones makes the undercoverage problems even more substantial. For example, in the Netherlands, only about two thirds of the telephone numbers can be found in the directory. Another problem is the existence of *do-not-call registers*. Although interviewing over the telephone is not the same as selling products or services, many market research organizations avoid calling people in this register. For example, in the United Kingdom, the do-not-call registers contain more than 18 million households (in 2010), which is much more than half of approximately 26 million households. The conclusion can be the telephone directory becomes less and less useful as a sampling frame for a telephone survey.

Another approach to select a sample of telephone numbers is to apply *random digit dialing* (RDD). Taking into account the structure of telephone numbers, a computer algorithm generates random, but valid, telephone numbers. Such an algorithm will produce both listed and unlisted numbers. This guarantees complete coverage. However, also here telephone numbers in the do-not-call register should not be used.

Random digit dialing also has drawbacks. In some countries, it is not clear what an unanswered number means. It can mean that the number is not in use, which is a case of overcoverage. Then no follow-up is needed. It can also mean that someone simply does not answer the phone, which is a case of nonresponse that has to be followed up.

Like in face-to-face surveys, telephone survey interviewers can use their skills to convince persons to participate. Another advantage is that interviewers can supply additional information about the survey and its questions. They can assist respondents in finding the proper answer to a question. The risk of item nonresponse is therefore reduced. Generally, this results in high response rates, although they may be somewhat lower than in face-to-face surveys. Interviewer assistance can also have a positive effect on the quality of the collected data.

All kinds of additional information can be used in the attempt to get contact and to obtain cooperation. Information like composition of the household or the ages of its members may help choose an optimal contact strategy. The amount of available information depends partly on the way in which the sample was selected. For example, Statistics Netherlands selects its samples for telephone surveys from the population register. Next, telephone numbers are added to the selected names addresses by the telephone company. So all register information (gender, age, marital status, location) is available for everyone in the sample.

If the sampling frame contains address information, it is possible to send an advance letter. Such a letter announces the telephone call of the interviewer and, therefore, takes away the surprise of an unexpected call. The letter should provide information about the survey and explain why it is important to participate. To

be effective, advance letters should also mention the research agency, and they must have an official letterhead. Such advance letters will help to increase response rates.

It should be noted that only listed telephone numbers can be linked to addresses and other sampling frame information. Unfortunately, not every telephone number is listed. This is a serious case of undercoverage. Note that if the sample is selected by means of Random Digit Dialing, no information at all is available about the selected addresses.

The fast rise of mobile telephones has not made it easier to conduct telephone surveys. Landline telephones are increasingly replaced by mobile telephones. Landline telephones are typically a means to obtain contact with households, whereas mobile telephones are linked to specific persons. Therefore, the probability of contacting a member of the household is greater for landline telephones. Moreover, if persons can only be contacted through their mobile telephones, it is often in situations that are not fit for interviewing. An additional problem is that sampling frames (telephone directories) in many countries do not contain mobile telephone numbers. A final complication to be mentioned here is that in countries such as the Netherlands people often switch from one telephone provider to another. Usually, this means that they get a different telephone number. It is sometimes possible, but not always easy, to keep the original number. For more information about the problems and possibilities of mobile telephones for interviewing, see the study by Kuusela, Vehovar, and Callegaro (2006).

If RDD is used to select the sample, there is no information at all about the selected persons. This does not help the interviewers in their preparation of calls. It becomes more difficult to persuade reluctant persons. Lack of information from a sampling frame also makes nonresponse correction (for example, adjustment weighting) more difficult.

There may be physical impediments preventing interviewers from making contact with respondents. One example is an answering machine. People may be at home but still have the answering machine switched on. It is not clear whether it is a good idea for the interviewers to leave a message. It may or may not help to get into contact at a next attempt. Groves and Couper (1998) note that sometimes answer machines give some relevant information about the people living at the address. This information may be useful for a next contact attempt.

If no contact is established with a selected person, and possibly also in the case of an initial refusal, one or more subsequent attempts may be made to obtain participation. Fortunately, repeated call attempts are not so expensive (as compared with face-to-face surveys). So, it is relatively easy to do this in practice. It is not uncommon that survey agencies may make six call attempts or more before the case is closed as nonresponse because of no-contact.

Some CATI systems (for example, Blaise) distinguish *call attempts* and *contact attempts*. A contact attempt consists of a series of call attempts within a short time interval. Several contact attempts, each with several call attempts, are made before a case is closed as nonresponse. For example, Statistics Netherlands

makes at most 3 or 4 contact attempts each consisting of at most 3 call attempts. The time interval for contact attempts is one hour in case of no-answer and 5 minutes in case of a busy number.

Many telephone companies have a *calling number identification* (CNID) service. It transmits the survey organization's number to the telephone of the selected person during the ringing signal. This information becomes visible on the telephone of this person. If persons do not recognize the number, or if the number is shown as an "unknown number", they may decide not to pick up the phone. Thus, CNID may be a cause of nonresponse.

The response rates of telephone surveys suffer from telemarketing activities. Typically people are called around dinner time (when the contact probability is high because people are at home) in attempts to sell products or services. This spoils the survey climate. This phenomenon has led to a more hostile attitude toward interviewers. It is therefore important that interviewers make clear at the very start of the interview that they are not selling anything. In some countries (for example, the United States, the United Kingdom, and the Netherlands), there are *do-not-call registers*. When people register, they will not be called any more by telemarketing companies. Such registers may help to improve the survey climate.

Incentives may help to increase response rates. Research has shown that incentives are most effective when they are given before the interview and not after it. To be able to give incentives, the addresses of the respondents must be available. This is typically not the case for RDD surveys. So the possibilities for giving incentives in RDD surveys are limited.

An effective call management system may reduce nonresponse from non-contact. Models may even be developed to predict the optimal time to call. Of course, this requires auxiliary information about the selected persons.

Call management systems can display the results of earlier attempts and other information on the computer screen of the interviewers. This information may help in persuading reluctant respondents. Thus, the refusal rate may go down. See the study by Wagner (2008) for details tuning CATI call management systems with the objective to reduce the nonresponse rate.

Response order effects can occur in telephone surveys. Because the interviewer reads out loud the answer options for a closed question, it is difficult for respondents to remember all options. Because it will be likely that only the last few options remain in their short-term memory, they restrict their judgment to these options. As a result, there is a preference for options near the end of the list (*recency effect*). In the case of face-to-face surveys, this effect may be reduced by using show cards. This is not possible in telephone surveys.

Telephone survey interviews may suffer from *acquiescence* (the tendency that respondents tend to agree with statements in questions, regardless of their content). They only superficially think about the statement offered in the question. They simply give a confirmatory answer. Acquiescence is typically a problem for agree/disagree, true/false, or yes/no questions. Acquiescence seems to be more common among respondents with a lower socioeconomic status. Typically acquiescence occurs in interviewer-assisted surveys. Respondents tend

to agree more with statements made in questions if interviewers are present. Without interviewers, respondents may feel more anonymous and, therefore, will be more inclined to answer sensitive questions honestly. All this suggests that acquiescence may be a problem in telephone surveys.

Survey questionnaires may contain questions asking for opinions about changes. For example, a question asks whether government should change its policy with respect to a specific issue. The easiest way to answer such a question without thinking is to say that everything should remain as it was. If the “no change” option is not explicitly offered, not many respondents will insist on giving this answer. However, if this option is available, the number of people selection will substantially increase. *Endorsing the status quo* occurs more in self-administered surveys than in interviewer-assisted surveys. Therefore, it will be less of a problem in telephone surveys.

Nondifferentiation may occur if respondents have to answer a series of questions with the same response options. This approach is problematic because respondents tend to select the same answer for all these questions irrespective of the question content. Nondifferentiation can be even more a problem if a series of questions with the same set of answer options is combined into a matrix question. See Table 5.2 for an example. Respondents simply select all answers in the same column. This is often the column corresponding to the middle response option. This phenomenon is called *straightlining*. Nondifferentiation occurs more in self-administered surveys than in interviewer assisted surveys. Therefore, it is not a serious problem in telephone surveys.

Respondents sometimes cannot give an answer because they simply do not know the answer. Therefore, it seems reasonable to have “don’t know” as one of the answer options. However, this may lead to *satisficing*: Respondents select “don’t know” to avoid having to think about a real answer. It is common in CATI software to offer “don’t know” implicitly. Only a list of substantive answer options is shown on the screen and read out to the respondents. See Figure 5.4 for an example. If respondents insist they do not know the answer, the interviewer can record this by using a special key combination. A more or less similar approach can be followed for traditional telephone surveys. First a list of substantive answers is read out. If a respondent insists he does not know the answer, it is recorded as “don’t know” on the form by the interviewer. An alternative approach is to use a filter question. First, respondents are asked whether they have an opinion about a specific issue. Only if they do, their opinion is asked about it. With these approaches, the treatment of “don’t know” does not seem to be a problem in telephone surveys.

If respondents do not want to think about a proper answer, they may decide to just pick an arbitrary answer. This type of *satisficing* particularly occurs in *check-all-that-apply questions*. It is a closed question for which more than one answer can be selected. It is common practice for computer-assisted surveys to use square check boxes for such questions. See Figure 4.10 in Chapter 4 for an example. The respondents are asked to check all appropriate items in the list of answer options. This can be a substantial cognitive task. Instead of checking all relevant answers, respondents may just check some arbitrary answers and stop

TABLE 5.2 Cost and quality aspects of telephone surveys

Costs	–	Involving interviewers makes a telephone survey expensive (but not as expensive as a face-to-face survey).
Coverage	–	Telephone surveys may suffer from severe undercoverage. This may be caused by unlisted telephone numbers and numbers in do-not-call registers.
Nonresponse	+	Telephone surveys have higher response rates than self-administered modes of data collection.
Response order effects	–	If the possible answers to a closed question are read out loud, there will be recency effects.
Acquiescence	–	Respondents tend to agree with statements in questions, regardless of their content.
Status quo endorsement	+	If interviewers are present, respondents are less inclined to take the easy way out (select no change).
Nondifferentiation	+	If interviewers are present, respondents are less inclined to select the same answer for a set of questions.
Answering “don’t know”	+	Don’t know is not explicitly offered, but it is accepted as an answer if respondents insist they do not know the answer.
Arbitrary answer	+	Because of the presence of interviewers, respondents are less inclined to give an arbitrary answer.
Socially desirable answers	–	Because of the presence of interviewers, respondents are more inclined to give a socially desirable answer. This happens particularly for sensitive questions.
Questionnaire design effects	+	Because the interviewers are in charge of completing the questionnaires, there will generally be no problems.
Checking	+/-	A CATI questionnaire with checks will improve the quality of the data. For telephone interviewing with a paper questionnaire, the data may contain errors.
Routing	+/-	A CATI questionnaire with routing will improve the quality of data. There are no missing data because of answering the wrong questions. Routing errors may occur in paper questionnaires.
Timeliness	+	The fieldwork of a telephone survey is not so time consuming. Data processing after the fieldwork is quicker for CATI than for traditional telephone surveys.

when they think they have checked enough answers. This problem typically occurs in mail and web surveys. It is not a problem for telephone surveys where interviewers ask for each item in the list separately whether it applies to them. This comes down to splitting one check-all-that-apply question into a set of yes/no questions.

Interviewer-assisted surveys perform less well with respect to answering questions about sensitive topics. The presence of an interviewer may prevent a

respondent to give an honest answer to a question about a potentially embarrassing topic. Self-administered surveys (mail, web) perform better. Note that a sensitive question may lead not only to item nonresponse but also to socially desirable answer. See De Leeuw (2008) for more information. So giving *socially desirable answers* may be a problem in telephone surveys.

Questionnaire design will not be an important issue in telephone surveys. Interviewers are in charge of navigating through the questionnaire, asking the questions, and recording the answers. They usually have received training to do their job.

Computer-assisted surveys may have *consistency checks* in the questionnaire. This helps to detect errors already during the interview. Problems can be repaired immediately. This improves the quality of the collected data. Therefore, checks should be included in a CATI survey. It is not possible to implement extensive checking in paper questionnaire forms. This means that such a quality improvement cannot be obtained for telephone surveys with traditional paper questionnaire forms.

It is possible to make use of *automatic routing* in computer-assisted interviewing questionnaires. Respondents will only have to answer relevant questions, whereas irrelevant questions are skipped. This reduces the number of questions that have to be answered. Moreover, it also avoids respondents to become irritated because they have to answer inapplicable questions. Automatic routing also reduces the workload of the interviewers. They are relieved of the task of making sure the correct route is followed through the questionnaire. Automatic routing can be implemented in CATI surveys. It is impossible to do this in paper questionnaires. Printed routing instructions can be included, but there is always a risk that respondents end up in the wrong part of the questionnaire.

The fieldwork of a telephone survey is not so time consuming. The length of the fieldwork period is determined by the number of cases that can be handled on a day. This depends on the size of the interviewer crew and the success rate of the call attempts. If paper forms are used, subsequent data processing will take time. This includes data entry and data editing. In case of a CATI survey, the answers to the questions are immediately entered in the computer and checked during the interview. After completion of the fieldwork, almost no subsequent data processing is required. So timeliness is even less of a problem for CATI surveys.

Table 5.2 contains a summary of the effects of various phenomena in telephone surveys. A plus (+) indicates a positive effect or no negative effect and a minus (−) a negative effect.

5.2.3 MAIL SURVEYS

Respondents complete the questionnaire themselves in mail surveys. There are no interviewers asking the questions and recording the answers. The respondents themselves read the questions and record the answers. Consequently, there are no interviewers attempting to persuade a reluctant person to fill in the form. There are no interviewers to explain unclear aspects and to assist in answering the

questions. As a result, response rates and data quality will generally be lower in mail surveys than in face-to-face surveys and telephone surveys.

Mail surveys are much cheaper than interviewer-assisted surveys. The reason is the absence of interviewers. They are usually a major cost component. Mail surveys have other costs, such as printing costs and mail costs. These costs are much lower than interviewer costs.

Completed questionnaire forms will be returned to the survey organization. The next step is to enter the collected data into the computer for further processing. This may require data entry personnel, particularly for large surveys. Mail questionnaire forms will not be without errors. One advantage of computer-assisted interviewing is that the answers can be checked during the interview. This is not possible for paper questionnaire forms. Therefore, often a data editing procedure is carried out. The collected data are checked, and the detected errors are corrected. Data editing of large population surveys requires substantial resources.

A list of addresses is required to be able to select a sample for a mail survey. Coverage problems depend on the extent to which this list covers the target population of the survey. For general-population surveys, such a list may be obtained from the postal service organization. They often maintain a list of points where post can be delivered. Usually home addresses can be distinguished from business addresses. A disadvantage of these address lists is that they do not contain the names of the people living there. Therefore, their name cannot be included in the address on the envelope. Instead, a phrase like “The residents of . . .” has to be used. According to Dillman et al. (2009), such a less personalized approach may increase nonresponse. Usually the coverage of address lists is good.

In countries like the Netherlands and the Scandinavian countries, the sample can be selected from the population register. Such a register contains both names and addresses. This makes it possible to implement a personalized approach.

Several literature overviews indicate that the response rates of mail surveys are generally lower than those of interviewer-assisted surveys. See, for example, Goyder (1987) and Hox and De Leeuw (1994). This mainly is caused by the absence of interviewers. Their efforts usually have a positive effect on response rates.

Nonresponse occurs in a mail survey if the questionnaire form is not returned to the survey agency. There is usually no information at all about the reasons for nonresponse. Here are some examples of what the reasons for nonresponse could be:

- The letter did not arrive at the indicated address (no-contact).
- The people were not at home during the survey period (no-contact).
- The letter was received but ignored (refusal).
- The people at the address did not understand the language the letter was written in (not-able).

Bethlehem et al. (2011) stress the importance of distinguishing different types of nonresponse. Every type of nonresponse can have a different impact on survey

outcomes and, therefore, should require its own treatment. Unfortunately, it is not possible to distinguish different types of nonresponse in mail surveys. This makes it more difficult to correct for the effects of nonresponse in such surveys.

To obtain a reasonable response rate, special efforts are required for contacting people and persuading them to participate in the survey. Furthermore, the design of the questionnaire is very important and all other procedures, like advance letters, cover-letters, reminders, and incentives. Dillman et al. (2009) describe this in detail.

The mail to the respondents must not only contain the questionnaire form but also a cover letter. This letter should explain why participation is important and for what purpose the data will be used. The letter should preferably come from a high official in the organization. The letter should not look like it has been photocopied, but it should resemble an original letter (including colored letterheads and a signature). It will also help increasing the response if the letter contains a clear statement that all collected data will be treated as confidential.

It should be as easy for the respondents to return the completed questionnaire. Therefore, it is advised to enclose a postage-paid, return-envelope in the letter to the sampled persons.

Reminders are important for increasing the response rate. If no response is obtained after two to four weeks, a letter can be sent. It reminds persons that their questionnaire has not yet been received and urges them to respond. This reminder letter should contain a replacement questionnaire (in case they lost the original copy).

An incentive can be included in the first letter to the sample persons. Incentives increase response rates. They work best when sent in advance. Examples of incentives are cash payment, lottery tickets, postage stamps, pens, or a donation to a charity organization in the respondent's name. Some research indicates that donations to charity organizations seem not to work as well as real monetary incentives.

Long questionnaires should be avoided. There is empirical evidence that long questionnaires reduce the response rate. For example, Dillman, Sinclair, and Clark (1993) show that shorter questionnaire forms increase the response in census mail surveys.

A final practical advice is to avoid periods in which there is other heavy mail traffic. Examples are the period just before tax forms have to be submitted and the period before the Christmas holiday.

A special form of a mail survey is one that uses questionnaire drop-off and pick-up. Personal delivery of a questionnaire by only slightly trained survey takers may increase the response. This approach also provides more information about nonrespondents.

Response order effects may occur for closed questions. Respondents have to pick the proper answer from a (sometimes long) list of possible answer options. Instead of thinking carefully about the appropriate answer, the first reasonable option is chosen. Mail survey questions may suffer from a special type of response order effect called the *primacy effect*. This is the tendency to pick an answer early in the list of options. Reading to a list of possible options and considering each

option requires a considerable effort. Therefore, respondents may stop at the first reasonable option.

Interviewer-assisted surveys may suffer from *acquiescence*. This is the tendency to agree with statements in questions, regardless of their content. Without interviewers, respondents may feel more anonymous, and therefore, they will be more inclined to answer sensitive questions honestly. This suggests that acquiescence is not a problem in mail surveys.

Surveys sometimes ask questions about opinions about changes. The easiest way to answer is to say that everything should remain as it was. Not many respondents will insist on answering “no change” if this option is not offered. If this option is available, the number of people selected will substantially increase. *Endorsing the status quo* occurs more in self-administered surveys than in interviewer-assisted surveys. Therefore, it may be a problem in mail surveys.

Nondifferentiation occurs if respondents have to answer a series of questions with the same response options. Respondents tend to select the same answer for all these questions irrespective of the question content. Nondifferentiation can be even more a problem if a series of questions with the same set of answer options is combined into a matrix question (see Figure 5.2). Respondents simply select all answers in the same column (*straightlining*). This is often the column corresponding to the middle (neutral) response option. Nondifferentiation occurs more in self-administered surveys than in interviewer-assisted surveys. Therefore, it may be a problem in mail surveys.

The treatment of “don’t know” in mail surveys requires careful consideration. Offering “don’t know” explicitly as one of the answer options has the advantage that respondents not knowing the answer can answer so. This approach accepts the existence of a group of persons that cannot answer the questions, and thus, “don’t know” is considered a substantive answer. This approach may suffer from *satisficing*. People not wanting to think about an answer or not willing to provide an answer have an escape by answering “don’t know.” Several authors have shown that explicitly offering “don’t know” substantially increases the percentage of respondents choosing this option. See, for example, Sudman and Bradburn (1982).

To avoid satisficing, one may decide to not offer the option “don’t know.” Consequently, respondents always have to provide a substantive answer, even if they do not know the answer. According to Couper (2008), this violates the norm of voluntary participation and therefore may frustrate respondents resulting in nonresponse. Also, Dillman (2007) strongly recommends not forcing respondents to answer. He warns about negative effects on respondent motivation, data quality, and response.

Satisficing may also be reduced by using a filter question. See the studies by Krosnick (1991) and Schuman and Presser (1981). This filter question asks whether respondents have an opinion about a specific issue. Only if they say they have, a next question asks what their opinion really is.

If respondents do not want to think about an answer, they may decide to just pick an *arbitrary answer*. This type of satisficing particularly occurs in check-all-that-apply questions. Instead of checking all relevant answers, they may just

check some arbitrary answer options and stop when they think they have checked enough. This problem typically occurs in self-administered surveys. So it can be a problem in mail surveys.

A mail survey may perform better than an interviewer-assisted survey with respect to answering questions about sensitive topics. The absence of an interviewer may encourage a respondent to give an honest answer to a question about a potentially embarrassing topic. So there is a lower risk of *socially desirable answers*.

Questionnaire design is of crucial important in mail surveys. There are no interviewers assisting respondents in navigating through the questionnaire, asking the questions, explaining, and recording the answers. The respondents are completely on their own. Unclear questions and navigation may result in wrong answers or no answers at all.

A lot of attention must be paid to the design of the questionnaire form. It must look attractive to the respondents. The more personalized it is, the better it works, and the more likely it is that respondents will complete it. The clearer it is, the less likely it is they will get confused, resulting in errors.

Navigation instructions (*routing instructions* and skip patterns) are a potential source of error. If respondents fail to follow the correct route through the questionnaire, wrong questions are answered and right questions are left unanswered, which comes down to item nonresponse. Navigation instructions must be clear and unambiguous. Dillman et al. (2009) advise indicating jumps to other questions by arrows or other graphs with a similar meaning. See Figure 5.2 for an example.

Computer-assisted surveys may have *consistency checks* in the questionnaire. This helps to detect errors already during the interview. Problems can be repaired immediately. This improves the quality of the collected data. It is not possible to implement extensive checking in paper questionnaire forms. This means that such a quality improvement cannot be obtained for mail surveys.

It is possible to make use of *automatic routing* in computer-assisted interviewing questionnaires. Respondents will only have to answer relevant questions, whereas irrelevant questions are skipped. This reduces the number of questions that have to be answered. Moreover, it also limits respondents from becoming irritated because they have to answer inapplicable questions. Automatic routing also reduces the workload of the interviewers. They are relieved of the task of making sure the correct route is followed through the questionnaire. It is impossible to do this in mail surveys. Of course, printed routing instructions can be included in the questionnaire form, but there is always a risk that respondents end up in the wrong part of the questionnaire.

The fieldwork of a mail survey may be time consuming. It takes time to send the empty questionnaire forms to the respondents. Respondents tend to be slow in returning the completed forms. Often reminders have to be sent several times. This may take weeks. Also subsequent data processing will take time. This includes data entry and data editing.

Table 5.3 contains a summary of the effects of various phenomena in mail surveys. A plus (+) indicates a positive effect or no negative effect and a minus (-) a negative effect.

TABLE 5.3 Cost and quality aspects of a mail surveys

Costs	+	Because of the absence of interviewers, costs are relatively low.
Coverage	+	Available address lists have a good coverage of the population.
Nonresponse	-	Mail surveys have lower response rates than interviewer-assisted modes of data collection.
Response order effects	-	If the possible answers to a closed question are read by the respondent, there will be primacy effects.
Acquiescence	+	The tendency to agree with statements in questions, regardless of their content, is less for self-administered modes of data collection.
Status quo endorsement	-	If no interviewers are present, respondents are more inclined to take the easy way out (select no change).
Non-differentiation	-	If no interviewers are present, respondents are more inclined to select the same answer for a set of questions.
Answering "don't know"	-	It is generally advised to offer "don't know" as one of the answer options. Particularly for opinion questions, respondents may use this option as an escape for not giving an substantial answer.
Arbitrary answer	-	If no interviewers are present, respondents are more inclined to give an arbitrary answer.
Socially desirable answers	+	If no interviewers are present, respondents are less inclined to give a socially desirable answer. This happens particularly for sensitive questions.
Questionnaire design effects	-	Questionnaire design is critical in mail surveys. A suboptimal design may have severe consequences.
Checking	-	It is not possible to include checks in mail survey questionnaire forms.
Routing	-	Routing takes the form of printed instructions for the respondents. There is no guarantee the proper route will be followed.
Timeliness	-	Collecting the completed questionnaire forms may take time, even more if reminders are sent. Data processing is slower for mail surveys than for computer-assisted surveys.

5.2.4 WEB SURVEYS

The basic feature of a web survey is that the questionnaire is designed as a website, which is accessed by respondents. Web surveys combine some aspects of self-administered surveys and computer-assisted surveys. On the one hand, a web survey looks like a mail survey but with the questionnaire on the computer screen instead of on paper. On the other hand, a web survey may include facilities like error checking and routing.

Web surveys have become very popular in a short time. This is not surprising as web surveys seem to have some attractive advantages: (1) it is a simple means to

get access to a large group of potential respondents, (2) questionnaires can be distributed at very low costs, (3) surveys can be launched very quickly, and (4) they offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation, and movies). So web surveys seem to be a fast, cheap, and attractive means of collecting large amounts of data. In some survey situations, a web survey is indeed an effective mode of data collection that produces high-quality data. However, conducting a web survey in a scientifically sound way is sometimes not so easy, particularly with large population surveys. Some of the advantages may be not so clear any more.

A web survey can be a cheap data collection instrument. No interviewers are involved as well as no mailing and printing costs. Of course, there are hardware and software costs, but a well-designed survey infrastructure can be used for many different surveys. The costs may increase for sample selection. If a list with e-mail addresses is available, sample selection is cheap and straightforward. The situation becomes more difficult if a sample has to be selected for a general-population survey. There is no sampling frame with e-mail addresses. So a different mode has to be used to recruit persons for the survey. One way to realize this is to send a letter to the sample persons containing a link to the website and a unique identification code to start the questionnaire. If the persons do not respond, they may be called by telephone in an attempt to encourage them to complete the questionnaire. And if this fails, it may be considered to visit them at home. It will be clear that such an approach implies a substantial increase in the costs of the web survey.

Depending on the target population of the survey, undercoverage may be a serious problem. In simple situations, like a survey among employees of a company or students at a university, there usually is a list of e-mail addresses. These lists cover the population completely. However, in many situations, the target population is usually wider than the Internet population. This typically applies to general-population surveys. For example, Internet coverage varies between 30% and 90% in European countries. Broadband Internet access is even lower. This prevents web surveys with advanced features (requiring broadband). See Chapter 8 for more details.

Unfortunately, Internet access is unevenly distributed over the population. A typical pattern found in many countries is that the elderly, the low-educated, and ethnic minorities are severely underrepresented among those having access to Internet.

If undercoverage in a web survey really is a problem, a possible simple solution could be to provide Internet access to those without Internet. An example of this approach is the Longitudinal Internet Studies for the Social Sciences (LISS) panel, described by Scherpenzeel (2008). This web panel has been constructed by selecting a random sample of households from the population register of the Netherlands. Selected households were recruited for this panel by means of CAPI or CATI. Cooperative households without Internet access were provided with equipment giving them access to the Internet.

Web surveys suffer from nonresponse. As a web survey is a self-administered survey, it has the potential for high nonresponse rates. An additional source of

nonresponse problems are the technical issues of respondents having to interact with the Internet [see Couper (2000), Dillman and Bowker (2001), Fricker and Schonlau (2002), and Heerwegh and Loosveldt (2002)]. Respondents need a browser to open and complete a web survey questionnaire. Many different browsers are available. Examples are Internet Explorer, Firefox, Safari, Opera, and Google Chrome. These browsers do not behave in exactly the same way. Therefore, a questionnaire may behave differently in different browsers. A feature may not even work in a specific browser, preventing respondents from recording their answers to the questions. This may result in nonresponse.

The Internet is in continuous development. Specific features may work in the most recent version of a browser but not in earlier versions. Unfortunately, not all people have the latest version of their browser installed. Again, as a result, a specific feature in the questionnaire may or may not work. Some questionnaire features (for example, use of animation and video) require an Internet connection with substantial bandwidth, but not every Internet user has a broadband Internet connection. So, these features will not work properly on their computer.

Slow modem speeds, unreliable connections, high connection costs, low-end browsers, and incompatible browsers may frustrate respondents, often resulting in prematurely interrupting completion of the questionnaire. To keep the survey response up to an acceptable level, every measure must be taken to avoid these problems. This requires a careful design of web survey questionnaire instruments.

Response order effects may occur in web surveys. Instead of thinking carefully about the appropriate answer, the first reasonable option of a closed question is chosen. Web survey questions may suffer from a special type of response order effect called a *primacy effect*. This is the tendency to pick an answer early in the list of options. Instead of reading the list of possible options and considering each option, respondents may stop at the first reasonable option.

Interviewer-assisted surveys may suffer from *acquiescence*. This is the tendency to agree with statements in questions, regardless of their content. Without interviewers, respondents may feel more anonymous, and therefore, they will be more inclined to answer sensitive questions honestly. Therefore, acquiescence is not a problem in web surveys.

Survey questions sometimes ask for opinions about changes. The easiest way to answer is to say that everything should remain as it was. Therefore, if there is a “no change” option, many respondents will choose it. This phenomenon of *endorsing the status quo* typically occurs in self-administered surveys. So it may be a problem in web surveys.

Nondifferentiation occurs if respondents have to answer a series of questions with the same set of response options. Respondents tend to select (without thinking) the same answer for all these questions irrespective of the question content. Nondifferentiation can be even more a problem if a series of questions with the same set of answer options is combined into a matrix question. Respondents simply select all answers in the same column (*straightlining*). This is often the column corresponding to the middle (neutral) response option. Nondifferentiation occurs more in self-administered surveys than in interviewer-assisted surveys. Therefore, it may be a problem in web surveys.

Various approaches are possible are to offer the option “don’t know” as an answer to a question. On the one hand, it can be explicitly shown in one of the options. This will encourage *satisficing*. Many will see this option as an escape for having to think about the proper answer. On the other hand, one may decide to not offer the option “don’t know.” Then respondents always have to provide a substantive answer, even if they do not know the answer. Experts advise against this approach as it may lead to nonresponse.

Vis-Visschers et al. (2008) investigated an approach whereby respondents were first offered the question with only substantive answer options. If they did not select an option and attempted to skip the question, the question was offered again, but then “don’t know” was included in the list. It turned out that some respondents did not understand this mechanism, as they complained that they could not answer “don’t know.” The other respondents less frequently selected “don’t know.”

DeRouvray and Couper (2002) experimented with a similar approach. The answer option “don’t know” was not offered for the question. If respondents attempted to skip the question without answering it, a new screen appeared offering two choices: (1) go back and answer the question, and (2) record the answer as “don’t know” and proceed to the next question. This approach resulted in the lowest “don’t know” rates.

If respondents do not want to think about an answer, they may decide to pick just an *arbitrary answer*. This type of *satisficing* particularly occurs in *check-all-that-apply questions*. Instead of checking all relevant answers, they may just check some arbitrary answer options and stop when they think they have checked enough. This problem typically occurs in self-administered surveys. So it can be a problem in web surveys.

A web survey may perform better than an interviewer-assisted survey with respect to answering questions about sensitive topics. The absence of an interviewer may encourage a respondent to give an honest answer to a question about a potentially embarrassing topic. Therefore, there may be less *socially desirable answers* in web surveys.

Designing a questionnaire for a web survey is to some extent similar to designing it for a mail survey. At first sight, one could say that a web survey questionnaire is nothing more than a paper form displayed on a computer screen. There are, however, also differences that may affect the answers to the questions. For example, a paper questionnaire page usually does not fit on the computer screen. This means the respondent has to scroll to see all parts of the page. Failing to do so may mean that some (not visible) questions are skipped, resulting in item nonresponse.

The designer of a web survey has the choice to display just one question per screen or to put more questions on the screen. One question per screen may be more appropriate if the questionnaire is to contain extensive routing instructions (skip patterns). However, this will increase the perceived length of the questionnaire, possibly resulting in (partial) nonresponse.

There are many more design issues that may affect response to a web survey. See Chapter 4. A detailed description is also given in a study by Couper (2008).

CAPI and CATI surveys questionnaires may contain extensive edit checks. These checks help to detect and correct inconsistencies in the answers of respondents. Such checks are not possible in mail surveys, and this may have a negative effect on the quality of the collected data. The designer of a web survey questionnaire can decide to include edit checks. This should increase data quality, but many (unfriendly) error messages may scare away respondents. So, reporting and treating errors should be implemented carefully.

Respondents are completely free in the way they complete a paper questionnaire. They can answer any question they like in any order. Web questionnaires can be designed such that routing is enforced like in CAPI or CATI surveys. This sees to it that only relevant questions are answered and that irrelevant ones are skipped. Routing can only be implemented in a meaningful way if the questionnaire is processed on the basis of one question per screen (the *question-based* approach). It is difficult, if not impossible, to implement routing for a *form-based* approach.

Conducting a web survey is not time consuming. No interviewers are involved. Questionnaires do not have to be sent by mail. There is no separate data entry phase. If the web survey questionnaire contains checks, there is no need for a subsequent data editing phase. In the case of web panels, it is even possible to conduct a survey in one day. Carrying out a web survey may become more time consuming if there is no proper sampling frame. If respondents have to be recruited by means of a telephone or face-to-face survey, the survey will take more time.

Table 5.4 contains a summary of the effects of various phenomena in web surveys. A plus (+) indicates a positive effect or no negative effect and a minus (–) a negative effect.

TABLE 5.4 Cost and quality aspects of a web surveys

Costs	+	There are no interviewer, printing, and mailing costs. Recruitment can be expensive without a good sampling frame.
Coverage	–	There is serious undercoverage in many situations.
Nonresponse	–	Web surveys have lower response rates than interviewer-assisted modes of data collection.
Response order effects	–	If the possible answers to a closed question are read by the respondent, there will be primacy effects.
Acquiescence	+	The tendency to agree with statements in questions, regardless of their content, is less for self-administered modes of data collection.
Status quo endorsement	–	If no interviewers are present, respondents are more inclined to take the easy way out (select no change).
Non-differentiation	–	If no interviewers are present, respondents are more inclined to select the same answer for a set of questions.
Answering “don’t know”	+	It is generally advised to offer “don’t know” as one of the answer options. Particularly for opinion questions, respondents may use this option as an escape for not giving a substantial answer.

(Continued)

TABLE 5.4. Continued

Arbitrary answer	–	If no interviewers are present, respondents are more inclined to give an arbitrary answer.
Socially desirable answers	+	If no interviewers are present, respondents are less inclined to give a socially desirable answer. This happens particularly for sensitive questions.
Questionnaire design effects	–	Questionnaire design is critical in web surveys. A suboptimal design may have severe consequences.
Checking	+	It is possible to include checks in web surveys. Presenting and treating errors is critical.
Routing	+	Automatic routing can be implemented for the question-by-question implementation.
Timeliness	+	A web survey can be conducted quickly. However, recruitment by means of mail/CAPI/CATI may take time.

5.3 Application

The Blaise system is a software system that supports various modes of data collection. It was developed by Statistics Netherlands as a tool for making survey data collection faster and at the same time for improving the quality of the data. Blaise can be used for large and complex questionnaires. It is used by many national statistical institutes in the world. Blaise is used in this section to illustrate the advantages and disadvantages of the different modes of data collection. More about the development of the Blaise system and its underlying philosophy can be found in the study by Bethlehem and Hofman (2006).

A questionnaire is designed in Blaise using a scripting language. It defines the questions to be asked, the possible answers to the questions, the order in which the questions must be asked, and the conditions under which the questions have to be asked. Moreover, relationships can be defined that have to be checked. Once the Blaise questionnaire definition is ready, the system can generate the software tools for various modes of data collection: CADI (*computer-assisted data input*) for data entry of paper forms, CAPI, CATI, and web.

A small election survey questionnaire is used as an example. Figure 5.1 contains the specification of this questionnaire in the Blaise system. The first part of the questionnaire specification is the *Fields section*. It contains the definition of all questions. A question consists of an identifying name, the text of the question as presented to the respondents, and a specification of valid answers. For example, the question about voting behavior has the name *Voted*, the text of the question is “*Did you vote for the parliamentary election on June 2, 2110?*”, and there are three answer options. Each option has a name (for example, *Yes*) and a text for the respondent (for example, “*Yes, did vote*”).

Almost all questions in this sample questionnaire are closed questions. There is one exception: the question *DatBirth* asks for the date of birth. The answer must be a date.

DATAMODEL ElectionSurvey "The 2010 Election Survey"**FIELDS**

Democracy

"On the whole, are you very satisfied, fairly satisfied, not very satisfied or not satisfied at all with the way democracy works in the country?":

(VerySat "Very satisfied",
SomeSat "Somewhat satisfied",
Neutral "Neither satisfied nor dissatisfied",
SomeDis "Somewhat dissatisfied",
VeryDis "Very dissatisfied")

NewsTV

"How much attention did you pay to the TV news about the election":

(Lot "A lot", Fair "Fair amount", Some "Some", Little "Little", None "None")

NewsRad

"How much attention did you pay to the radio news about the election":

(Lot "A lot", Fair "Fair amount", Some "Some", Little "Little", None "None")

NewsPap

"How much attention did you pay to the news in newspapers about the election":

(Lot "A lot", Fair "Fair amount", Some "Some", Little "Little", None "None")

NewsWeb

"How much attention did you pay to the news on the Internet about the election":

(Lot "A lot", Fair "Fair amount", Some "Some", Little "Little", None "None")

Voted

"Did you vote for the parliamentary election on June 2, 2010?":

(Yes "Yes, did vote",
CouldNot "Could not vote",
DidNot "Could vote, but did not vote")

Party

"Which party did you vote for?":

(Con "Conservative Party",
Soc "Social Democratic Party",
Lib "Liberal Party",
Green "Green Party",
Oth "Other party"), DONTKNOW

SecParty

"Which party was your second choice?":

(Con "Conservative Party",

```

Soc    "Social Democratic Party",
Lib    "Liberal Party",
Green "Green Party",
Oth    "Other party",
None  "None"), DONTKNOW
WhyNot
  "What is the reason you did not vote?":
  (NoTime "No time, too busy",
  NotInt  "Not interested",
  PhysLim "Physical limitations",
  OthReas "Other reason")
DatBirth
  "What is your date of birth?": DATETYPE
MarStat
  "Are you presently married, living with a partner,
  divorced,
  separated, widowed, or have you never been married?":
  (Married "Married",
  Partner  "Living with a partner",
  Divorce  "Divorced",
  Separat  "Separated",
  Widowed  "Widowed",
  NevMarr  "Never been married")

RULES
Democracy
NewsTV NewsRad NewsPap NewsWeb
Voted
IF Voted = Yes THEN
  Party SecParty
  Ord(Party) <> Ord(SecParty) "Your first choice was ^Party.
  Your second choice must
  be a different party."
ELSEIF Voted = DidNot THEN
  WhyNot
ENDIF
DatBirth
IF (Voted = Yes) OR (Voted = DidNot) THEN
  DatBirth <= (1992, 6, 2) "You are too young to vote!"
ENDIF
MarStat
ENDMODEL

```

FIGURE 5.1 A questionnaire definition in Blaise

By default the answer “don’t know” is forbidden for all questions in Blaise. To allow this answer, the keyword *DONTKNOW* must be added to the question definition. This has been done for the questions *Party* and *SecParty* in Figure 5.1. It means that in a computer-assisted data collection mode, this answer option is implicitly available. It can be selected with the function key $\langle \text{Ctrl-K} \rangle$.

The second part of the Blaise specification is the *Rules section*. Here, the order of the questions is specified as well as the conditions under which they are asked. According to the rules section in Figure 5.1, every respondent must answer the questions *Democracy*, *NewsTV*, *NewsRad*, *NewsPap*, *NewsWeb*, and *Voted*. Only persons who voted (*Voted=Yes*) have to answer the questions *Party* and *SecParty*. Respondents who could vote but did not vote (*Voted=DidNot*) are asked why they did not vote (*WhyNot*). Finally, all respondents have to provide their date of birth (*DatBirth*) and marital status (*MarStat*).

The Rules section contains two checks. The first one checks that the second choice for a party is different from the first choice (*Ord(Party) <> Ord(SecParty)*). Note that a text label is attached to the check. This text appears in the error message on the screen. The second check produces an error message if a voter is younger than 18 years (*DatBirth <= (1992, 6, 2)*).

The election survey questionnaire has been formatted as a paper questionnaire in Figure 5.2. It is a disadvantage of a paper questionnaire that there is no software in charge of the proper route through the questionnaire. There are printed instruction like “Go to 7,” but there is no guarantee that these instructions will be followed. There is always a risk that respondents end up in the wrong part of the questionnaire.

The four questions about paying attention to news about the election campaign all have the same answer options. Therefore, they have been formatted as a matrix question. This saves space in the questionnaire, but there is also a risk of straightlining: respondents may make it easy for themselves by selecting all answer options in the same column. This is not the only form of satisficing that may occur. Another is that respondents may not have an opinion or do not want to think about an opinion. As result, they choose the middle option for all four questions.

The question asking for the data of birth contains clear instructions as to how the date must be written. These instructions are important. Without them an answer may be confusing. For example, if someone writes 4-3-1951, it is unclear whether this means March 4 or April 3.

The two questions about party choice contain “don’t know” as an explicit answer option. If there are respondents who really do not know the answer, this may be the best approach. Of course, this creates an escape route for those not wanting to give an answer.

After the completed paper forms have been sent back to the survey agency, the data must be entered into the computer. Some kind of data entry tool can be helpful to do this in an efficient way. Blaise implements an approach that is called CADI. It is a combination of data entry and data editing. The CADI program can be automatically generated from the questionnaire definition.

1. **On the whole, how satisfied are you with the way the democracy works in the country?**

Very satisfied
 Somewhat satisfied
 Neither satisfied nor dissatisfied
 Very satisfied
 Somewhat satisfied

2. **How much attention did you pay to news about the election campaign in each of the following media?**

	A lot	Fair amount	Some	Little	None
Television	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Radio	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Newspapers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Internet	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3. **Did you vote for the parliamentary election of June 2, 2010?**

Yes → *Go to 5*
 Could not vote → *Go to 7*
 Could vote, but did not vote

4. **What is the reason you did not vote?**

No time, too busy
 Not interested
 Physical limitations
 Other reason
 Don't know

→ *Go to 7*

5. **Which party did you vote for?**

Conservative Party
 Social-democratic Party
 Liberal Party
 Green Party
 Other party
 Don't know

6. **Which party was your second choice?**

Conservative Party
 Social-democratic Party
 Liberal Party
 Green Party
 Other party
 None
 Don't know

7. **What is your date of birth?**

dd mm yyyy

8. **Are you presently married, living with a partner, divorced, separated, widowed, or have you never been married?**

Married
 Living with a partner
 Divorced
 Separated
 Widowed
 Never been married

FIGURE 5.2 The paper version of the questionnaire

Question	Answer	Label
Democracy	1	Verysat
NewsTV	1	Lot
NewsRad	2	Fair
NewsPap	1	Lot
NewsWeb	5	None
Voted	1	Yes
Party	1	Con
SecParty	1	Con
WhyNot		
DatBirth		
MarStat		

FIGURE 5.3 The data entry program

Figure 5.3 shows an example of a screen of the CADI program for the election survey. The questions are indicated by their short identifying names. If necessary, the complete question text can be displayed by pressing a special function key.

Data entry typists enter the answers to the questions. They are completely free in the order in which they enter data. They are not constrained by routing instructions. The idea is to first copy all data from the form to the computer, after which the answers are checked, and corrections can be made.

Special symbols in front of the input fields denote problems in the data. The answers to two questions *Party* and *SecParty* in Figure 5.3 are inconsistent. A diagnostic error message can be displayed by moving the cursor to the field and pressing a special function key (“*Your first choice was Conservative Party. Your second choice must be a different party*”).

The two symbols for the questions *DatBirth* and *MarStat* indicate that these questions are on the route through the questionnaire and therefore have to be answered. These symbols will disappear once the answers have been entered into the respective fields.

Note that because the paper questionnaire and the CADI program have been generated from the same source (the Blaise questionnaire specification), the form and the program are consistent with each other.

The Blaise system also supports CAPI and CATI. A program for computer-assisted interviewing can be automatically generated from the questionnaire specification. Figure 5.4 shows an example of a screen for the election survey.

The screen is divided into two parts. The top half contains the current question to be answered. The bottom half is a condensed view of the questionnaire form. It shows which questions have already been answered and which

The 2010 Election Survey

Forms Answer Navigate Options Help

Did you vote for the parliamentary election on June 2, 2110?

1. Yes, did vote
 2. Could not vote
 3. Could vote, but did not vote

Democracy	1	VerySat	WhyNot
NewsTV	2	Fair	DatBirth
NewsRad	1	Lot	MarStat
NewsPap	5	None	
NewsWeb	1	Lot	
Voted	1		
Party			
SecParty			

New 1/1 Modified Dirty Insert ElectionSurvey

FIGURE 5.4 The computer-assisted interviewing program

questions still must be answered. This gives the interviewers an overview of where they are in the questionnaire.

Routing is forced in this questionnaire. It is only possible to move forward to the next question on the route if the current question has been properly answered. Changes can be made in the questionnaire by moving back to a previous question.

If an error is encountered in the answers, an error message will be displayed on the screen. Figure 5.5 contains an example. The message also shows the questions involved in the error. The interview can only go back to one of these questions and correct its answer. It is not possible to proceed to the next question on the route as long as the error has not been corrected.

This approach of forced routing and forced error correction has shown to improve the quality of the collected data. The data contain fewer errors, and there is less item nonresponse.

The questions *Party* and *SecParty* do not have an option “Don’t know.” This option is implicitly available for these questions. If the respondents insist they really do not know they answer, the interviewer can press $\langle \text{Ctrl-K} \rangle$ to record the answer as “Don’t know.”

The Blaise system uses the same computer-assisted interviewing program for CAPI and CATI. There are other tools that are specific for either CAPI or CATI. For example, there is an extensive call management system for CATI. This system sees to it that the right telephone number is called at the right time. In



FIGURE 5.5 A detected consistency error

case of a busy signal, a subsequent call attempt is scheduled after a short while. If there is no answer, typically, a new call attempt will be scheduled at a different day and/or time of the day. Interviewers can also make appointments with respondents to call back at a specific date and time. This call management system helps to increase response rates.

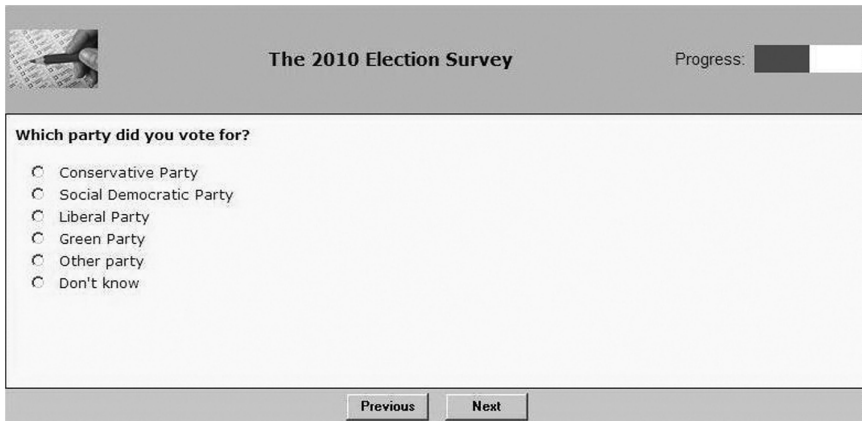
It is also possible to carry out a web survey with Blaise. The system distinguishes two approaches:

- The *question-based approach*. There is only one question on the screen. After this question has been answered, the answer is sent to the web server and checked. The system determines the next question on the route, and this question will appear on the screen. This approach should be used if the questionnaire contains routing instructions and checks. The question-based approach requires the respondent to be on-line while answering the questions.
- The *form-based approach*. The whole questionnaire is displayed on the screen as one form. This approach can be used if there are no routing instructions and checks. All questions are answered, after which the answers are sent to the web server. The respondent does not need to be on-line while answering the questions.

It is also possible to mix both approaches. The idea is to divide the questionnaire into several subquestionnaires each without routing. There is a form for each subquestionnaire. After a form has been completed, it is sent to the web server, where it is checked. Then a next form will be sent to the respondent.

Figure 5.6 contains an example of a screen for the election survey. Because the questionnaire contains several checks and route instructions, it has been implemented in the question-based approach.

The web questionnaire has been made somewhat more attractive by adding a logo. Also note that there is a progress bar in the top-right corner. This may help to keep respondents motivated, as long as the progress bar does not behave too wild because of large jumps in the questionnaire.



The screenshot shows a web survey interface. At the top left is a small image of a hand holding a pen over a document. The title "The 2010 Election Survey" is centered at the top. To the right of the title is a progress indicator labeled "Progress:" followed by a black and white bar. Below the title bar is a question: "Which party did you vote for?". The question has six radio button options: "Conservative Party", "Social Democratic Party", "Liberal Party", "Green Party", "Other party", and "Don't know". At the bottom of the question area are two buttons: "Previous" and "Next".

FIGURE 5.6 The web survey

The problem of “don’t know” is treated by including this answer as one of the answer options for the question. At the same time, respondents have to answer this question. They cannot proceed without answering it. This may help prevent skipping questions too quickly. Nevertheless, there is always the risk that respondents answer “don’t know” as the easy way out.

The routing is forced in this web questionnaire. Respondent can only move back to the previously answered question. They can only move forward to the next questionnaire on the route if they have properly answered the current question. The routing mechanism is the same as that of the CAPI/CATI program.

It is possible to include checks in the Blaise web questionnaire. This may help to increase data quality. However, it should be realized that many unfriendly error messages may frustrate respondents resulting in (partial) nonresponse. It is therefore advised to pay careful attention to the way error messages are presented to respondents.

It is an attractive feature of the Blaise system that software tools for various modes of data collection can be generated from the same questionnaire specification. This forces these instruments to be consistent with each other. This makes the system suitable for mixed-mode data collection.

5.4 Summary

A researcher can choose from various modes of data collection. They all have their advantages and disadvantages with respect to cost, timeliness, and data quality.

Interviewer-assisted modes (CAPI, CATI) are expensive, but they produce good-quality data. Other modes (mail, web) are less expensive, but a price may have to be paid in terms of data quality.

With respect to response rates, the best option is to use an interviewers-assisted mode of data collection. This applies to face-to-face and telephone interviewing, and to their computer-assisted analogues CAPI and CATI. The presence and efforts of interviewers often leads to higher response rates. Response rates are low in web surveys.

Although mail and web surveys are not as good as the interviewer-assisted modes in terms of measurement errors, they perform better with respect to sensitive questions. Respondents tend to answer this type of questions better if no interviewers are present.

Web surveys grow in popularity. Until now, their response rates have now disappointingly low. This is partly because of the self-administered nature of this mode of data collection. Nonresponse may, however, also be caused by technical problems, like slow modems (no broadband connection) and old or incompatible browsers.

Web surveys may also suffer from undercoverage problems, particularly with large population surveys. This problem may solve itself in the future. In the mean time, mixed-mode surveys may help to get into contact with those without access to the Internet.

Another problem of web surveys is that often a proper sampling frame is lacking. This problem could be solved by recruiting respondents by means of another mode of data collection (mail, CATI, CATI). A disadvantage of this approach is that it makes the web survey much more expensive and conducting it will also take more time.

Because respondents are on their own when they answer the questions, it is of crucial importance that the design of the questionnaire is such that it helps them to perform this task properly.

KEY TERMS

Accurate: An estimator that always results in estimates close to the true value (if the survey is repeated).

Acquiescence: The phenomenon that respondents tend to agree more with statements in questions if interviewers are present.

Blaise: A software package for computer-assisted interviewing and survey processing developed by Statistics Netherlands.

Computer-assisted interviewing (CAI): A form of interviewing in which the questionnaire is not printed on paper. Questions are asked by a computer program.

Computer-assisted personal interviewing (CAPI): A form of face-to-face interviewing in which interviewers use a laptop computer to ask the questions and to record the answers.

Computer-assisted self-administered questionnaires (CSAQ): A form of data collection in which respondents complete the questionnaires on their own computer. See also CASI.

Computer-assisted self-interviewing (CASI): A form of data collection in which respondents complete the questionnaires on their own computer. See also CSAQ.

Computer-assisted telephone interviewing (CATI): A form of telephone interviewing in which interviewers use a telephone to ask the questions and to record the answers.

Computer-assisted web interviewing (CAWI): A form of self-interviewing in which respondents complete the questionnaires on the Internet. CAWI is a synonym for web survey.

Face-to-face survey: A survey where interviewers visit the homes of the respondents (or another location convenient for the respondent). Together, the interviewer and the respondent complete the questionnaire.

Mail survey: A survey where paper questionnaire forms are sent to the respondents. After completion of the questionnaires, they are returned to the research organization.

Mixed-mode survey: A survey in which various modes of data collection are combined. Modes can be used concurrently (different groups are approached by different modes) or sequentially (nonrespondents of a mode are reapproached in a different mode).

Nondifferentiation: A form of satisficing that typically occurs if respondents have to answer a series of questions each with the same set of response options. They tend to select the same answer for all these questions irrespective of the question content.

Nonresponse: The phenomenon that elements in the selected sample, which are also eligible for the survey, do not provide the requested information or that the provided information is not usable.

Precise: An estimator with a small variance. The precision is a quantification of the reliability.

Primacy effect: The tendency of respondents have a preference for the to pick an answer early in the list of answers of a closed question. This typically happens in face-to-face and telephone surveys.

Probability sampling: A form of sampling where selection of elements is a random process. Each element must have a positive and known probability of selection.

Random digit dialing (RDD): A form of sample selection for a telephone survey where random telephone numbers are generated by some kind of computer algorithm.

Recency effect: The phenomenon that respondent have a preference for the last option in the list of answers of a closed question. This typically happens in mail and web surveys.

Reliable: An estimator that would result in (approximately) the same estimates if the survey is repeated.

Response order effect: The tendency that the answer selected by the respondents depends on its location in the list of answer options. Primacy and recency effects are special cases.

Satisficing: The phenomenon that respondents do not do all they can to provide a correct answer. Instead they attempt to give a satisfactory answer with minimal effort.

Socially desirable answer: The phenomenon that respondents do not give the true answer but an answer that is more socially desirable.

Status quo endorsement: The tendency to answer that everything should be kept the same. This typically occurs in opinion questions about change.

Straightlining: The tendency that respondents give the same answer to all single questions in a matrix question. They simply check all answer options in the same column.

Telephone interviewing: A form of interviewing in which interviewers call selected persons by telephone. If contact is made with the proper person, and this person wants to cooperate, the interview is started and conducted over the telephone.

Undercoverage: The sampling frame does not cover completely the target population of the survey. There are persons in the population who do not appear in the sampling frame. They will never be selected in the sample.

Valid: An estimator that estimates what the estimator is intended to estimate.

Web survey: A survey where respondents complete the questionnaires on the Internet.

EXERCISES

Exercise 5.1. Which mode of data collection is most expensive?

- a. A face-to-face survey.
- b. A telephone survey.
- c. A mail survey.
- d. A web survey.

Exercise 5.2. Should an error message be included in a web survey?

- a. Yes, because they always improve data quality.
- b. Yes, but careful attention should be paid to their design.
- c. No, because they increase item nonresponse.
- d. No, because they lead to socially desirable answers.

Exercise 5.3. Why is Random Digit Dialing (RDD) sometimes preferred for a telephone survey instead of random sampling from a telephone directory?

- a. RDD sampling provides more auxiliary information about nonrespondents.
- b. Response rates are lower in an RDD survey.
- c. RDD guarantees full coverage of the population.
- d. An RDD sample is less expensive than a sample from a directory.

Exercise 5.4. What does acquiescence mean?

- a. This is the tendency to not answer sensitive questions.
- b. This is the tendency to give more extreme answers.
- c. This is the tendency to disagree with what the interviewers say.
- d. This is the tendency to agree with what the interviewers say.

Exercise 5.5. How can the primacy effect in closed questions in web surveys be reduced?

- a. By randomizing the order of the answer options.
- b. By putting the answer options in the reverse order.
- c. By reducing the number of answer options.
- d. By increasing the number of answer options.

Exercise 5.6. Which of the following options is not an advantage of computer-assisted interviewing (CAI) as compared with traditional modes of data collection?

- a. Data quality is higher because of included checks.
- b. The software is in charge of routing through the questionnaire.
- c. CAI leads to higher response rates.
- d. Data can be processed more quickly.

Exercise 5.7. What is the effect of the mode of data collection on an opinion question with the possible answers: strongly disagree, disagree, neutral, agree, and strongly agree?

- a. More people will select the neutral category in mail and web surveys.
- b. More people will select the neutral category in CAPI and CATI surveys.
- c. More respondents will select the extreme options strongly disagree and strongly agree.
- d. The mode of data collection does not influence the answer patterns.

Exercise 5.8. In which situation should a web survey be implemented using the question-based approach (and not the form-based approach)?

- a. If the questionnaire contains many questions.
- b. If the respondent must be able to complete the questionnaire off-line.
- c. If the questionnaire contains matrix questions.
- d. If there are checks and routing instructions in the questionnaire.

Exercise 5.9. How should “don’t know” be treated in a web survey questionnaire?

- a. Do not present as it one of the answer options, and force respondents to answer.
- b. Do not present as it one of the answer options and do not force respondents to answer. If the question is skipped, record it as “don’t know.”
- c. Do not present as it one of the answer options and do not force respondents to answer. If the question is skipped, give the respondent two options: (1) answer the question or (2) record the answer as “don’t know.”
- d. Present it as an answer option, but it is less obvious elsewhere on the screen.

Exercise 5.10. What kind of problem may check-all-that-apply questions cause in web surveys?

- a. Acquiescence.
- b. Selecting an arbitrary answer.
- c. Straightlining.
- d. Giving a socially desirable answer.

REFERENCES

- Bethlehem, J. G., Cobben, F., & Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons, Hoboken NJ.
- Bethlehem, J. G. & Hofman, L.P.M.B. (2006), Blaise—Alive and Kicking for 20 Years. *Proceedings of the 10th Blaise, Users Meeting*, Statistics Netherlands, Voorburg/Heerlen, the Netherlands, pp. 61–88.
- Biemer, P. P. & Lyberg, L. E. (2003), *Introduction to Survey Quality*. John Wiley & Sons, New York.
- Couper, M. P. (2000), Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, pp. 464–494.
- Couper, M. P. (2008), *Designing Effective Web Surveys*. Cambridge University Press, Cambridge, U. K.
- Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls II, W. L., & O’Reilly, J. M. (eds.). (1998), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.
- De Leeuw, E. D. (2008), Choosing the Method of Data Collection. In: De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (eds.), *International Handbook of Survey Methodology*. Lawrence Erlbaum Associates, New York, pp. 113–135.
- DeRouvray, C. & Couper, M. P. (2002), Designing a Strategy for Reducing “No Opinion” Responses in Web-Base Surveys. *Social Science Computer Review*, 20, pp. 3–9.
- Dillman, D. A. (2007), *Mail and Internet Surveys: The Tailored Design Method*. John Wiley & Sons, Hoboken, NJ.
- Dillman, D. A. & Bowker, D. (2001), The Web Questionnaire Challenge to Survey Methodologists. In: Reips, U. D. & Bosnjak, M. (eds.), *Dimensions of Internet Science*, Pabst Science Publishers, Lengerich, Germany.

- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009), *Internet, Mail and Mixed-mode Surveys: The Tailored Design Method*. John Wiley & Sons, Hoboken, NJ.
- Dillman, D. A., Sinclair, M. D., & Clark, J. R. (1993), Effects of Questionnaire Length, Respondent-friendly Design and a Difficult Question on Response Rates for Occupant-addressed Census Mail Surveys. *Public Opinion Quarterly*, 57, pp. 289–304.
- Fricker, R. & Schonlau, M. (2002), Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Field Methods*, 15, pp. 347–367.
- Goyder, J. (1987), *The Silent Minority: Nonrespondents on Sample Surveys*. Westview Press, Boulder, CO.
- Groves, R. M. & Couper, M. P. (1998), *Nonresponse in Household Interview Surveys*. John Wiley & Sons, New York.
- Heerwegh, D. & Loosveldt, G. (2002), An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys. Paper presented at the International Conference on Improving Surveys, Copenhagen, Denmark, 2002.
- Hox, J. J. & De Leeuw, E. D. (1994), A Comparison of Nonresponse in Mail, Telephone and Face-to-face Surveys. Applying Multilevel Modeling to Meta-analysis. *Quality & Quantity*, 28, pp. 329–344.
- Krosnick, J. A. (1991), Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5, pp. 213–236.
- Kuusela, V., Vehovar, V., & Callegaro, M. (2006), Mobile phones—Influence on Telephone Surveys. Paper presented at the Second International Conference on Telephone Survey Methodology, Florida.
- Scherpenzeel, A. (2008), An Online Panel as a Platform for Multi-disciplinary Research. In: Stoop, I. & Wittenberg, M. (eds.), *Access Panels and Online Research, Panacea or Pitfall?* Aksant, Amsterdam, the Netherlands, pp. 101–106.
- Schuman, H. & Presser, S. (1981), *Questions and Answers in Attitude Surveys*. Academic Press, New York.
- Sudman, S. & Bradburn, N. (1982), *Asking Questions: A Practical Guide to Questionnaire Design*. Jossey-Bass, San Francisco, CA.
- Vis-Visschers, R., Arends-Tóth, J., Giesen, D., & Meertens, V. (2008), *Het Aanbieden van 'Weet niet' en Toelichtingen in een Vragenlijst*. Report DMH-2008-02-21-RVCS, Statistics Netherlands, Heerlen, the Netherlands.
- Wagner, J. (2008), *Adaptive Survey Design to Reduce Nonresponse Bias*. Doctoral Dissertation, University of Michigan, Ann Arbor.

Designing a Web Survey Questionnaire

6.1 Introduction

A web survey is a survey in which data are collected using the World Wide Web. As for all surveys, the aim of a web survey is to investigate a well-defined population. Such populations consist of concrete elements, such as individuals, households, or companies. It is typical for a survey that information is collected by means of asking questions of the representatives of the elements in the population. To ask questions in a uniform and consistent way, a questionnaire is used. There are various ways in which a questionnaire can be offered on the Internet:

- Poppers can be used to direct respondents to the questionnaire while they are visiting another site. This approach is particularly useful when the objectives of the survey relate to the website being visited, such as evaluating the website.
- E-mails are sent to people in a panel, a mailing list of customers, or other people who might qualify for the survey. The e-mail contains a link that directs them to the web survey questionnaire.
- Respondents can be directed to the website after a recruitment interview, either by telephone or face-to-face.

Popup surveys are not considered in this chapter. Such surveys are usually conducted for simple evaluation purposes and consist of only a few straightforward questions. Instead focus is on complex surveys, where the web survey questionnaire may be directed to individuals, households, or businesses.

To make advances in the social sciences and to make informed decisions for public policy and businesses, a high-quality data collection process is essential for capturing representative information. The design of the questionnaire is a crucial factor in the survey process, as this determines the quality of the collected data. In all data collection modes, the questionnaire is important; yet it can be argued that self-administered web or e-mail surveys rely even more heavily on the quality of the questionnaire.

It should be noted that designing a web survey is not the same as designing a website. In other words, web surveys are not the same as websites. Their goals are different, and their structures should be different as well. The web design aspects of the questionnaire do not solve all the problems related to web surveys. However, a well-designed, web-based survey questionnaire can help in reducing nonsampling errors, such as measurement errors and nonresponse. A badly designed web questionnaire increases such errors and therefore reduces data quality.

Web surveys are self-administered. As a result, they are similar to mail or fax surveys. In terms of data collection, the major differences between web surveys and other forms of data collection are the same as between self-completion (mail or fax surveys) and interviewer-assisted surveys (face-to-face and telephone surveys). It is, however, interesting to note that some literature has found that differences in the mode of data collection do not always imply differences in survey results. For instance, Cobanoglu, Warde, and Moreno (2001) have shown that mean scores for data collection via a web-based questionnaire are the same as for other self-completion methods, such as mail and fax surveys. More recently, Biffignandi and Manzoni (2011), using an experimental design comparing paper and web surveys, found no difference in evaluation scores relative to the data collection mode.

Even if the completion mode (self-administered versus interviewer-assisted) potentially has no effect on survey results, many issues should be taken into account in designing web survey questionnaires capable of capturing information correctly. Only then can neutrality and objectivity be preserved.

A point to be taken into account in choosing the mode of data collection is that self-administered questionnaires have particular advantages. One is that a longer list of answer options can be offered if the questionnaire is printed on paper or displayed on the computer screen. This is particularly true for web surveys. A high level of detail is difficult to obtain for other modes of data collection. Furthermore, a self-administered survey is more effective for addressing sensitive issues (such as medical matters or drug use). Answers to web survey questions suffer less from social desirability bias as respondents answer more truthfully. This means that web survey data on “threatening” issues, where respondents may feel a need to appear socially acceptable, are likely to represent much better how the survey population really feels.

Some surveys are carried out regularly. Examples are periodical surveys on prices, production, or international trade conducted by national statistical institutes. The transition of such surveys from a traditional paper mode to the web mode should take into account the fact that when the visual layout of the questions is not consistent with past experience and expectations, respondents may perceive (or even effectively undergo) a greater response burden and confusion. In

transition situations (and even when a mixed-mode approach is adopted), the key factor is to determine whether a questionnaire design should be preferred that fully takes advantage of the possibilities offered by the Internet or that a more “plain” questionnaire should be used that is similar to previous questionnaires.

When constructing a web questionnaire, many design principles of paper questionnaires can be applied. Actually, the basic principles for paper questionnaire design can be copied to the web. Examples are principles like using short questions, avoiding combined questions, and avoiding (double negative) questions. Nevertheless, data collection with on-line questionnaires is still a relatively new methodology. Research has been carried out and is still in progress to retest the principles of traditional paper questionnaires in the context of Internet research and to identify (and solve) new problems. It should be noted that the design of a web survey questionnaire is more complex than the design of a paper questionnaire. Web surveys allows for a wide range of textual options, format, and sophisticated graphics, none of which are usually attainable with e-mail surveys. Tools and procedures are available that allow for improving quality and simplifying compilation. For example, web surveys provide additional formats and response control such as preventing multiple answers when only one is called for, and links that provide the respondents with direct reference to definitions or examples at multiple points in the survey. Moreover, the questionnaire may include route instructions that see to it that respondents only answer relevant questions, whereas irrelevant questions are skipped.

Another advantage of using the Internet for data collection is that it provides a lot of extra information (so-called *paradata*) about the questionnaire completion process, such as completion time, the number of accesses to the questionnaire website, the number of clicks, and completion patterns. Moreover, data about the respondents can already be imported into the questionnaire before they start answering the questions. Examples are the values of stratification variables (type and size of the company) and the variables that are included in the sampling frame (address and household composition).

In conclusion web questionnaire design provides new insights into certain traditional basic principles of questionnaire design and it draws attention to new methodological issues. In this chapter, the focus is on design issues that are specific to web survey questionnaires. It gives an overview of the various possible questions types and shows how these question types should be formatted. Special attention is paid to handling “don’t know” in web surveys, taking into account the effects it can have on the answers to the questions.

6.2 Theory

6.2.1 THE ROAD MAP TOWARD A WEB QUESTIONNAIRE

Adequate *questionnaire construction* is critical for the success of a survey. Inappropriately formatted questions, incorrect ordering of questions, incorrect answer scales, or a bad questionnaire layout can make the survey results meaningless, as they may not accurately reflect the views and opinions of the

participants. Before entering into the detailed problems related to questionnaire design, some important general guidelines are mentioned. These guidelines help to obtain a good questionnaire design and to complete the web survey successfully.

6.2.1.1 Conduct a Pretest. Before going into the field, the questionnaire should be tested on a small group of respondents from the target population. This helps to check whether the survey accurately captures the intended information.

6.2.1.2 Pay Attention to the Way the Survey is Presented. The way the potential respondents are approached and the survey is presented to them are vital to obtaining their participation. Failure to carry out a proper contact approach may easily result in refusal to respond.

6.2.1.3 Include Instructions. Wherever needed, question-specific instructions should be incorporated into the survey instrument. Avoid offering instructions on separate sheets, in booklets, or on web pages. If instructions are included in a question in the paper version of the questionnaire, they should also appear in the electronic version of the questionnaire. Help facilities can be more powerful in web surveys than in interviewer-assisted surveys. A first advantage is that the help information is always presented in the same, consistent way. It does not depend on the interviewers. A second advantage is that web survey respondents need not ask a person for help. They can just click on a button or link to open a help window. Respondents in an interviewer-assisted survey may be reluctant to ask the interviewer for clarification because this would mean admitting ignorance. Moreover they also may feel embarrassed by asking for clarification about everyday concepts, although the terms may be used in atypical or ambiguous ways. It seems unlikely that web survey respondents will be embarrassed to ask a computer for more information. Nevertheless, the literature on social presence suggests computers do trigger similar self-presentation concerns and so could potentially prevent requests for clarification. See the study by Tourangeau, Couper, and Steiger (2003) for a discussion on social presence in web surveys.

6.2.1.4 Pay Attention to Technical Aspects. On the one hand, the limitations of the available hardware and software may restrict what can be done with a web survey, at both the design stage and the data collection stage. On the other hand, powerful hardware and software may open new possibilities.

6.2.1.4.1 Design stage aspects. Nowadays many different software and hardware environments are available for the survey researcher. It is possible to design advanced and sophisticated survey instruments. However, the limited computer facilities of the respondents prohibit running the survey software on their machines. See Dillman (2009) for more details. If the questionnaire becomes inaccessible, or difficult to complete, for some groups of respondents,

response rates will drop. And incorrect visualization of questions will result in measurement errors and, thus, in biased results.

Another important aspect is that the current stage of the technology makes it possible to control access to the questionnaire with a unique identification code for each sample element. This approach is recommended because it guarantees that only selected elements can obtain access to the questionnaire. The unique code can also help to prevent someone from completing a questionnaire more than once. The unique identification code should be included in the e-mail or letter that is sent to the selected survey elements. Such codes are often part of the link to the survey website. Using code access to the questionnaire usually has no effect on survey participation.

■ EXAMPLE 6.1 The library study

The library of the faculty of economics of Bergamo University conducted a customer satisfaction survey in 2007. The target population of the survey consisted of all students of the faculty. Because all students get an e-mail address when they enroll at the university, it was decided to use a web survey.

The objective of the survey was the evaluation of the library services (opening times, books availability, interlibrary delivery, room space, and equipment such as computers). The questionnaire also contained some questions about the use of the Internet, particularly about the use of e-mail.

Students were contacted by e-mail. The e-mail message contained a link to the questionnaire and a unique identification code. The students needed to enter the code to get access to the questionnaire. Figure 6.1 contains the access screen.

The image shows a web interface for Bergamo University's library services. At the top left is the university's circular logo. To its right, the text 'Università di Bergamo' is displayed in a large, bold, black font, with 'Servizi bibliotecari' in a smaller font below it. Below the header is a search bar. The main form area contains a 'Username' label followed by a text input field, and a 'Password' label followed by a text input field. Below these fields is a button labeled 'Invia'.

FIGURE 6.1 Using a unique code to get access to the questionnaire

A total of 1,273 students completed the questionnaire. A simple progress indicator was used to inform students about their progress in answering the questions. It was a simple, non-graphic message showing the question number and the total number of questions. Figure 6.2 shows this indicator for the question about checking e-mail.

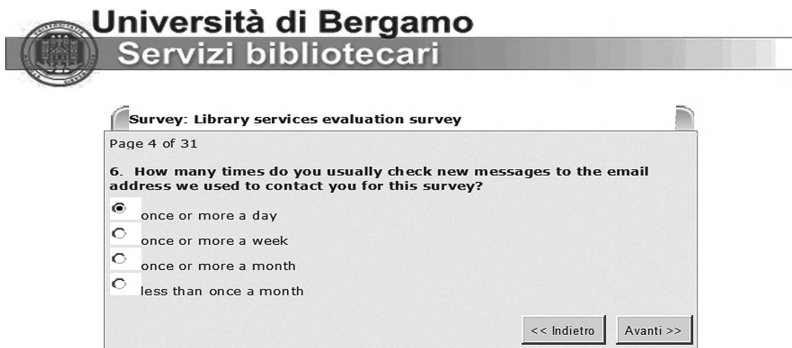


FIGURE 6.2 A question screen with a progress indicator

The results of the question in Figure 6.2 gave some insight into the e-mail checking behavior of the students. Table 6.1 summarizes these results.

TABLE 6.1 Frequency of checking e-mail

Checking	Number	Percentage
Once or more every a day	533	41.9%
Once or more every week	552	43.3%
Once or more every month	145	11.4%
Less than once a month	43	3.4%
Total	1,273	100.0%

It is clear that many survey respondents check their e-mail fairly often. It should be noted that students who never check their e-mail, or only will do so now and then, will not be well represented in the survey.

It is important to keep respondents motivated while they are in the process of completing a self-administered questionnaire. A feature that may help to accomplish this in a web survey is to include *progress indicators*. These are textual or graphical devices that give feedback about the progress of the respondent in the questionnaire.

EXAMPLE 6.2 A progress indicator

Figure 6.3 shows an example of a progress indicator that was used in a Radio Listening Survey. This indicator took the form of a black bar that increases in length as the respondent progresses through the questionnaire.

The image shows a screenshot of a web survey interface. At the top, there is a header area with a radio icon on the left, the title "Local Radio Listening Survey 2010" in the center, and a "Progress:" label on the right with a black progress bar. Below the header, the main content area contains question 4: "4. Did you listen to your local radio during the last seven days?". There are two radio buttons: "Yes" and "No". At the bottom of the interface, there are two buttons: "Previous" and "Next".

FIGURE 6.3 The progress indicator of the Radio Listening Survey

The questionnaire of the Radio Listening Survey was fairly straightforward, with a limited number of questions and without complex routing. This makes a progress indicator ideal as a tool for motivating respondents to complete the questionnaire.

Progress indicators are widely used in web surveys. The prevailing view among survey designers seems to be that this information is appreciated by respondents. They will be more likely to continue answering questions if they have some sense of where they are in the process. A hypothesis related to this device is that giving respondents information about progress increases their involvement in the survey. The advantage is that they will be more likely to complete the task if they see they are making progress. However, a negative effect is also possible: if progress is slow, respondents may become discouraged, resulting in breaking off their task. Some studies have been carried out to test hypotheses concerning the effect of progress indicators on survey participation. The results suggest that when progress seems to surpass the expectations of the respondent, feedback can significantly improve completion rates. When progress seems to lag behind what they expect, feedback reduces engagement and completion rates. Progress indicators are useful in surveys related to individuals or households, and even in simple surveys. In complex surveys and in business surveys (especially in surveys requesting business data, such as financial data), progress indicators are less useful because data must be collected before completing the questionnaire. Moreover, different people may be involved in completing the questionnaire.

6.2.1.4.2 Data collection stage aspects. Progress in the development of Internet technology has greatly improved the quality and possibilities of online

access of households and businesses participating in surveys. However, the survey designer has to be aware that one of the major problems of web surveys is that it is never clear how the questionnaire will appear on the computer of the respondents. Their browsers may not be able to display certain pictures or symbols. A low screen resolution may require a respondent to scroll to make different parts of the questionnaire visible. A low bandwidth may slow navigation through the questionnaire. There are many different computer configurations, including the platform or operating system. There are differences in hardware (desktop, laptop, and mobile phone), screen size, Internet browser, versions of the same Internet browser, processor speed, available memory, and so on. All these difference may cause problems for large, complex, sophisticated web survey questionnaires. However, if the questionnaire is kept simple, one may expect that respondents will not encounter technical problems, whatever the environment they use. Nevertheless, survey designers should be aware of potential problems and, therefore, should test their survey instruments in many different environments.

6.2.1.5 Make it Attractive to Complete the Questionnaire. Web surveys are self-administered surveys. There are no interviewers to motivate respondents to complete the questionnaire. Respondents will be motivated if the questionnaire is attractive. The decision to participate is to a large extent determined by respondent motivation and satisfaction. It is thus important that the design of the web questionnaire provides the respondents with as much pleasure and satisfaction as possible to increase his or her interest. This helps in convincing respondents to answer all survey questions and to complete the questionnaire. Item nonresponse and partial nonresponse is minimized. The so-called *interactional information system approach* (a questionnaire not only collects data but also provides respondents with relevant information) may be useful in generating and maintaining interest in the survey, thereby increasing response rates. This approach is in line with experimental studies that highlight the importance of placing interest-related questions early in the questionnaire as this would prevent attrition from lack of interest. See the study by Shropshire, Hawdon and Witte (2009).

6.2.1.6 Apply Questionnaire Design Principles. As previously stated, web questionnaire design can partly be based on the principles for paper questionnaire construction. These criteria are a valuable starting point, but also two specific aspects of web surveys should be taken into account. First, it should be born in mind that even small differences in question wording or of a stimulus embedded in the question display may greatly affect the answers given. Second, it should be noted that checks can be included in a web survey questionnaire. This feature is also available in computer-assisted surveys modes like computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI) but not in a paper-based approach. Every change in web survey questionnaire may make it easier or harder to complete it, or it may convey different messages about what kind of information is expected. In conclusion, this may have a serious impact on the collected data.

The recipe for designing a web survey questionnaire in a scientifically sound way involves the following specific ingredients:

- a. The proper technology (hardware, platform, software, and so on).
- b. Well-organized survey management, both in the design phase as well as in the data collection phase.
- c. Availability of skills for defining question and response format, being aware of the possible impact of the offered response choices and of the possible ways in which the questionnaire can be visualized and customized using the available Internet technology.
- d. Anticipating during the survey the collection of other data (paradata and auxiliary information) that may be linked to the interview. This aspect is typical of web surveys and, if used appropriately, may improve participation and the quality of the survey process.

Technical aspects will not be discussed in the subsequent chapters. Attention also will not be paid to organizational issues, as these issues depend very much on the context in which the survey is being conducted and on the objectives of the survey.

The focus is on various aspects of questionnaire design (point c), with special attention on the issues that are typical for web questionnaires. In addition, section 6.2.5 introduces the concept of paradata. An attempt is made to answer the question of how auxiliary information and data from other linked sources can be used during the survey process (point d).

Web questionnaires started as simple electronic analogues of paper questionnaires. Over time, web questionnaires have evolved. There is ample research about applying new technological innovations. Clear criteria have been formulated, but some issues are still under investigation. Criteria will be discussed that are related to the following themes:

- Formatting the text of the questions
- Formatting the answers of the questions (closed questions and open questions)
- Paradata collection.

6.2.2 THE LANGUAGE OF QUESTIONS

The design of a question requires decisions to be made with respect to its format and layout. The format of the question includes aspects like the wording of the question and the type of answer (the answer format) that must be given. The layout of the question includes aspects like font type, font size, and use of colors. All these aspects are called the language of questions.

Two very basic types of survey questions can be distinguished: open questions and closed questions. A *closed question* offers a (not very long) list of answer options. The respondent has to select one option. A closed question may be used if there are a limited number of possible answers and the researcher wants to avoid the respondent to overlook an answer. Such a question also may be used if

the researcher wants to define his or her own classification of possible answers. A closed question is a tool to measure a categorical (qualitative) variable. Figure 6.4 shows an example of a closed question (from a radio listening survey).

A special case of a closed question is a *check-all-that-apply* question. This is an open question for which more than one answer option may be selected. Another special case is a closed question with ordered categories. An example is an opinion question with the answer options *Strongly agree*, *Agree*, *Neutral*, *Disagree*, and *Strongly disagree*.

The other basic question type is the open question. For such a question, the respondents may enter any answer they like. There are no restrictions other than the length of the answer. Open questions should be used where there are a very large number of possible different answer options, where the researcher does not know all possible answer options, or where one requires the respondents to give the answer in their own words. Figure 6.5 shows an example of an open question.

If the list of answer options is very long, if the complete list of answer options is unknown, or if there may be unanticipated answers, one may decide to use a question type that is a mixture of a closed and an open question. The main options are listed, and all possible answers can be dealt with by selection the option “Other, please specify” and entering the answer if it were an open question. See Figure 6.6 for an example.

The question language of a survey includes the wording of the text of the questions, instructions for answering them, and visual aspects such as font size, font type, color, layout, symbols, images, animation, and other graphics. Couper

In the last seven days, what type of music did you listen to most?

- Chart / Top 40
- Dance
- Rock
- R & B
- Hip-hop
- Country
- Folk
- Easy listening
- Jazz
- Classical
- New age
- Other music

FIGURE 6.4 A closed question

2. In the last seven days, what type of music did you listen to most?

FIGURE 6.5 An open question

3. In the last seven days, what type of music did you listen to most?

Chart / Top 40
 Dance
 Rock
 R & B
 Hip-hop
 Country
 Folk
 Easy listening
 Jazz
 Classical
 Other, please specify:

FIGURE 6.6 A hybrid question

(2000 and 2008) includes visual language in the question language because it is intended to supplement the written language. In fact, the way the questionnaire is visualized can have a great impact on the way the questions are answered. Abundant nonfunctional use of graphical effects may draw attention away from the text, alter the meaning of words, and complicate easy and straightforward understanding of the questions. In summary, the various visualization aspects together affect the perception of the survey questionnaire and the response burden. A more detailed description of the *visual language* (as proposed by Redline and Dillman (1999) includes three different types of languages:

- *Graphic language.* This language consists of fonts, font sizes, and font enhancements (bold, italics, and underline), borders, and tables. When used in a functional way, it helps respondents to move their eyes across the page and to comprehend the questionnaire.
- *Symbolic language.* This language is sometimes used in questionnaires for navigation purposes. Symbols like arrows help in leading respondents through the survey questions in the proper order and in answering the proper questions.
- *Numeric language.* This language is used in questionnaires for numbering questions and sometimes for numbering answer options.

Visual language is an auxiliary language. It may help to make the questionnaire more attractive. However, the questionnaire designer should be aware that this may affect the way in which respondents may interpret questions. For instance, if graphics or pictures are used, bear in mind that respondents tend to interpret questions in the context of those graphics and pictures.

EXAMPLE 6.3 Use of Pictures in Questions

Couper et al. (2004a) describe an experiment in which questions were offered in three different ways: (1) without a picture, (2) with a picture indicating low-frequency behavior, and (3) with a picture indicating

high-frequency behavior. One of these questions asked how many times one went out to eat in the last month. The low-frequency behavior picture was one of an intimate, exclusive restaurant and the high-frequency behavior picture was one of a fast food restaurant.

When the low-frequency picture was included in the questionnaire, the frequency of going out to eat was 9.9%. When no picture was included, the percentage was higher: 12.0%. And when the high-frequency picture was included, the percentage was 13.6%. Apparently, use of pictures partly determines what people mean by “going out to eat.”

Many experiments with traditional surveys have shown that minor changes in question wording may substantially change responses. This also applies to web surveys. However, not only the wording of questions is important in web surveys, but also the presentation of the questions. Minor changes in presentation may seriously affect the answers given.

6.2.3 ANSWERS TYPES (RESPONSE FORMAT)

The two basic question types were already mentioned: open questions (people answer the question in their own words) and closed questions (the answer is selected from a list of possible answer options). The answer options can take the form of an unordered list, as in Figure 6.4. The answer options can also take the form of an ordered list like a rating scale (*Strongly agree, Agree, Neutral, Disagree, and Strongly disagree*). Another frequently occurring example is a yes/no-question where the respondent has to select either *Yes* and *No*. It is also possible to allow the respondent to select more than one answer option. Then the closed question takes the form of a *check-all-that-apply question*.

There are various ways to visualize questions on the computer screen. It is important that the visual format corresponds to what is expected of the respondents. The HTML-language that is used to define web pages allows for the following constructs:

- Radio buttons
- Check boxes
- Drop-down boxes and list boxes
- Text boxes and text areas.

Radio buttons, check boxes, and drop-boxes are alternative solutions for closed questions. Text boxes and text areas are used for open questions. In the following sections, different answer types are described and discussed. It is noted again that the visual aspects of questionnaire design are very important. There is more to web questionnaire design than just mimicking the paper questionnaire on the computer screen. Small changes in the design and layout may have large consequences for the answers to the questions.

6.2.3.1 Radio Buttons. Radio buttons should be used for closed questions where respondents can select only one answer option. The answer options must be mutually exclusive, and together they must cover all possible answers.

Initially, no option is selected. This forces the respondents to select an option. They cannot continue to the next question without thinking about the current question. An answer is selected by clicking the radio button corresponding to the answer option. If the respondent clicks another answer option, the previously selected option is de-selected. Not more than one option can be selected.

To avoid confusion among respondents, radio buttons should always be used to implement closed questions with one answer. It is technically possible to use radio buttons also for check-all-that-apply questions, but this is not recommended.

One limitation on the use of radio buttons is that their size is not related to the size of the font attached to it. Radio buttons keep the same size irrespective of the font size. It is advised to use a font of approximately the same size as the radio button size.

The answer options should be presented in a logical order. Alphabetical ordering is not recommended because it creates problems for multilingual questionnaires. This would mean a different order for a different language.

A point of discussion is always whether to include answer options “Don’t know” or “Does not apply.” However, the danger of including “Don’t know” is that people may select it to avoid having to think about an answer. Yet, if people really do not know the answer, they must have the possibility to answer so. This is generally more a point of discussion of opinion questions than of factual questions.

EXAMPLE 6.4 Answering a Closed Question

A simple example shows what happens when a closed question is answered if answer options are denoted by radio buttons. The question is “What kind of product were you thinking about while filling in this questionnaire?”. There are five possible answers: “Electronics”, “Clothes/accessories”, “Watches”, “Cars”, and “Other products”. The respondents can only give one answer because they can only think of one product. Therefore, radio buttons are the appropriate format to define this question. The possible answers are mutually exclusive. The option “Other products” guarantees that a respondent always has an option that is appropriate.

What kind of product were you thinking about while filling in this questionnaire?

- Electronics
- Clothes / accessories
- Watches
- Cars
- Other products

FIGURE 6.7 A closed question in its initial state

Initially, the question looks the same as in Figure 6.7. No answer has yet been selected. This is done to prevent respondents from skipping the questionnaire (because it already has been answered). This would lead to bias toward the default answer option. Figure 6.8 shows the situation after the respondent clicked the radio button for the option “Cars.”

What kind of product were you thinking about while filling in this questionnaire?

- Electronics
- Clothes / accessories
- Watches
- Cars
- Other products

FIGURE 6.8 A closed question after selecting an answer

If the respondent decides that he or she made a mistake, and that “Electronics” is the proper answer, he or she can just click on the radio button corresponding to this option. The result is shown in Figure 6.9. The answer “Cars” has been de-selected automatically.

What kind of product were you thinking about while filling in this questionnaire?

- Electronics
- Clothes / accessories
- Watches
- Cars
- Other products

FIGURE 6.9 A closed question after selecting another answer

In these examples, there has been just a list of possible answer options. However, it is also possible to offer more complex structures. Figure 6.10 contains an example of nested options. Such a hierarchical structure should be avoided as it may cause confusion among respondents. It is preferred to keep all options at the same level.

Correct:

- Let Internet Explorer decide how pop-ups should be opened
- Always open pop-ups in a new window
- Always open pop-ups in a new tab

Incorrect:

- Let Internet Explorer decide how pop-ups should be opened
- Always open pop-ups:
 - in a new window
 - in a new tab

FIGURE 6.10 A closed question with nested answers

The Internet is a useful tool for socialization.

- Strongly disagree
- Disagree
- Neither agree or disagree
- Agree
- Strongly agree

FIGURE 6.11 A Likert scale question

A closed question can be used to measure a value on a scale. This very often takes the form of a *Likert scale*. When responding to a Likert scale question, respondents specify their level of agreement to a statement. The scale is named after its inventor, the psychologist Rensis Likert. Often five-ordered response options are used, although some researchers advocate using seven or nine levels. A typical Likert questions has the response options “*Strongly agree*,” “*Agree*,” “*Neither agree nor disagree*,” “*Disagree*,” and “*Strongly disagree*.” Figure 6.11 shows an example of a Likert scale question.

A Likert scale question measures an ordinal variable, i.e., the answer options have a natural ordering. It is even possible to assign numerical values to the response options (1 = strongly agree, 2 = agree, etc.), so that in fact a numerical variable is measured and for example mean scores can be computed.

A Likert scale question has the advantage over a simple yes/no- question in that it allows respondents to give a more differentiated answer than just yes or no. It even allows respondents to have no opinion at all. The other side of the coin is that respondents can select the neutral middle option to avoid having to give an opinion.

Use of Likert scales is not without problems. The description of the response categories must be as accurate as possible. All respondents must interpret the descriptions in the same way. This is not always easy to realize. What does “strongly agree” mean? What is the difference with “agree”? Likewise, respondents find it hard to distinguish “good” from “very good” and “very good” from “excellent.”

Sometimes closed questions ask for the frequency with which activities are carried out. Words like “often” or “sometimes” should be avoided as it is unclear what they mean. They could be understood differently by different respondents. Problems could increase even more in multilingual questionnaires if these words have to be translated in different languages. For a question asking about frequencies, a solution is to relate the activities to concrete time periods, like “every day,” “at least one a week,” and so on.

Figure 6.11 shows just one way to display a Likert scale question on the screen. Figure 6.12 shows another way to do it. The response categories are now placed horizontally. This may correspond better to how respondents visualize a scale.

It is also possible to replace the response category labels by numbers. For example, the radio buttons in Figure 6.12 could be numbered 1 to 5. If this is done, the question text must explain that 1 means “strongly disagree” and that 5 corresponds to “strongly agree, (or vice versa: 1 for “strongly agree” and 5 “for strongly disagree”).” Experience has shown that numbered scales are difficult to handle for people. For example, scales that are marked “1 to 5, with 5 being the

It is easy for me to remember how to perform tasks using spreadsheets

Strongly disagree
 Disagree
 Neither agree or disagree
 Agree
 Strongly agree

FIGURE 6.12 A Likert scale question with horizontal categories

How would you rate the quality of service provided by your Internet service provider?

1 Poor 2 Fair 3 Good 4 Very good 5 Excellent

How would you rate the quality of service provided by your Internet service provider?

1 Poor 2 3 4 5 Excellent

FIGURE 6.13 Likert scale questions with numbered and labeled categories

highest, require more cognitive efforts than scales with labels such as “poor” or “excellent.” Some studies (see, for instance, the study by Christian, Parsons, and Dillman, 2009) show that response times are longer for scales with numeric labels, but there are no differences in response patterns. Numbered response categories may help if they are combined with labels. Figure 6.13 contains two examples. In the first example, all categories are labeled, whereas in the second example, only the endpoints are labeled.

Respondents interpret the endpoint label for the low end of the scale as more negative or extreme when negative numbers are used. Research also indicates that scales with numeric labels produce results that are similar to scales without numeric labels. This suggests a hierarchy of features that respondents pay attention to, with text labels taking precedence over numerical labels and numerical labels taking precedence over purely visual cues, such as color. See also the study by Tourangeau, Couper, and Conrad (2007).

Some researchers prefer five-point scales (or seven-point scales) because they offer respondents a “neutral” middle point. Other researchers prefer an even number of response options (for example, a four-point scale) because they “force” people to select a negative or a positive answer.

EXAMPLE 6.5 Asking about the Use of Mobile Phones

Sometimes an extra response option is added to the scale indicating that the question does not apply. This is illustrated in Figure 6.14. A five-point Likert scale has been used. Note that all response options have been

numbered. Only the extreme options and the middle option have a text label.

Which features do you have on your mobile phone? Indicate for each feature you have how frequently you use it.

	1	2	3	4	5	6
	Not available	Do not use		Average use		Very frequent use
MP3	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Digital camera	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
PDA	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text messages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

FIGURE 6.14 Adding an extra option to a Likert scale question

Note that the first response option is not part of the Likert scale. It has been added for respondents that do not have the specific options in their mobile phone.

In fact, this is a matrix question containing four separate questions. To make it easier for the respondents to find the proper answer for each question, the rows have alternative background colors.

The treatment of “Don’t know” is always a point of discussion in surveys. The general rule is applied here to make a distinction between factual questions and opinion/attitudinal questions. In the case of a factual question, the respondents should know the answer. Therefore, “Don’t know” is not offered as one of the response options. The situation is different for opinion and attitudinal questions. If respondents are asked about their opinion about a specific issue, it is possible that they do not have an opinion. Therefore, “Don’t know” should be offered as a possible response option. There is a risk, however, that respondents select this option to avoid having to express their opinion. This is called satisficing.

If “Don’t know,” and possibly “No opinion,” are included in the list of response options, the question arises at which position to put in the list. Tourangeau, Couper and Conrad (2004) have conducted experiments where these nonsubstantive options were at the bottom of the list and visually separated from the other options by means of a dividing line (see Figure 6.15). The result was that more respondents selected “Don’t know” because their attention was drawn to it. Without the dividing line, they observed an upward shift in the answers, as many respondents tended to select an answer in the visual middle of the list and they considered “Don’t know” and “No opinion” part of the rating scale.

DeRouvy and Couper (2002) experimented with questions where the “Don’t know” option was displayed in a smaller and lighter font so that its visual prominence was reduced. This did not affect the number of respondents selecting this option.

Think of how much the government is doing to make sure women have the same job opportunities as men. Would you say the government is doing too much, about the right amount, or too little about this?

- Far too much
 - Too much
 - About the right amount
 - Too little
 - Far too little
-
- Don't know
 - No opinion

FIGURE 6.15 Including nonsubstantive options

Not offering “Don’t know” in opinion questions is not an option. This would imply that respondents always have to provide a substantive answer, even if they do not know the answer. According to Couper (2008), this violates the norm of voluntary participation. Respondents should have the possibility of not answering a question. Forcing respondents to answer may frustrate respondents resulting in break-off. Also Dillman (2007) strongly recommends not forcing respondents to answer. He warns about detrimental effects on respondent motivation, on data quality, and on the risk of break-off.

Figure 6.14 contains an example of a matrix question. A *matrix question* (sometimes also called a *grid question*) brings together several questions with the same set of answer options. Matrix questions seem to have some advantages. A matrix question takes less space on the questionnaire form than a set of single questions and it provides respondents with more oversight. Therefore it can reduce the time it takes to answer questions. Couper, Traugott, and Lamias (2001) indeed found that a matrix question takes less time to answer than a set of single questions.

However, according to Dillman, Smyth, and Christian (2009), answering a matrix question is a complex cognitive task. It is not always easy for respondents to link a single question in a row to the proper answer in the column. Moreover, respondents can navigate through the matrix in several ways, row-wise, column-wise, or a mixture of the two. This increases the risk of missing answers to questions, resulting in a higher item nonresponse. Shading the rows of the matrix, like in Figure 6.14, may help to reduce this problem.

Dillman (2009) recommends limiting the use of matrix questions as much as possible. If they are used, they should not be too wide nor too long. Preferably, the whole matrix should fit on a single screen. This is not so easy to realize as different respondents may have different screen resolutions on their computer screens. If respondents have to scroll, either horizontally other vertically, they may easily get confused, leading to wrong or missed answers.

Several authors (see, for example, Krosnick (1991) and Tourangeau et al. (2004)) express concern about a phenomenon that is sometimes called *straight-lining*. Respondents give the same answer to all questions in the matrix. They simply check all radio buttons in the same column. Often this is the column corresponding to the middle response option. For example, respondents could

make it easy for themselves by selecting all radio buttons in column 4 (average use) for the question in Figure 6.14.

If the response options of a closed question form an ordinal scale, this scale must be presented in a logical order. This also applies to matrix questions, where the scale is displayed in a horizontal fashion. The leftmost option must be either the most positive one or the most negative one. Tourangeau et al. (2004) found that response times were considerably longer if the response options were not ordered, or if the midpoint (e.g., no opinion) was the last option. Respondents tend to use the visual midpoint in a response scale as an anchor or reference point for judging their own position. They get confused when the visual midpoint does not coincide with the midpoint of the Likert scale.

6.2.3.2 Drop-Down Boxes. A different way to select one option from a list of answer options is using a *drop-down box*. This device may be considered if the list of answer options is very long. Radio buttons are less effective for such lists. They would require a lot of space, and the respondents lack oversight. Figure 6.16 shows an example of a drop-down box in its initial state. The question asks for the country of birth, and the list of countries is invisible.

To open the list, the respondent has to click on “Select a country.” If this list is very long, it only becomes partially visible. See Figure 6.17 for an example. It depends on the browser used how long this list is. For example, 20 items are shown in Firefox 3.6 and 30 items in Internet Explorer 8. If the list is longer,

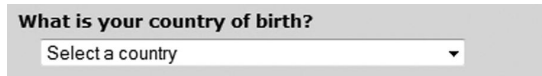


FIGURE 6.16 A drop-down box in its initial state



FIGURE 6.17 A drop-down box after opening the list

scroll bars are provided to make other items visible. The respondent selects an answer by clicking on it in the list.

Drop-down boxes have several disadvantages. In the first place, it requires more actions to select an answer (as compared with radio buttons). Three actions have to be performed: clicking the box, scrolling to the right answer, and clicking this answer. In the second place, there can be serious primacy effects if only part of the list is displayed: respondents tend to select an option in the visible part of the list. In the third place, it is unclear how much space the question requires on the screen.

It is possible to modify the behavior of the drop-down box, so that it always shows a fixed number of items in the list. This has the advantage that it is clear how much space the question requires. However, also here there may be serious primacy effects; see Couper et al. (2004b).

Note that in fact “Select a country” is the first item in the list. It could be removed from the list, but then the first country (Afghanistan) would be visible in the initial state. This could be the cause of a primacy effect. The text “Select a country” sees to it that all items in the list are treated equally, and it provides a clue as to what the respondents should do.

The general advice is that, where possible, radio buttons should be preferred. The advantages and disadvantages of various answer formats of closed questions are also discussed by Couper (1999), Heerwegh and Loosveldt (2002), and Dillman (2007).

6.2.3.3 Check Boxes. Check boxes are used for answering closed questions for which more than one answer is allowed. Such a question is also called a *check-all-that-apply question*. Figure 6.18 shows an example of such a question. A check box is shown on the screen as a square box. It can have two states: an empty white box indicates the option is not selected and a tick mark indicates the option is elected. The state of the check box can be changed by clicking on it with the mouse.

Check boxes permit the user to select multiple answers from the set of answer options. For example, two options have been selected in the check box in Figure 6.18. All answer options must be mutually exclusive and together cover all possible answers. Note that the option “Other” in Figure 6.18 guarantees that always an answer can be selected.

For specific questions, it may be possible that none of the answer options apply to the respondent. Of course, that can be dealt with by checking no option. However, this makes it easy for a satisficing respondent to just skip the question without attempting to answer it. This can be avoided by including the option “None of the above” and forcing the respondent to at least check one answer. Of course, the questionnaire software must prevent selecting “None of the above” in combination with another answer.

If the list of answer options of a check-all-that-apply question is long, selecting the proper answer may mean a lot of work for respondents. Instead of checking all relevant answers, they may just check some arbitrary answers and stop when they think they have checked enough answers. Moreover, satisficing

How did you follow the news about the 2010 election campaign? Please check all that apply.

Watching TV
 Listening to the radio
 Reading newspapers
 Looking on the Internet
 Other

FIGURE 6.18 A check-all-that-apply question

In the last seven days, what type of music did you listen to most?

Yes No

Watching TV
 Listening to the radio
 Reading newspapers
 Looking on the Internet
 Other

FIGURE 6.19 A check-all-that-apply question with radio buttons

respondents tend to read only the first part of the list, and not the complete list. This causes a bias toward answers in the first part of the list. A solution to this problem may be replacing check boxes by radio buttons, like in Figure 6.19. Smyth, Dillman, Christian, and Stern (2006) have shown that the format in Figure 6.19 leads to more selected options and that respondents take more time to answer the questions. This is an indication that the format as in figure 6.18 may cause satisficing.

If there is a series of check-all-that-apply questions all with the same set of possible answers, one may consider combining them in a matrix question (grid question). It was already mentioned that a matrix question has some advantages. It takes less space on the questionnaire form than a set of single questions, and it provides respondents with more oversight. Therefore, it can reduce the time it takes to answer questions. However, answering a matrix question is a complex cognitive task. Respondents can navigate through the matrix in several ways, row-wise, column-wise, or a mixture of the two. This increases the risk of missing answers to questions, resulting in a higher item nonresponse. Dillman (2009) recommends limiting the use of matrix questions as much as possible. If they are used, they should not be too wide to too long. Preferably the whole matrix should fit on a single screen. This is not so easy to realize as different respondents may have set different screen resolutions on their computer screens. If respondents have to scroll, either horizontally other vertically, they may easily get confused, leading to wrong or missed answers.

6.2.3.4 Text Boxes and Text Areas. Text boxes and text areas are used in web survey questionnaires to record answers to open questions. Open questions

have the advantage that the respondents can reply to a question completely in their own words. No answers are suggested. There are also several disadvantages. It takes more time to answer such questions. Processing the answers is much more complex. The answers have to be analyzed, coded, and put into manageable categories. This is time consuming and error prone.

As there is not much guidance as to what is expected of the respondents, they may easily forget things in their answers, put the focus differently, or have difficulty putting their thoughts into words. This may be the cause of measurement errors. If the question lacks focus, the answer will also lack focus. For example, the question “When did you move to this town?” could elicit responses like “when I got married,” “last year,” “when I started my study at the university,” “when I bought a house,” and “last year.”

Open questions must be treated with considerable caution. Nevertheless, in some specific research contexts, they may offer several advantages. No other solution may be available to retrieve such information. Respondents answer open questions in their own words, and therefore, they are not influenced by any specific alternatives suggested by the interviewer. If there are no clear ideas as to which issues may be the most important to the respondents, open questions are required. They may reveal findings that were not originally anticipated.

An open question can be implemented by means of a *text box*. An example is the first question in Figure 6.20. It provides space for just one line of text. The length of the text box can be specified by the questionnaire designer. This implementation of an open question should typically be used in situations where short answers are expected.

The second question in Figure 6.20 is an implementation of an open question by means of a *text area*. This provides for an answer space consisting of several lines of text. The width and height of the area can be specified by the questionnaire designer. The scrollbars even suggest that the text can be longer than the size of the box. Research shows that indeed the second format leads to longer answers than the first format.

1. Why did you choose this university?

2. Your answer to this question is very important for understanding what brings people to this university. Why did you choose this university?

FIGURE 6.20 Formatting an open question

How did you follow the news about the 2010 election campaign? Please check all that apply.

Watching TV

Listening to the radio

Reading newspapers

Looking on the Internet

Other, please specify:

FIGURE 6.21 Combining an open and a closed question

If the researcher has good prior knowledge of the question topic, and can generate a set of likely (but not exhaustive) response options, an alternative approach could be a hybrid question. This is a combination of a closed and an open question (see Figure 6.21). All known options are listed and a different answer can be taken care of by the option “Other, please specify.”

6.2.4 BASIC CONCEPTS OF VISUALIZATION

6.2.4.1 Answer Spaces. The *answer space* is the area on the computer screen where respondents type their answers to the (open) questions. Answer space requires careful design. They should be easy to locate by standing out visually, and it should also be clear to respondents what is expected of them. In specific situations, it may be helpful to include extra instructions.

Research results show that increasing the size of the answer box has little effect on early respondents of the survey but substantially improves the quality of the answers of late respondents. Including instructions and explanations improves the data quality for both early and late respondents.

A consistent questionnaire design is important. If the same type of question is asked, the same type of answer space must be used. This reduces the cognitive task of the respondents. Research shows that respondents use all available information to help them to formulate an answer. That is, in addition to the questions themselves, they use the information provided by the response categories and the answer space. See Sudman et al. (1996 and 1973). Inconsistencies in the web questionnaire will confuse them and may lead to lower data quality or nonresponse.

For questionnaires in general, it is advised to surround answer spaces by a frame in a contrasting color. This clearly separates the answer spaces from the rest of the questionnaire page. This is particularly important in the following situations:

- The questionnaire has a light background color so there is not enough contrast to distinguish white answer spaces from the rest of the page.
- The questionnaire is subject to key-from-image (KFI) processing. KFI involves separating the questionnaire forms into single pages, scanning the pages on high-speed scanners, and storing a digital image of each page in a central repository. Data entry operators then key the information from the images into the census system.

- The route through the questionnaire depends on the answer to a previous question. It is particularly important that such a filter question is answered correctly. See also Example 6.6.

EXAMPLE 6.6 Adding a frame to an answer space

Question 25 of the R&D survey of the Italian Statistical Institute (ISTAT) asks whether the company had any branch or subsidiary abroad performing research and development. In addition, firms having such a branch or subsidiary have to provide information about R&D expenditures and personnel. Thus, a different answer scheme is provided related to the answer to the main question. So the filter question in Figure 6.22 (with as possible answers yes and no) determines whether the subsequent matrix question has to be answered.

Istat.it Rilevazione statistica sulla ricerca e sviluppo nelle imprese 2008

RSI

L'indagine Istruzioni Registrazione Accesso al modello Password Cor

Company Name:PRE-MET S.P.A. | Company Code:000015272

Indicare obbligatoriamente una risposta al quesito 1

25 - THE COMPANY HAD, IN 2008, ANY BRANCH OR SUBSIDIARY ABROAD PERFORMING RESEARCH AND DEVELOPMENT ACTIVITIES?
 If yes, please supply the following information for each single branch or subsidiary with reference to year 2008:
 (line should correspond to a single branch or subsidiary liable)

Yes No

COUNTRY OF PERFORMANCE		R&D EXPENDITURE	R&D PERSONNEL
a.	2501		
b.	2502		
c.	2503		
d.	2504		
e.	2505		
TOTAL	2506		

Previous Save Save and Exit Next Compute totals

FIGURE 6.22 Stressing the importance of a filter question with a frame

Some research suggests framed answer spaces should be preferred because they decrease the cost of keying forms or increase accuracy if questionnaires are optically scanned and verified. It is sometimes also believed that framed answer spaces often require less interpretation on the part of the data entry typist or data editor. There is, however, no experimental evidence that this is the case. Cognitive testing of questionnaire instruments has revealed that respondents do not have a strong preference for open answer spaces or framed spaces, as long as the answer spaces are sized appropriately for the information being requested. In deciding in favor of or against framed answer spaces, it is better to rely on testing on respondents, data entry typists, and data editors.

As a general rule, it is desirable to use the same type and physical dimensions of answer spaces when requesting similar information. For example, if percentages or euro amounts are asked for in different parts of the questionnaire, it will help respondents if the same type of answer space is used and if the same additional information and instructions are included (Couper, Traugott, and Lamias, 2001; Christian, Dillman, and Smyth, 2007).

If respondents are asked to enter values or amounts, they should do so with the proper unit of measurement. They should be helped with that. Errors caused by entering amounts in the wrong unit of measurement are not uncommon. For example, a company enters its turnover in dollars instead of in thousands of dollars. To avoid or reduce these problems, question instructions should make clear what is expected from respondents. This can be accomplished by adding words or symbols near the answer space. For example, for some survey questionnaires, “000” is printed next to the answer space to indicate that respondents should report in thousands of euros (or dollars). Other survey questionnaires have “.00” next to the answer space to make clear that responses are to be rounded to the nearest euro/dollar. There is no empirical evidence what works best for respondents. The main point here is that answer spaces should be consistent within a questionnaire.

EXAMPLE 6.7 Asking Values in R&D Survey

The R&D Survey of ISTAT, the Italian Statistical Institute, is a compulsory business survey that collects data on research activities and expenses. This survey asks responding companies to report values in thousands of euros. To make this clear to the respondents, the column heading contains the instruction (“Report in thousands of euros”). Moreover, the zeroes necessary to form the complete number are already next to the answer space. See Figure 6.23 for an example.

Istat.it Rilevazione statistica sulla ricerca e sviluppo nelle imprese 2008

RS1

L'indagine Istruzioni Registrazione Accesso al modello Password Cont

Company Name: PRE-MET S.P.A. | Company Code: 000015272

Indicare obbligatoriamente una risposta al questo 1

7 - INTRAMURAL R&D EXPENDITURE BY TYPE OF RESEARCH, YEAR 2008 (Thousand Euros)

TYPE OF RESEARCH	EXPENDITURE
Basic Research	701 <input type="text"/> .000,00€
Applied Research	702 <input type="text"/> .000,00€
Experimental Development	703 <input type="text"/> .000,00€
TOTAL (1)	704 <input type="text"/> .000,00€

Previous Save Save and Exit Next Compute totals

(1) This should correspond to the Total at question 2

FIGURE 6.23 Helping the respondent to record values in the proper unit of measurement

If respondents are asked to enter percentages, this should be clearly mentioned. The percent symbol should be placed adjacent to the input field. If relevant, it should be emphasized that percentages have to sum up to 100%. This can be done by placing “100%” at the bottom of the column. It is also possible to compute automatically the sum of the percentages and check whether this total equals 100%.

6.2.4.2 Use of Color. When respondents are presented with visual information in the questionnaire, they quickly decide which elements to focus on (Lidwell, Holden, and Butler, 2003; Ware, 2004). Regular and simple visual features are easier to perceive and remember. This is the so-called Gestalt principle of simplicity. Moreover, respondents are more likely to perceive answer spaces or response categories as being related to one another if they have the same color. This is the Gestalt principle of similarity.

To facilitate the comprehension process, white answer spaces should be displayed against a light-colored or shaded background questionnaire screen. As a result, the small answer space tends to “rise” above the colored background. It is an identifiable graphic area that becomes an object of interest. Therefore, it is considered more prominent. White answer boxes against colored backgrounds are especially important for use in optical imaging and scanning systems. In this case, there is a contrast between the white answer space and the colored background. Therefore, it is not necessary any more to put a frame around the answer frame.

Dividing lines tend to focus visual attention on the area around answer spaces, rather than the on answer spaces themselves. Therefore, it is recommended not to use them unless necessary.

6.2.4.3 Use of Images. Research literature has shown that small changes in the visual presentation of questions can lead to substantial variations in the response distribution and in the amount of time taken to completing questions in self-administered surveys. If the burden of answering the questions becomes too high, there is a serious risk respondents may interrupt the completion of the questionnaire resulting in (partial) nonresponse. Therefore, it is important to know why respondents break off questionnaire completion and which questions formats require most time to complete.

One issue is the inclusion of images in the questionnaire, as they definitely affect the answers of the respondent. Thus, the decision to use images, and the choice of images, has consequences in terms of measurement error.


For example, if a picture of a person is included, several aspects of this person (male or female, ill or healthy, sportsman or dull civil servant, and so on) might greatly affect how the question is perceived and therefore the answer that is given.

EXAMPLE 6.8 Surveys about mobile phones

Including images can stimulate participation in the survey and clarify the characteristics of products about which questions are asked. These images should draw attention and stimulate interest of respondents in a neutral manner.

The example in Figure 6.24 is taken from a study by Arruda Fihlo and Biffignandi (2011). In a first study, the focus of the survey was on the

This is a simple mobile phone with a memory capacity of 100 telephone numbers. This phone allows for both voice and text communication.



How much do you dislike or like (on a scale from 1 to 7) this mobile phone?

1 2 3 4 5 6 7
 Dislike very much Like very much


FIGURE 6.24 A mobile phone survey question (study 1)

characteristics of the mobile phone, not on the brand. The objective was to gain insight into new product development, and to decide which features are important for consumers. The image in the question was kept simple and without displaying a brand name. Respondents were asked to give their opinion by means of a Likert scale.

In Study 2, respondents were asked to compare brands. Therefore, mobile phone images were related to brand names. The comparison is between a new brand (called *Mandarina*) and a well-established brand (*Nokia*). Figure 6.25 contains the two questions for the *Mandarina* mobile phone.

To create a perception of a simple, unknown brand, the *Mandarina* image was kept simple, whereas the well-established brand was represented by

This is a mobile phone from *Mandarina*, a new brand that is launching several new models onto the market. On a scale from 1 to 7 how much do you dislike or like the following *Mandarine* models?



A *Mandarina* mobile phone with an integrated MP3 player. It costs 600 euro.

1 2 3 4 5 6 7
 Dislike very much Like very much


A *Mandarina* mobile phone with an integrated electronic agenda. It costs 600 euro.

1 2 3 4 5 6 7
 Dislike very much Like very much

FIGURE 6.25 Questions for the *Mandarina* mobile phone (study 2)

a more appealing image. Figure 6.26 contains the two questions for the Nokia mobile phone.

This is a Nokia mobile phone, a well-established brand with a good reputation. On a scale from 1 to 7 how much do dislike or like the following Nokia models?



A Nokia mobile phone with an integrated MP3 player. It costs 600 euro.

1 2 3 4 5 6 7
 Dislike very much Like very much

A Nokia mobile phone with an integrated electronic agenda. It costs 600 euro.

1 2 3 4 5 6 7
 Dislike very much Like very much

FIGURE 6.26 Questions for the Nokia mobile phone (study 2)

6.2.5 WEB QUESTIONNAIRES AND PARADATA

6.2.5.1 Definition of Paradata. The concept of paradata emerged in the era of computer-assisted interviewing (CAI). *Paradata* are defined as data that are generated during the fieldwork of the survey. Early examples of paradata related to keystroke files and audit trails that were automatically generated by many CAI systems. They were a by-product of the CAI system and used for technical purposes, such as error diagnosis and recovery from failure. Already early in their existence it was realized that paradata can also provide insight into the process of asking and answering questions, and how respondents interact with computer systems. Therefore they may help to improve data collection. Early applications

of paradata can be found in the studies by Couper, Hansen, and Sadosky (1997) and Couper, Horm, and Schlegel (1997). Couper (1998) was the first to coin the term “paradata.” Heerwegh (2003) describes the use of paradata in web surveys.

Paradata can be particularly useful in web surveys. There are no interviewers, so other means have to be found for obtaining insight into the activities of respondents while they are completing the questionnaire form. Examples of paradata that can be collected are the keys pressed by respondents, the movement of the mouse, changes they have made, the time it takes to answer a question or the complete questionnaire, use of help functions, and so on.

Two types of paradata can be collected during web questionnaire completion: server-side paradata and client-side paradata. *Server-side paradata* are collected by software tools running at the server where the survey resides. These are usually data measured at the level of the questionnaire. Examples of server-side paradata include the download time of the survey questionnaire, the number of times the survey web page was accessed, the time spent in each visit, identifiers of respondents, the type of browser used, and the operating system on the computer of the respondent. Server-side paradata are usually stored in *logfiles* for later analysis.

EXAMPLE 6.9 Audit trials

In 2006, a pilot study was carried out by Statistics Netherlands to examine the feasibility of introducing an electronic questionnaire for the Structural Business Survey. Details are described in the study by Sniijkers and Morren (2010). Figure 6.27 shows a sample screen (in Dutch) of the downloadable questionnaire form.

Audit trails were used to examine the response behavior in this survey. Audit trails are a specific kind of paradata. They contain information on all events occurring on the computer while the respondent completes the questionnaires. For example, they record all key presses, all mouse clicks, and the time at which they occurred. This information can be used to answer questions, such as follows:

- Do respondents complete the questionnaire in one session or in several sessions?
- How long does it take to complete a questionnaire as a whole?
- How do respondents navigate through the questionnaire?
- Do respondents use the print function? How often and at what moments in the completion process do they print the questionnaire?
- Do respondents first browse through the questionnaire before they start filling it in (to get an overview of the questionnaire), or do they start filling it in right away question-by-question?

FIGURE 6.27 A screen of the Structural Business Surveys

The survey was conducted with the Blaise system. This survey software can automatically generate an audit trail. The audit trail data were imported into the statistical analysis package SPSS. Figure 6.28 shows a small part of the data for one respondent

	id	date	time	action	field	value
1	12345	22-MAY-2007	10:43:17	0	0	
2	12345	22-MAY-2007	10:43:58	1	40	
3	12345	22-MAY-2007	10:44:19	2	40	01-04-2005
4	12345	22-MAY-2007	10:44:19	14	40	
5	12345	22-MAY-2007	10:44:29	14	27	
6	12345	22-MAY-2007	10:44:48	14	21	
7	12345	22-MAY-2007	10:44:51	1	48	
8	12345	22-MAY-2007	10:45:15	3	48	
9	12345	22-MAY-2007	10:50:20	8	48	
10	12345	22-MAY-2007	10:51:05	10	48	
11	12345	22-MAY-2007	10:51:05	99	0	
12	12345	20-JUN-2007	8:52:28	0	0	
13	12345	20-JUN-2007	8:53:09	8	15	
14	12345	20-JUN-2007	8:54:50	13	15	
15	12345	20-JUN-2007	8:54:53	11	15	
16	12345	20-JUN-2007	8:55:02	9	15	
17	12345	20-JUN-2007	9:35:58	99	0	
18	12345	29-JUN-2007	15:10:38	0	0	

FIGURE 6.28 Blaise audit trail data imported into SPSS

Each line represents data on one event. The respondent is identified (id), the data and time, a code for the specific action (for example, entering an answer field, exiting a field, clicking on help, and clicking on the save button), the question (field) for which the event occurred, and the answer entered by the respondent.

Client-side paradata are collected by software tools running on the computers of the respondents. These data describe how the respondents answer the questions in the questionnaire. They give insight into which questions are answered, in what order, and whether all relevant questions have been answered. Many activities can be recorded, like keys that have been pressed and mouse movement across the screen. As time stamps can be associated with each activity, it is for example possible to measure how long it takes to answer a question. All these data may provide insight into how easy or difficult it is for the respondents to complete the questionnaire. Recording paradata is often implemented by embedding Javascript code in the survey questionnaire.

6.2.5.2 Use of Paradata. Initially, researchers collected paradata to study the response behavior of different groups. Analysis of these paradata provides insight into this response behavior, and therefore, it may help to direct resources and efforts aimed at improving response rates.

Analysis of paradata is now increasingly used to understand or even control respondent behavior. An interesting new area of research is the application of methods and concepts from cognitive psychology to the development of web questionnaires and of new computerized methods of data collection. This type of research is referred to here (according to Biffignandi, 2010) as the “behavioral approach.” In this approach, the focus turns toward understanding why someone responds to a web survey whereas others do not, and whether and in what ways these two groups may differ on key variables of interest.

Using client-side paradata, it is possible to better understand how respondents construct their answers, including data about the time it takes to answer a question and about possible changes in their answers. Results suggest that the visual layout of survey questions not only affects the number but also the types of changes respondents make. The theory does not concern itself with the impact of nonresponse on estimates although it is not unlikely that different phenomena may cause a different type of nonresponse bias.

The objective of the behavioral research approach is to answer questions such as who is more likely to respond, why does nonresponse occur, who is likely to be hard to reach, and how does interest in the survey topic affect the willingness to participate. As an ultimate goal, behavioral studies aim to acquire an understanding of how respondents construct their answers in their natural setting, which in turn facilitates web questionnaire design and tailoring.

From the methodological point of view, behavioral analyses are related to the Cognitive Aspects of Survey Methodology movement (CASM). In many empirical studies, the *theory of planned behavior* (TPB) is applied. This theory was proposed by Ajzen (1985) and is described by Ajzen in his 1991 study. The main objective of TPB is to obtain a more comprehensible picture of how intentions are formed. This theory is an extension of the theory of reasoned action (Ajzen and Fishbein, 1980). The theory of planned behavior specifies the nature of relationships between beliefs and attitudes. According to these models, people's evaluations or attitudes toward behavior are determined by their accessible beliefs about such behavior, where a belief is defined as the subjective probability that the behavior will produce a certain outcome. Specifically, the evaluation of each outcome contributes to the attitude in direct proportion to the person's subjective belief in the possibility that the behavior produces the outcome in question.

The central factor of this theory is the individual intention to perform a given behavior. The first postulate is that intention is the result of three conceptual determinants. Human behavior is guided by the following three kinds of considerations:

- *Behavioral beliefs*: They produce a favorable or unfavorable *Attitude toward Behavior*. It is the degree to which a person has a favorable or unfavorable evaluation or appraisal of the behavior in question (Ajzen, 1991). When new issues arise requiring an evaluative response, people can draw on relevant information (beliefs) stored in memory. Because each of these beliefs carries evaluative implications, attitudes are automatically formed.
- *Normative beliefs*: They result in *subjective norms* (SNs). This refers to individuals' perceptions of others' opinions of their behavior. SN has been shown to be a predictor of behavior (Bagozzi, Davis, Wasshaw, 1992; Fishbein & Ajzen, 1975, 2010; Mathieson, 1991). In the context of web surveys, subjective norm would be the amount of influence a person's superiors (i.e., employers, parents, or spouse) would have in influencing a choice to participate in the survey.
- *Control beliefs*: They give rise to *perceived behavioral control* (PBC). It is the perceived social pressures to perform, or not, a certain behavior (Ajzen, 1991), i.e., the subject's perception of other people's opinions of the proposed behavior. This pressure can have, or have not, an influential role. For example, in France, the failure of a company is negatively perceived, whereas in the United States, a person can undergo several failures and yet often undertake new attempts. Perceived behavioral control is presumed to not only affect actual behavior directly, but also affect it indirectly through behavioral intention. PBC refers to the perception of an individual of whether or not he or she can perform a particular behavior. In the context of a web survey, PCB would be defined as whether an individual could use the web tools to participate and engage in the survey successfully. Therefore, PBC would be similar to computer self efficacy (CSE), see the study by

Bagozzi, Davis and Warshaw (1992). CSE is defined as the judgment of one's capability to use a certain aspect of information technology (Agarwal, Sambamurthy, and Stair, 2000; Compeau, and Higgins, 1995; Gist, 1989; Gist, Schwoerer, and Rosen, 1989). Attitudinal models (such as the above-mentioned TPB) use multiple constructs to predict and explain behavior. The concept of perceived ease or difficulty of performing a behavior (Ajzen, 1991) was introduced into the TPB to accommodate the nonvolitional elements inherent, at least potentially, in all behaviors (Ajzen, 2002a).

This theory assumes that human social behavior is reasoned, controlled, or planned in the sense that it takes into account the likely consequences of the considered behavior (Ajzen and Fishbein, 2000). This model has been applied for the prediction of many types of human behavior, such as electoral choice and intention to stop smoking.

Combining attitude toward behavior, the subjective norm, and perceived behavioral control leads to *behavioral intention* (Ajzen, 2002b). As a general rule, the more favorable the attitude toward behavior and subjective norm is, and the greater the perceived behavioral control is, the stronger the person's intention to perform the behavior in question should be. Finally, given a sufficient degree of actual control over the behavior, people are expected to carry out their intentions when the opportunity arises (Ajzen, 2002b).

In its simplest form, the Theory of Planned Behavior can be expressed as the following statistical function:

$$(6.1) \quad BI = (W1)AB[(b) + (e)] + (W2)SN[(n) + (m)] + (W3)PBC[(c) + (p)],$$

where

- BI = behavioral intention
- AB = attitude toward behavior
- (b) = the strength of each belief
- (e) = the evaluation of the outcome or attribute
- SN = social norm
- (n) : the strength of each normative belief
- (m) : the motivation to comply with the referent
- PBC = perceived behavioral control
- (c) = the strength of each control belief
- (p) = the perceived power of the control factor
- $W1, W2, W3$ = empirically derived weights/coefficients

For instance, Tourangeau (2003) proposes an extended TPB model to explain the intentions of potential respondents in participating in web surveys. His model is shown graphically in Figure 6.29.

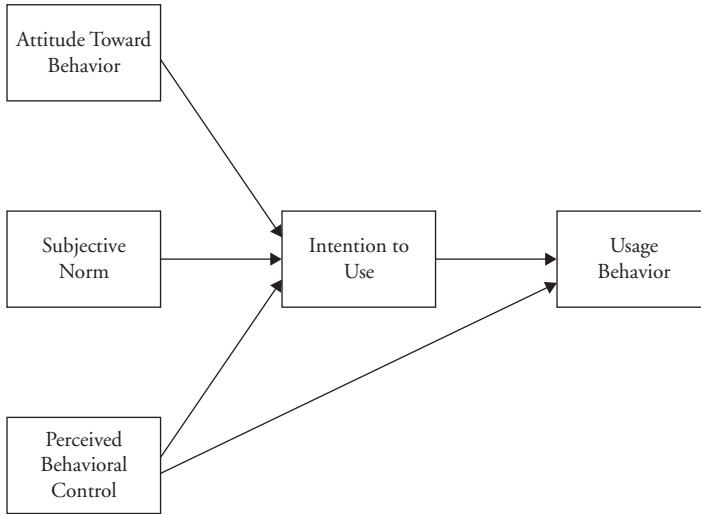


FIGURE 6.29 Theory of Planned Behavior

As a proxy of the burden on the respondent, often the time spent to complete the web questionnaire is used, i.e., the actual number of minutes spent working on the questionnaire. Although this is not a perfect measure, it is one that indicates the intensity and duration of the time spent performing the task of participating in the survey.

6.2.6 TRENDS IN WEB QUESTIONNAIRE DESIGN AND VISUALIZATION

6.2.6.1 The Cognitive Approach to Web Questionnaire Design. The World Wide Web has features that are not available for paper questionnaires. These features enable a new way of thinking about the design and self-administration of survey questions. In particular, web survey interaction can be conceived of as a dialogue consisting of turns of interaction between a user (respondent) and the system (the interviewing agent). From this viewpoint, inspired by the collaborative comprehension of elements of cognitive psychology and psycholinguistics, each system action (presenting a question and prompting for an answer) and each respondent action (clicking to proceed, reading a question, and typing a number as an answer) corresponds to some physical or mental action in a face-to-face interview. Conceiving interaction this way not only highlights and clarifies the function of each action, but it also opens up possibilities that survey designers can implement in web surveys. The task of the so-called cognitive approach is to assess how respondents go about answering questions.

There is widespread agreement about the cognitive processes involved in answering questions optimally. See, for example, Cannell, Miller, and Oksenberg

(1981), Tourangeau and Rasinski (1988), and Willis (2005). Respondents are presumed to execute the following steps:

1. Interpreting the question and deducing its intent.
2. Searching their memories for relevant information, and then integrating whatever information comes to mind into a single judgment.
3. Translating this judgment into a response, by selecting one of the alternatives offered by the question.

Each of these steps can be complex, involving considerable cognitive work. A wide variety of motives may encourage respondents to perform these activities, including the desire for self-expression, interpersonal response, intellectual challenge, self-understanding, altruism, or emotional catharsis. They can also be motivated by the desire to assist the survey sponsor, (e.g., to help employers improving working condition, to help businesses designing better products, or to help governments formulating better-informed policies). To the extent that such motives inspire a respondent to perform the task in a thorough and unbiased manner, the respondent may be said to be *optimizing*. The web questionnaire designer hopes all respondents will optimize throughout a questionnaire. This is often an unrealistic expectation. Some people may agree to complete a questionnaire as a result of a relatively automatic compliance process or because they are required to do so. Thus, they provide answers, with no intrinsic motivation to produce high-quality answers. Other respondents may satisfy whatever desires motivated them to participate after answering a first set of questions, and become fatigued, disinterested, or distracted as a questionnaire progresses further.

Rather than make the effort necessary to provide optimal answers, respondents may take subtle or dramatic shortcuts. In the former case, respondents may simply be less thorough in comprehension, retrieval, judgment, and response selection. They may be less thoughtful about a question's meaning, search their memories less comprehensively, integrate retrieved information less carefully, or select a response choice less precisely. All steps are executed but less diligently as in the case of optimization. Instead of providing the most accurate answers, respondents settle for merely satisfactory answers. This response behavior is termed *weak satisficing* (Krosnick, 1991, borrowing the term from Simon, 1957).

In other cases, respondents skip the retrieval and judgment steps altogether. That is, respondents may interpret each question superficially and select what they believe will appear to be a reasonable answer. The answer is selected without reference to any internal psychological cues specifically relevant to the attitude, belief, or event of interest. Instead, the respondent may look to the wording of the question for a cue, pointing to a response that can be easily selected and easily defended if necessary. If no such cue is present, the respondent may select an answer completely arbitrarily. This process is termed *strong satisficing*.

It is useful to see optimizing and strong satisficing as the two ends of a continuum indicating the degrees of thoroughness with which the response steps

are performed. The strong satisficing end involves little effort in the interpretation and answer reporting steps and no retrieval or integration at all. In between are intermediate levels, like weak satisficing.

The likelihood of satisficing is thought to be determined by three major factors: task difficulty, respondent ability, and respondent motivation (Krosnick, 1991). *Task difficulty* is a function of both question-specific attributes (e.g., the difficulty of interpreting a question and of retrieving and manipulating the requested information) and attributes of the questionnaire's administration (e.g., the pace at which an interviewer reads the questions and the presence of distracting events). *Ability* is shaped by the extent to which respondents can perform complex mental operations, practiced at thinking about the topic of a particular question, and equipped with pre-formulated judgments on the issue in question. *Motivation* is influenced by need for cognition (Cacioppo, Petty, Feinstein, and Jarvis 1996), the degree to which the topic of a question is personally important, beliefs about whether the survey will have useful consequences, respondent fatigue, and aspects of questionnaire administration (such as interviewer behavior) that either encourage optimizing or suggest that careful reporting is not necessary. Efforts to minimize task difficulty and maximize respondent motivation are likely to pay off by minimizing satisficing and maximizing the accuracy of self-reports.

Web surveys can incorporate audio and video to more closely approximate an interviewer-assisted survey. A so-called *cognitive interview* provides basic information for understanding the impact of audio, video, and even interviewers' faces. All this information allows for further improvements in a web questionnaire. Web survey researchers are increasingly conducting experiments where they introduce pictures of faces of interviewers. This kind of cognitive interview is labor intensive. Current results suggest that bringing features of human dialogue into web surveys can exploit the advantages of both the interviewer-assisted and the self-administered interviews. Nevertheless, it should be noted that introducing interviewers may also introduce biases.

EXAMPLE 6.10 Cognitive Interviewing

A simple example shows how cognitive interviewing can be used in web surveys. The Samplion Survey Research Institute (SSRI) wants to improve the design of its web survey questionnaires. The hope is that this will increase response rates. An experiment is set up to compare various survey designs.

A lab experiment is organized involving a company that provides this kind of services. The lab is equipped with audio and video recording equipment, using multiple cameras and two-way communication with an observation room. Each room contains a computer on which a web survey questionnaire can be completed.

A group of 50 respondents takes part in this experiment. They are divided in five subgroups:

1. 10 respondents are invited to complete the web questionnaire without any special external stimulus, just with the standard survey design. This is the control group.
2. 10 respondents are invited to complete the questionnaire with audio support: A male voice introduces the survey and asks the respondent to participate in the survey.
3. 10 respondents are invited to complete the questionnaire with video and audio support: An animated male face appears on the screen and asks the respondent to participate in the survey.
4. 10 respondents are invited to complete the questionnaire with audio support: A female voice introduces the survey and asks the respondent to participate in the survey.
5. 10 respondents are invited to complete the questionnaire with video and audio support: An animated male face appears on the screen and asks the respondent to participate in the survey.

The behavior of the respondents can be viewed and recorded. By studying this and other information, it may become clear in which circumstances and at which points in the questionnaires respondents face problems, make mistakes, or do not know what to do. By comparing the groups, insight may be obtained into the effects of audio and video support.

6.3 Application

This section describes a web survey on the values of consumers with respect to purchases of trendy and vogue products. This survey is described in detail by Biffignandi and Manzoni (2011).

Interviewed people were asked to complete a questionnaire made up of two different parts. In the first part, the respondents were asked to indicate which consequences a certain product attribute is linked to. In the second part, they had to indicate which consumers' values are linked to certain consumers' consequences. Both product attributes and consequences were identified in a previous study using a different sample.

The survey was administered to three experimental groups determined by different data collection modes (paper questionnaire vs. computer-based questionnaire) and different measurement scales used in the questionnaire. There was a 5-point Likert numerical scale and a 2-point Likert scale (Yes/No). The following hypotheses were tested:

TABLE 6.2 The design and the consumers' values experiment

Group	Mode of data collection	Type of Likert scale
1	Paper	Dichotomy scale
2	Computer-based	Dichotomy scale
3	Computer-based	5-level Likert numerical scale

H₁: The mode of data collection (paper vs. computer-based) influences the respondents' responses.

H₂: The type of Likert scale (5-point vs. 2-point) influences the respondents' responses.

H₃: The 2-point Likert scale is equal to the 5-point Likert scale if the latter is recoded such that the first four items are assigned to value 0 and the last item is assigned to value 1.

A field experiment was designed using various combinations of the mode of data collection and different scales. The experimental design is summarized in Table 6.2.

For the analysis of the survey results, the questionnaire for group 3 (computer-based, with a 5-point Likert scale) was recoded using four different dichotomization criteria: 1 versus 2 + 3 + 4 + 5, 1 + 2 versus 3 + 4 + 5, 1 + 2 + 3 versus 4 + 5, and 1 + 2 + 3 + 4 versus 5.

The results for the different groups (including the recoded versions) were compared by adopting the computer-based dichotomy questionnaire as a benchmark. With respect to comparing the web-based and paper-based questionnaires, the results showed that the mode of data collection had no influence on the responses of the respondents.

With respect to comparing the 2-point Likert scale and the 5-point Likert scale, analysis showed that the recode 1 + 2 + 3 + 4 versus 5 provided the same results as the 2-point scale. This confirms the theory that respondents are unsatisfied until completely satisfied. In other words, respondents are aware of not being unsatisfied only after realizing they are satisfied. As an example, Figure 6.30 contains an example of one of the survey questions. The meaning of the scale values was explained in the survey questionnaire.

I have just bought a HIGH QUALITY product.					
I feel gratified	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5
I will be sure it will be noticed when I wear/use it.	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5
It will make my life easy	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5
I am satisfied	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5

FIGURE 6.30 A matrix question in the Consumers' Values Survey

6.4 Summary

A web survey is a tool that can collect information in a cheap and timely manner, producing good quality data with a reduced response burden. To achieve this, the questionnaire designer should not only apply traditional survey research principles, but also more recent research results that are more specific for online data collection. Together, they provide a general framework for designing web surveys.

A well-designed web survey requires research that pays attention to several different aspects. One aspect is to define the research tasks and the environment required for this. Another aspect is to select the appropriate hardware and software. There should be a balance between technological sophistication and user accessibility.

With respect to the survey questionnaire, careful attention must be paid to labeling the answers for closed questions and all-that-apply questions. Navigation through the questionnaire must be easy and intuitive, such that only relevant questions are answered, and irrelevant ones are skipped.

Another important aspect is the way the questionnaire is visualized on the screen. Proper use of symbols, colors, and images may help in this respect. Offering answer spaces that are sized appropriately for the information being requested improves the likelihood that respondents will provide answers and good quality information.

A final aspect to pay attention to is to collect paradata. This information may help to detect problems in the survey instrument. Analysis of these data and, more generally, research on cognitive interviewing can help in improving web survey design in the future.

Summing up, the web questionnaire design process is challenging work that requires knowledge of different fields, such as (1) the subject-matter topic of the survey, (2) computer technology, (3) textual language (labeling, wording, graphics, and so on), (4) use of paradata, and (5) cognitive methods. The most promising direction in the field of web survey design research is user-centered design methods. This approach seems to be worthwhile in terms of conducting effective web survey success and collecting good quality data.

KEY TERMS

Check-all-that-apply question: A question allowing respondents to select more than one answer from a list of answer options. Check boxes are used to select answer options.

Check box: A graphical user-interface element. It is a small rectangular box allowing respondents to select an associated option. An option is selected or deselected by clicking on it. A selected option is denoted by a tick mark.

Closed question: A question allowing respondents to select exactly one answer from a list of answer options. Radio buttons or drop-down boxes can be used to select answer options. Radio buttons are preferred.

Drop-down box: A list of answer options from which one answer can be selected. Initially, the list is invisible. The list can be opened by clicking on its title. Longer lists will only be partly visible. Other parts of the list can be shown by using a scroll bar.

Likert scale: A type of closed question where the answers constitute a scale. It allows respondents to indicate how closely their feelings match the question or statement on a rating scale. Likert scales are useful measuring the degree of respondents' feelings or attitudes concerning some issue.

Numerical question: A type of open question that allows respondents to enter a number.

Open question: A question that does not have a set of possible answers associated with it. The verbatim response of the respondent is recorded. Such a question is used if respondents are required to respond completely and freely.

Ordinal question: A type of closed question that allows respondents to rank their answer to a question. Ordinal questions are used for determining priorities or preferences of respondents.

Radio button: A graphical user-interface element. It is a small circle box allowing respondents to select the associated option. An option is selected by clicking on it. A selected option is denoted by a dot in the circle. An option is de-selected by selecting another option.

Theory of Planned Behavior (TPB): A psychological theory that links attitudes to behavior.

EXERCISES

Exercise 6.1. Closed questions must have:

- a. Only numerical answers
- b. Text areas for each option
- c. Mutually exclusive categories
- d. Check boxes

Exercise 6.2. Suppose a customer satisfaction survey on motorbike preferences is conducted. What is the best way to obtain an answer on the question "Indicate which motorbike model your prefer"?

- a. A set of check boxes
- b. An open question
- c. A set of radio buttons

d. A drop-down box

Exercise 6.3. A web survey contains the following question:

<p>Do you agree or disagree with the following statement: School is a place where I usually feel great?</p> <p><input type="checkbox"/> Strongly agree</p> <p><input type="checkbox"/> Agree</p> <p><input type="checkbox"/> Neither agree or disagree</p> <p><input type="checkbox"/> Disagree</p> <p><input type="checkbox"/> Strongly disagree</p>

Which of the following statements is correct?

- a. The box format is correct.
- b. The answers form a Likert scale.
- c. The wording is not correct.
- d. This is a check-all-that-apply question.

Exercise 6.4. Which of the following statements about a web questionnaire is correct?

- a. It is an interviewer-assisted survey.
- b. It allows for collecting server-side paradata.
- c. The use of open questions is prohibited.
- d. As much as possible, matrix questions should be used.

Exercise 6.5. Which of the following statements are advantages of web questionnaires over paper questionnaires?

- a. They allow for customization.
- b. Radio buttons and check boxes can enforce different treatment.
- c. They do not allow for partial completion.
- d. “Don’t know” answers are not possible.

Exercise 6.6. For what kind of answer formats are radio buttons used in web surveys?

- a. For selecting one answer.
- b. For selecting more than one answer.

- c. For recording any text.
- d. For recording amounts and values.

Exercise 6.7. How should “don’t know” be treated in web surveys?

- a. It should never be offered as an answer option.
- b. It should always be offered as an answer option.
- c. It should only be offered as an answer option in opinion questions.
- d. It should only be offered as an answer option in factual questions.

Exercise 6.8. What can be said about the use of pictures in web survey questions?

- a. They should never be used as they may suggest a wrong reference framework for the respondent.
- b. They should always be used as they suggest the right reference framework for the respondent.
- c. They should always be used as they make the questionnaire more attractive.
- d. They should only be used to help the respondents to understand the question correctly.

REFERENCES

- Agarwal, R., Sambamurthy, V., & Stair, R. (2000), Research Report: The Evolving Relationship Between General and Specific Computer Self-efficacy—An Empirical Assessment. *Information Systems Research*, 11, pp. 418–430.
- Ajzen, I. (1985), From Intentions to Actions: A Theory of Planned Behavior. In Kuhl, J. & Beckman, J. (eds.), *Action-control: From Cognition to Behavior*, Springer Verlag, Heidelberg, Germany, pp. 11–39.
- Ajzen, I. (1991), The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, 50, pp. 179–211.
- Ajzen, I. (2002a), Perceived Behavioral Control, Self-efficacy, Locus of Control, and the Theory of Planned Behavior. *Journal of Applied Social Psychology*, 32, pp. 665–683.
- Ajzen, I. (2002b), Residual Effects of Past on Later Behavior: Habituation and Reasoned Action Perspectives. *Personality and Social Psychology Review*, 6, pp. 107–122.
- Ajzen, I. & Fishbein, M. (1980), *Understanding Attitudes and Predicting Social Behavior*. Prentice Hall, Englewood Cliffs, NJ.
- Ajzen, I. & Fishbein, M. (2000), Attitudes and the Attitude-behavior Relation: Reasoned and Automatic Processes. In: Stroebe, W. & Hewstone, M. (eds.), *European Review of Social Psychology*, 11, pp. 1–33.
- Arruda Filho, E. J. M. & Biffignandi, S. (2011), *Utilitarian Guilt Pushed by Hedonic Preference: Perceived Values to a Society Choice’s Justification*. Research Report, CASI/DMSIA, Bergamo, Italy.

- Bagozzi, R. P., Davis, F. D., & Warshaw, P. R. (1992), Development and Test of a Theory of Technological Learning and Usage. *Human Relations*, 45, pp. 659–686.
- Biffignandi, S. & Manzoni, A. (2011), *The Mode and Scale Measurement Impact in a Consumer Behaviour Survey*. Research Report, CASI/DMSIA, Bergamo, Italy.
- Biffignandi, S. (2010), Modeling Non-sampling Errors and Participation in Web Surveys, *Proceedings of the 45th SIS Scientific Meeting*, Padova, Italy.
- Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996), Dispositional Differences in Cognitive Motivation: The Life and Times of Individuals Varying in Need for Cognition. *Psychological Bulletin*, 119, pp. 197–253.
- Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981), Research on Interviewing Techniques. In: Leinhardt, S. (ed.), *Sociological Methodology*, Jossey-Bass, San Francisco, CA.
- Christian, L. M., Dillman, D. A., & Smyth, J. D. (2007), Helping the Respondents Get It Right the First Time: The Influence of Words, Symbols. *Public Opinion Quarterly*, 71, pp. 113–125.
- Christian L. M., Parsons, N. L., & Dillman, D. A. (2009), Measurement in Web Surveys: Understanding the Consequences of Visual Design and Layout. *Sociological Methods and Research*, 37, pp. 393–425.
- Cobanoglu, C., Warde, B., & Moreno, P. J. (2001), A Comparison of Mail, Fax and Web-based Survey Methods. *The Market Research Society*, 43, pp. 441–452.
- Compeau, D. R. & Higgins, C. A. (1995), Computer Self-efficacy: Development of a Measure and Initial Test. *MIS Quarterly*, 19, pp. 189–211.
- Couper, M. P. (1998), Measuring Survey Quality in a CASIC Environment. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 41–49.
- Couper, M. P. (1999), Usability Evaluation of Computer Assisted Survey Instruments. *Proceedings of the Third ASC International Conference*, Edinburgh, U. K., pp. 1–14.
- Couper, M. P. (2000), Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, pp. 464–494.
- Couper, M. P. (2008), *Designing Effective Web Surveys*. Cambridge University Press, New York.
- Couper, M. P., Hansen, S. E., & Sadosky, S. A. (1997), Evaluating Interviewer Use of CAPI Technology. In: Lyberg, L., Biemer, P., Collins, M. De Leeuw, E., Dippo, C., Schwarz, N. & Trewin, D. (eds.), *Survey Measurement and Process Quality*, John Wiley & Sons, New York, pp. 267–286.
- Couper, M. P., Horm, J., & Schlegel, J. (1997), Using Trace Files to Evaluate the National Health Interview Survey CAPI Instrument. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VO., pp. 825–829.
- Couper, M. P., Traugott M. W., & Lamias, M. J. (2001), Web Survey Design and Administration. *Public Opinion Quarterly*, 65, pp. 230–253.
- Couper, M. P., Tourangeau, R., & Kenyon, K. (2004a), Picture This! Exploring Visual Effects in Web Surveys. *Public Opinion Quarterly*, 68, pp. 255–266.
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Crawford, S. D. (2004b), What They See Is What They Get: Response Options for Web Surveys. *Social Science Computer Review*, 22, pp. 111–127.

- DeRouvray, C. & Couper, M. P. (2002), Designing a Strategy for Reducing “No Opinion” Responses in Web-Based Surveys. *Social Science Computer Review*, 20, pp. 3–9.
- Dillman, D. A. (2007), *Mail and Internet Surveys. The Tailored Design Method*. John Wiley & Sons, New York.
- Dillman, D. A. (2009), *Mail and Internet Surveys. The Tailored Design Method*. John Wiley & sons, New York.
- Dillman, D., Smyth, J., Christian, L., M. (2009), *Internet, Mail and Mixed Mode Surveys. The Tailored Design Method*. John Wiley & Sons, Hoboken, NJ.
- Fishbein, M. & Ajzen, I. (1975), *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA.
- Fishbein, M. & Ajzen, I. (2010), *Predicting and Changing Behavior. The Reasoned Action Approach*. Routledge, London, U.K..
- Gist, M. (1989), The Influence of Training Method on Self-efficacy and Idea Generation Among Managers. *Personnel Psychology*, 42, pp. 787–805.
- Gist, M. E., Schwoerer, C. E., & Rosen, B. (1989), Effects of Alternative Training Methods on Self-efficacy and Performance in Computer Software Training. *Journal of Applied Psychology*, 74, pp. 884–891.
- Heerwegh, D. (2003), Explaining Response Latency and Changing Answers Using Client-side Paradata from a Web Survey. *Social Science Computer Review*, 21, pp. 360–373.
- Heerwegh, D. & Loosveldt, G. (2002), *An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys*. Paper presented at the International Conference on Improving Surveys, Copenhagen, Denmark.
- Krosnick, J. A. (1991), Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5, pp. 213–236.
- Lidwell, W., Holden, K., & Butler, J. (2003), *Universal Principles of Design*. Rockport, Gloucester, MA.
- Mathieson, K. (1991), Predicting User Intentions: Comparing the Technology Acceptance Model With the Theory of Planned Behavior. *Information Systems Research*, 2, pp. 173–191.
- Redline, C. D. & Dillman, D. A. (1999), *The Influence of Symbolic, Numeric and Verbal Languages on Navigational Compliance in Self-Administered Questionnaires*. Paper presented at the International Conference on Survey Nonresponse, Portland, OR.
- Shropshire, K. O., Hawdon, J. E. & Witte, J. C. (2009), Topical Interest Web Survey Design: Balancing Measurement, Response, and Topical Interest. *Sociological Methods & Research*, 37, pp. 344–370.
- Simon, H. A. (1957), *Models of Man: Social and Rational*. John Wiley & Sons, New York.
- Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006), Comparing Check-all and Forced-choice Question Formats in Web surveys. *Public Opinion Quarterly*, 70, pp. 66–77.
- Snijders, G. & Morren, M. (2010), *Improving Web and Electronic Questionnaires: The Case of Audit Trails*. Paper presented at the European Conference on Quality in Official Statistics (Q2010), Helsinki, Finland.
- Sudman, S., Bradburn N., & Schwarz, N. (1996), *Thinking About Answers*. Jossey-Bass, San Francisco, CA.
- Sudman, S. & Bradburn, N. M. (1973), *Asking Questions*. Jossey-Bass, San Francisco, CA.

- Tourangeau, R. (2003), Cognitive Aspects of Survey Measurement and Mismeasurement. *International Journal of Public Opinion Research*, 15, pp. 3–7.
- Tourangeau, R. & Rasinski, K. A. (1988), Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*, 103, pp. 299–314.
- Tourangeau, R., Couper, M. F., & Steiger, D. M. (2003), Humanizing Self—Administered Surveys: Experiments on Social Presence in Web and IVR Surveys. *Computers in Human Behavior*, 19, pp. 1–24.
- Tourangeau, R., Couper, M. P., & Conrad, F. (2004), Spacing, Position and Order, Interpretative Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, 68, pp. 368–393.
- Tourangeau, R., Couper, M. P., & Conrad, F. (2007), Color, Labels, and Interpretative Heuristics for Response Scales, *Public Opinion Quarterly*, 71, pp. 91–112.
- Ware, C. (2004), *Information Visualization : Perception for Design*. Morgan Kaufmann, San Francisco, CA.
- Willis, G. B. (2005), *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Sage, Thousand Oaks, CA.

Mixed-Mode Surveys

7.1 Introduction

A mixed-mode (or multimode) survey is a survey in which various data collection modes are combined. Examples of data collection modes are interviewer-assisted data collection (face-to-face and by telephone) and self-administered data collection (by mail or by the Internet).

A web surveys is often one of the collection modes in a mixed-mode survey. For example, survey participants may be offered the choice to either complete a paper questionnaire or complete it on the web.

A special type of a mixed-mode survey is a survey for which records of respondents and nonrespondents can be linked to a register or other administrative database. An example of such a register is the Social Statistical Database of Statistics Netherlands. By linking survey records to this database, effective nonresponse correction can be applied. See the study by Bethlehem, Cobben, and Schouten (2011) for more information about this application. Rogelberg and Stanton (2007) call the analysis of this type of mixed-mode data *Archival Analysis*.

Using the Internet as one of the data collection modes in a mixed-mode survey offers new opportunities, but also it creates challenges. Both opportunities and challenges are discussed in this chapter.

There have been mixed-mode surveys in the era before web surveys appeared. There are early applications where telephone interviewing was used as an alternative for self-completion of mail questionnaires. There also have been examples of large-scale, mixed-mode surveys with mail, telephone interviewing, and face-to-face interviewing as modes of data collection.

EXAMPLE 7.1 A mixed-mode survey on customer satisfaction

The management of the library of the University of Bergamo carried out a survey on customer satisfaction. To recruit people for this survey, a simple random sample was selected from the administrative database of the students of the university. Students were called by telephone and invited to participate in the survey. If they agreed, they were offered the choice to answer the survey questions in a telephone interview or to provide their e-mail address so that they could fill in the questionnaire on the Internet.

A link to the questionnaire on the Internet and a unique access code was sent to each student that provided an e-mail address.

In this survey, telephone interviewing was chosen in the recruitment phase and two concurrent modes (telephone and web) in the response phase. This is a simple example of a mixed-mode survey. Such an approach is useful if a list of e-mail addresses is not available.

Note that in this example every student had a university-provided e-mail address, but at the time of the survey (2006), students did not use the university e-mail address very often. Consequently, many e-mails turned out to be undeliverable or remained unread. Therefore, the mixed-mode approach was chosen.

Nowadays, the Internet is used more and more as one of the data collection modes in mixed-mode surveys. See, for example, the studies by Biemer and Lyberg (2003), Christian, Dillman, and Smyth (2005) and De Leeuw (2005). There are two main reasons why survey researchers decide to conduct a mixed-mode survey with the Internet as one of the modes. One is that it is considered a means to reduce the costs of the survey, and the other is to increase response rates.

Traditionally, large household surveys are interviewer-assisted surveys. Data are collected by means of telephone or face-to-face interviewing. This makes these surveys expensive. Interviewer costs are often a major cost component. There are no interviewers in a web survey. This makes web data collection an attractive alternative mode in a mixed-mode survey. Such strategies have been discussed by Dillman (2000).

Survey organizations in many countries are faced with reducing response rates. This has an impact on the quality of the survey results. Mixed-mode surveys have been suggested as a means to increase response rates. See, for example, the study by Groves et al. (2004). The idea is to offer potential respondents different modes of data collection (e.g., mail or web) from which they can choose one. If people have the possibility of selecting their preferred mode, this may increase the response rate. Another possibility could be to determine beforehand what the most effective mode will be for people, dependent on their personal characteristics. For example, one could apply face-to-face interviewing for the elderly and the Internet for young people.

Mixing modes can have many advantages. It should be noted, however, that first experiences with the use of the web for data collection show that response rates are lower than for those of more traditional surveys (Sheehan, 2001). Also, changing from a single-mode survey to a multimode survey may require substantial efforts and resources.

■ EXAMPLE 7.2 The Safety Monitor

Statistics Netherlands has conducted an experiment with its Safety Monitor to determine whether a mixed-mode survey can replace a computer-assisted personal interviewing (CAPI) or computer-assisted telephone interviewing (CATI) survey without affecting the quality of the results.

The Dutch Safety Monitor is an annual survey. It measures the actual and perceived safety of the people in the country. Respondents are asked questions about feelings of safety, quality of life, and level of crime experienced.

The old Safety Monitor applied two modes of data collection. If sampled persons had a known telephone number, they were approached by CATI. If this was not the case, they were approached by CAPI. The sample size was approximately equal to 6,000 persons. This survey will be denoted by SM2.

The new Integrated Safety Monitor had four data collection modes. All individuals in the sample received a letter in which they were asked to complete the survey questionnaire on the Internet. The letter also included a postcard that could be used to request a paper questionnaire.

Two reminders were sent to those that did not respond by web or mail. If still no response was obtained, nonrespondents were approached by means of CATI if a listed telephone number was available. If not, these nonrespondents were approached by CAPI. This four-mode survey is denoted by SM4.

The response rate of SM4 turned out to be 59.7%. The response rate for SM2 was 63.5%. So including the web as one of the modes did not increase the response rate. Table 7.1 shows the composition of the response by mode for both surveys.

TABLE 7.1 Composition of the response in SM4 and SM2 Safety Monitors

Data collection mode	SM4	SM2
Web	41.8%	—
Mail	16.2%	—
CATI	30.5%	71.6%
CAPI	11.5%	28.4%
Total	100.0%	100.0%

More than half of the response (58%) in the SM4 was obtained in the SM4 with a self-administered mode of data collection (web or mail).

It should be noted that although the four-mode survey did not increase the response, substantial cost savings could be realized because interviewers were deployed in only 42% of the cases. Focusing on just interviewer costs, and ignoring all other costs (which are much lower), Beukenhorst and Wetzels (2009) found that the costs of SM4 were only 60% of the costs of SM2.

More detailed account of these experiments is given by Beukenhorst and Wetzels (2009) and Kraan et al. (2010).

This chapter gives an overview of mixed-mode surveys, as well as the related concerns and solutions. The focus is on what mixed-mode surveys are, why they should be used, how they should be used, and what future trend looks like. One of the serious concerns of mixed-mode surveys is the occurrence of mode effects. This is the phenomenon that the same question is answered differently when asked in a different mode. Therefore, in this chapter, mode effects are discussed as well as how to deal with them.

7.2 Theory

7.2.1 WHAT IS MIXED MODE?

Before explaining what a mixed-mode survey is, the concept of mode is introduced. *Mode* refers either to the approach used to contact potential respondents or to the way the data are collected. So if a sample of individuals is sent a letter in which they are requested to complete a questionnaire on the Internet, the mode for recruitment is the mail and the mode for data collection is the web. There are many different modes of data collection. Examples are face-to-face, telephone, mail, touchtone data entry (TDE), fax, and the Internet. Modes may use the same basic technology but differ in how they are used.

The data collection instrument refers to the technology used to record the answers to the questions. The instrument may be a paper questionnaire, a laptop with a computer-assisted interviewing program, or a web questionnaire. The same instrument may be used for different modes. For example, the same paper questionnaire could be used for both a mail survey and a face-to-face survey. And it is not unlikely that the same computer program is used for both CAPI and CATI.

De Leeuw (2005) describes two basic approaches to implement a mixed-mode survey.

A first approach is using different modes concurrently (parallel). The sample is divided into groups, and each group is approached with a different mode. A *concurrent mixed-mode* data collection is illustrated in Figure 7.1.

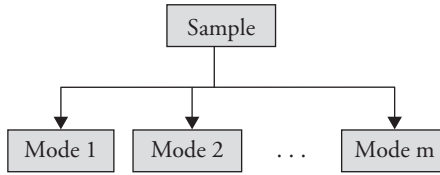


FIGURE 7.1 Concurrent mixed-mode data collection

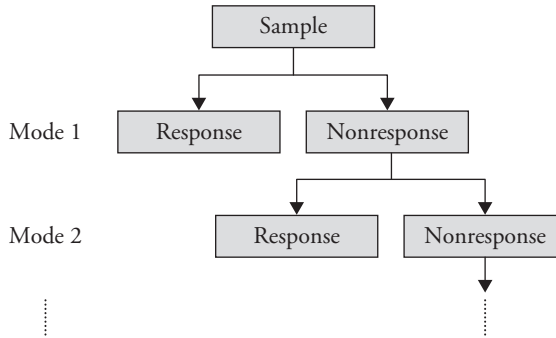


FIGURE 7.2 Sequential mixed-mode data collection

Obtaining a high response rate can be a reason to implement a survey with a mixed-mode design. Nonresponse research has shown that the response of specific groups may depend on the mode of data collection. For example, young people may prefer the web for completing a questionnaire, whereas the elderly appreciate being visited by interviewers. This could imply a web survey for the young and a CAPI survey for the elderly. Of course, this approach requires the age of each person in the sample to be known in advance. This is not always possible.

As mentioned, different strategies can be applied to assign modes to the individuals in the sample. One is to let the respondents choose their own favorite mode, and the other is to pre-assign modes to individuals based on their characteristics. The latter requires these characteristics to be available before data collection starts.

A second mixed-mode approach described by De Leeuw (2005) is the sequential approach. All individuals in the sample are approached using one mode. The nonrespondents are then followed up by a different mode than the one used in the first approach. This process can be repeated for several modes. The sequential mixed-mode approach is illustrated in Figure 7.2.

If the main objective is to keep survey costs as low as possible, a sequential mixed-mode survey could start with a mail questionnaire or a questionnaire on the web. Nonrespondents can, for example, be followed up by CATI. Nonrespondents after CATI could be followed up by CAPI. So the survey starts with the cheapest mode and ends with the most expensive one.

If quality and response rates are of vital importance, one could think of a mixed-mode design starting with CAPI. The nonresponse is then followed up by CATI. Finally, the remaining nonrespondents could be asked to complete the questionnaire on the web.

Another application of a sequential mixed-mode design is a survey in which only one data collection mode is used, but for which recruitment takes place with a different mode. A typical example is a web survey for which the sample is selected from an address register. Selected individuals are sent a letter with a link to the questionnaire on the web. This is a combination of mail recruitment and web data collection.

A mixed-mode survey can take many different forms. It can be a concurrent or sequential design, different modes can be used, and the order of the modes in a sequential design has to be defined. It also makes a difference whether interviewer-assisted modes are used or self-administered modes. All these choices have an impact on the costs of the survey and on the quality of the results. Other factors may play a role in designing a mixed-mode survey, such as

- The complexity of the information being collected
- The time it takes to complete the interview (time burden)
- The sensitivity of the topics covered in the questionnaire.

With respect to data collection, there is a substantial difference between interviewer-assisted modes of data collection (e.g., CAPI and CATI), on the one hand, and self-administered data collection (e.g., web and mail) on the other. Interviewers carry out the fieldwork in a CAPI or CATI mode. There are no interviewers in a self-administered mode. Therefore, quality of collected data may be lower as a result of higher nonresponse rates and more errors in the answers to the questions.

De Leeuw (2008) as well as Dillman, Smyth, and Christian (2009) discuss the differences between various data collection modes. They observe that a positive effect of the presence of interviewers is that they are in control of the interview. They lead the respondent through the interview. They see to it that the right question is asked at the right moment. If necessary, they can explain the meaning of a question. They can assist respondents in getting the right answers to the question. Interviewers can motivate respondents, answer questions for clarification, provide additional information, and remove causes for misunderstanding. All this will increase the quality of the collected data.

The presence of interviewers also can have a negative effect. It will lead to more socially desirable answers for questions about potentially sensitive topics. Giving socially desirable answers is the tendency that respondents give answers that will be viewed as more favorable by others. This particularly happens for sensitive questions about topics like sexual behavior and use of drugs. If a true answer would not make the respondents look good, they will refuse to answer or give a different answer. This phenomenon is described by Tourangeau and Yan (2007). A meta-analysis by De Leeuw (1992) shows that the effects of

socially desirable answers are stronger in interviewer-assisted surveys. Respondents tend to give more truthful answers in self-administered surveys.

EXAMPLE 7.3 A mixed-mode survey of manufacturing firms

Biffignandi and Fabrizi (2006) conducted a sample survey on manufacturing firms. Data collection was based on a multistage and multimode strategy. Respondents were first contacted by telephone. If contact was established, they were offered the choice to complete the questionnaire on the web or to fill in a paper form that was sent by mail or fax.

Reminder strategies were optimized for each data collection mode separately. There were only two reminders. This decision was based on the analysis of experience in previous surveys: Biffignandi et al. (2004) showed that the effect of successive reminders decreased rapidly. The decision was also a compromise between costs and response rates.

Web respondents received two reminders (after 10 and 20 days) by e-mail. Fax and mail respondents were reminded by telephone (after 14 and 28 days). For simplicity, reminder periods were the same for fax and mail respondents.

The data collection procedure started with telephone recruitment. For the respondents, there was a choice of web, mail, or fax. For reminders, e-mail or telephone was used. Figure 7.3 summarized the mixed-mode design of this study.

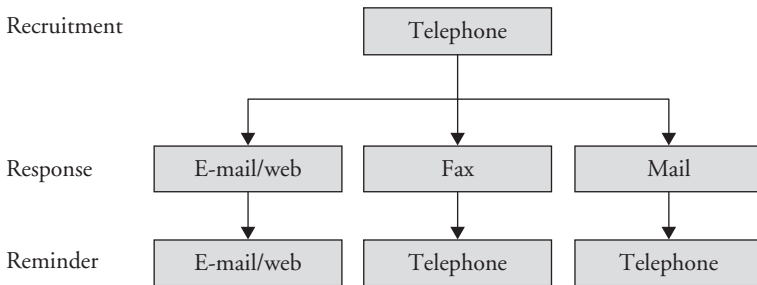


FIGURE 7.3 The mixed-mode design of the manufacturing sector survey

Knapp and Kirk (2003) have compared three types of self-administered data collection instruments: a web questionnaire, a paper questionnaire, and interactive voice response (IVR).

In the case of IVR, a computer system calls the respondents and asks the questions. The respondents answer like in a normal telephone call, after which their answer is processed by means of a voice recognition system, or they use the keypad of their telephone to type an answer. Knapp and Kirk (2003) found no differences among the three modes.

Kreuter, Presser, and Tourangeau (2008) showed that sensitive questions were answered more truthfully in web surveys as compared with CATI surveys. IVR had an intermediate position in this. For nonsensitive questions, the differences between modes were much smaller. This confirms that the mode effects are particularly important for sensitive questions. The presence of interviewers leads to more socially desirable answers to sensitive questions. Another effect is that there is less item nonresponse in web surveys.

A word of warning is necessary when comparing the research results in the literature. Sometimes experiments are conducted on specific populations. For example, Tourangeau et al. studied the effects in a population consisting of recent graduates of one university. They focused on questions that were relevant for that population. The population consisted of well-educated people. This could, for example, mean that they were much more familiar with computers and completing forms on the Internet. So one should be careful in generalizing the results to the general population. Fortunately, Tourangeau and Yan (2007) show that the detected mode effects seem to be consistent with earlier literature on social desirability bias.

Summing up the findings on the differences among web, IVR, and CATI, one conclusion is that there is no best mode.

Each mode has its advantages and disadvantages, with respect to unit nonresponse, item nonresponse, and measurement errors. Because the findings for IVR always seem to take a middle position between web and CATI, only the two extremes are compared:

- With respect to unit nonresponse, CATI surveys have higher unit response rates than web surveys. So CATI is better.
- With respect to item nonresponse, CATI surveys have lower item response rates than web surveys. So web surveys are better.
- With respect to social desirability bias, CATI surveys are more affected than web surveys. So web surveys are better.

Chapter 5 compares web surveys with other types of surveys. Many more aspects are described there for which surveys can differ. Examples are response order effects, acquiescence, and satisficing. These examples as well as other aspects described in Chapter 5 also apply to the various modes in a mixed-mode survey. All this makes it not very easy to design a mixed-mode survey. The following observations can be made:

- Each mode of data collection has its own specific limitations and sources of errors.
- There is no unique, best mixed-mode design.
- The choice of modes may depend on the source of error that is most important for a specific survey.
- There are substantial differences between self-administered modes of data collection and interviewer-assisted modes of data collection with respect to response rates and measurement errors.

- Government household surveys dealing with nonsensitive topics such as work, education, and expenditure may be relatively immune to changes in the mode of data collection.
- Government business surveys that deal with topic like innovation, finance, and employment may also be relatively immune to changes in the mode of data collection.
- The answers to sensitive topics are substantially affected by the survey mode.

7.2.2 WHY MIXED MODE?

It is widely recognized that survey response rates have been declining almost in any country and area of research. See, for example, the international comparison by De Leeuw and De Heer (2002). This trend is mainly due to:

- Increasing noncontact rates. Demographic changes (there are more single-person households), socioeconomic changes (there are more couples with both persons having a job), and new technological developments (mobile phones, Skype, and answering machines) make it increasingly difficult to contact persons that have been selected for the survey.
- Increasing refusal rates. Refusals are growing, because there is an increasing reluctance of the general public to take part in surveys. This may be caused by:
 - *An increase in the number of requests to participate in surveys.* For example, aside from official statistics surveys, many research institutes and commercial marketing research companies are conducting surveys. This is a problem for both household surveys and businesses surveys.
 - *A perceived heavy response burden.* In particular, companies complain about the administrative burden imposed by government. They think they have to fill in too many forms. Although the share of statistical surveys in this administrative burden is usually very small, companies often have the perception of a high response burden.
 - *A decreasing trust in surveys.* There is already much information available in registers and other administrative sources. There are also many surveys that collect even more data. There are good surveys, but there are also many bad surveys. It becomes more and more difficult for people to distinguish the good from the bad. All this reduces trust in surveys.

Decreasing response rates often have a negative impact on survey results. There are many examples where nonresponse leads to biased estimates of population characteristics. The lower the response, the larger the bias will be. Therefore, it is important to reduce nonresponse as much as possible. This requires extra efforts during data collection. One can provide extra training for interviewers, more and better supervisors, and incentives for participation in the survey. One should bear in mind, however, that these efforts increase the costs of the survey. In addition, recent literature emphasizes that a researcher should not just focus on raising response rates if this means ignoring other phenomena that may reduce survey quality.

Mixed-mode surveys are becoming more and more popular. Probably the two most important reasons for considering such a survey approach are reducing survey costs and keeping response at an acceptable level. Another factor that may play a role are the problems with current sampling frames for telephone surveys. These problems are making it almost impossible to reach certain parts of the population (young people with only a mobile phone). Some of these aspects are discussed in some more detail in the following sections.

7.2.2.1 Response Rates. In many surveys it is assumed that offering the right mode of data collection to the selected individuals will increase the response rates. This can be achieved by letting respondents choose their own mode or by predicting before the start of the fieldwork which mode of data collection is most fit for the specific individuals or businesses.

The response rates are not the same for each mode of data collection. The highest response rates are obtained for face-to-face interviewing, closely followed by telephone interviewing. Response tends to be lower for self-administered modes. The lower the response is, the higher the risk of substantial nonresponse bias.

Note that in Example 7.2 the mixed-mode design did not lead to an increased response rate. The web survey mode is not an easy mode from the point of view of response rates. The process of starting up a computer, connecting to the Internet, typing in the address of the survey website, entering the unique respondent code and answering the questions may be perceived as more cumbersome than filling in a paper form. Indeed, first experiences show that the response rates of web surveys based on probability-based sampling are not higher than around 40%.

7.2.2.2 Costs. If the focus of the survey design is on reducing the costs of the survey, a mixed-mode survey may be a way to realize this. In particular, an expensive interviewer-assisted survey can be replaced by a mixed-mode survey where one or more of the modes is a self-administered one.

A mixed-mode design aimed at keeping costs at a low level is a sequential design that starts with the cheapest mode, for example, mail or web. Non-respondents are followed up with a less cheaper mode (CATI). The final mode could be the most expensive one. As described in Example 7.2, Beukenhorst and Wetzels (2009) showed the direct data collection costs of the Dutch Safety Monitor could be reduced by 40% in this way.

Several authors, for example, Hochstim (1967), Mooney, Giesbrecht, and Shettie (1993) and Voogt and Saris (2005), argue that a sequential mixed-mode design offers advantages with respect to both response rates and costs. Indeed, De Leeuw (2005) shows that conducting a follow-up by telephone after an initial mail questionnaire mode improves response rates. The situation is, however, not clear if the web is included as one of the modes. Beukenhorst and Wetzels (2009) showed that sequential mixed-mode did not increase the response rate of the Safety Monitor.

7.2.2.3 Data quality. Chapter 5 of this handbook compares different modes of data collection. It is shown there that each data collection mode has its own advantages and disadvantages with respect to data quality. Therefore, it depends on the actual mixture of modes what the combined effect on the quality of the survey data will be. Some effects are summarized here:

- *Response order effects.* Respondents in interviewer-assisted surveys show a preference for the last options in the list of answer options of closed questions (recency effect). In self-administered modes of data collection, there is a preference for the first options in the list (primacy effect).
- *Acquiescence.* Respondents tend to agree with statements in questions, regardless of their content. They simply answer “yes.” There is less acquiescence in self-administered surveys than in interviewer-assisted surveys.
- *Status quo endorsement.* If respondents are asked to give their opinion about changes, they tend to select the option to keep everything the same. There seems to be less status quo endorsement in interviewer-assisted surveys.
- *Nondifferentiation.* This occurs when respondents have to answer a series of questions with the same set of response options. Respondents tend to select the same answer for all these questions irrespective of the question content. This is a form of satisficing. There is more nondifferentiation in self-administered surveys.
- *Answering “don’t know”.* This is a form of satisficing where respondents choose this answer to avoid having to think about a real answer. Not making it possible to answer “don’t know” may also cause measurement errors as respondents not knowing the answer are forced to give one. The way “don’t know” is treated in a survey may depend on the mode of data collection.
- *Arbitrary answer.* Respondents may decide to just pick an arbitrary answer in order to avoid having to think about a proper answer. They may also give an arbitrary answer if giving the proper answer is considered undesirable. This behavior is sometimes also called “metal coin flipping.” This phenomenon typically occurs in web surveys for check-all-that-apply questions.
- *Socially desirable answers.* This is the tendency that respondents give answers that will be viewed as more favorable by others. This particularly happens for sensitive questions. If a true answer would not make the respondents look good, they will refuse to answer or give a different answer. The literature shows that the effects of socially desirable answers are stronger in interviewer-assisted surveys. Respondents tend to give more truthful answers in self-administered surveys.

These phenomena have different effects in different data collection modes. As a consequence, the same questions may be answered differently in different modes. In the context of mixed-mode surveys, they are called *mode effects*.

It should be noted that changing a survey design, for example, from a single-mode (like a face-to-face survey), to a mixed-mode survey (including the web as

one of the modes), will lead to changes in mode effects. This will hinder comparing statistics over time. Observed changes in figures may be caused by real changes in the phenomena measured, but these changes may also be caused by changes in mode effects.

7.2.2.4 Coverage Problems. Coverage errors occur if the target population of the survey does not coincide with the sampling frame used. Undercoverage is a type of coverage problem that may have serious consequences. *Undercoverage* denotes the phenomenon that elements of the target population are not represented in the sampling frame. Therefore, these elements cannot be selected in the sample for the survey. If elements in the sampling frame differ from those not in the sampling frame, the survey may produce incorrect figures.

Undercoverage occurs in a telephone survey if the sample is selected from a telephone directory. People with unlisted numbers are excluded from the survey and often people with only a mobile phone. Undercoverage may also occur if the Internet is used to select persons for a survey and the target population is wider than just those with access to Internet. See Chapter 8 for an extensive description of undercoverage in web surveys.

A mixed-mode survey may help to reduce undercoverage problems. One approach is to divide the population into subpopulations and to assign to each subpopulation the data collection mode that is most appropriate for that group. For example, if a sample of individuals is selected from a population register, the age of all selected individuals is known. Therefore, it can be decided to approach young people using the web and the elderly by means of a visit of an interviewer. Another example is a survey in which first everybody is asked to complete the questionnaire on the Internet. Next, nonrespondents and those without Internet are given the possibility to complete a paper questionnaire.

It should be noted that the extent and effects of undercoverage may change if the survey design is modified. There can be groups of people in the new survey that were not included in the previous survey. So changes over time can be real changes in figures, but they can also be an error caused by a change in coverage. In fact, the surveyed population is not the same as before and therefore the statistics can be different.

7.2.2.5 Selection Errors. In some mixed-mode surveys, it is left to the respondents to choose the mode of data collection. For example, respondents receive an invitation letter in which they are invited to participate in the survey. If they agree, they can choose to complete the questionnaire on the Internet or they can fill in the paper questionnaire that is included in the letter. Another example is a sequential mixed-mode survey where the selected persons are first asked to complete the questionnaire on the Internet. Next, those not responding are called by telephone.

The effect of these approaches is that specific groups choose specific modes. If the collected data are compared for the different modes, it is not clear to what cause differences can be attributed. On the one hand, differences can be caused by mode effects, and on the other, differences may be caused by real differences

between the two groups. Unfortunately, these effects cannot be disentangled. Hence, it is unclear whether an observed difference is a “real” difference or just an artifact caused by mode effects.

7.2.2.6 Cognitive Efforts. Tourangeau (1984) introduced his cognitive response model to describe the process of answering survey questions. This model helps in explaining mode effects. More details can be found in the studies by Roberts (2007), Bowling (2005), Jäckle, Roberts, and Lynn (2010), and Ariel et al. (2008). Only a short overview of this model and its relation to mode effects is given here. The cognitive response model consists of four steps:

1. *Comprehension.* Respondents attempt to understand the meaning of the question. Comprehension is influenced by the presentation of the questions. Aural presentation by an interviewer may lead to recency effects and visual presentation in a self-administered mode to primacy effects. The presence or absence of interviewer also can have an effect. For example, an interviewer can always help to explain the meaning of a question. If the respondent has a paper questionnaire form, he or she can page through it and look to other questions. This may help understanding what the survey is all about.
2. *Retrieval.* To be able to answer the question, respondents must collect relevant information. Their long-term memory is an important, and sometimes the only, source for this. The process of retrieving can differ substantially across modes. In the case of self-administered modes of data collection, respondents can take as long as they want or need to perform this task. In the case of interviewer-assisted data collection, respondents will feel pressure to answer as quickly as possible.
3. *Judgment.* The respondents assess whether the retrieved information is adequate for answering the question. They do this by comparing the available information with the meaning of the question. The presence of interviewers can have a positive effect. They may help if the respondents are unable to reach a positive judgment. They do this by making suggestions or probing.
4. *Response.* The respondents report or record the answers to the question. To do this they have to put their answer in the proper format, for example, by selecting the right answer option of a closed question. In the case of interviewer-assisted data collection, respondents may decide to change their initial answer if they consider it socially undesirable.

Mixed-mode surveys have many methodological problems. There are advantages and challenges. The effectiveness of mixed-mode surveys depends on its design (concurrent, sequential, or a combination) and the data collection modes used. The main question always is whether a mixed-mode survey allows for accurate, unbiased estimation of population characteristics. The next section is devoted to this.

7.2.3 METHODOLOGICAL ISSUES

There are many unanswered questions with respect to the reliability of the outcomes of mixed-mode surveys. Mixed-mode surveys can be defined in so many ways, and there are so many phenomena that may affect the outcomes that it is impossible to provide simple, general answers. There have been many studies, but their conclusions almost always apply to specific situations (i.e., specific target populations or specific survey designs). The number of studies related to mixed-mode surveys with the web as one of the data collection modes is still limited. In most of these cases, web surveys are compared with mail surveys.

There is no general theoretical model (yet) for mixed-mode surveys that can help to take a decision about the best survey design. Therefore, still a lot of experiments are needed to collect information about this type of survey. This section attempts to answer three basic methodological questions with respect to mixed-mode surveys:

- *How to design a questionnaire for a mixed-mode survey.* Should the same questionnaire be used in each mode, or should there be versions that are optimal for each specific mode?
- *How to mix modes.* Is it better to have a sequential or a concurrent design? And which modes should be used in a specific survey context?
- *How to compute response rates.* How can the performance of various modes (in terms of response rates) be compared? And how can a mixed-mode survey be compared with a single-mode survey?
- *How to make statistical inference for a mixed-mode survey.* How can estimation procedures account for mode effects? Is it possible to disentangle mode effects from selection effects and other effects?

These questions will be addressed in the remaining subsections.

7.2.3.1 Questionnaire Design. Each mode of data collection has its advantages and disadvantages. The effects of different phenomena may vary over modes. This means that the same question may be answered differently. Consequently, observed differences in figures may be not be “true” differences but deviations caused by measurement problems. These mode effects should be avoided as much as possible.

The survey researcher is faced with the question of whether to use the same questionnaire across modes. If he does, there are mode effects. He could also use a different questionnaire for each mode. These questionnaires must be designed such that, although they are different, they measure the same concepts. The questionnaires must be cognitively equivalent. This is not so easy to realize. Three approaches are discussed here to deal with this problem.

The first approach is the *unimode approach* as proposed by Dillman (2007). The idea is to use the same questionnaire in each mode but to define the questions in such a way that mode effects are minimized. Some examples of his guidelines are as follows:

- The text of a question must be the same across modes.
- The number of answer options of a closed question must be kept as small as possible.
- The text of the possible answers to a closed question must be the same across modes.
- The order of the answer options to a closed question must be randomized.
- Include all answer options also in the text of a closed question.
- Develop equivalent instructions for skip patterns.

It may not be easy to develop a questionnaire that completely satisfies all unimode guidelines. In particular for attitudinal questions, it may turn out to be necessary to define mode-dependent versions of questions.

A properly designed unimode questionnaire should remove the mode effects related to question interpretation. However, this approach cannot take advantage of the specific features that every mode offers. For example, a researcher may have to abandon the idea of randomizing the order of the answer options of a closed question because it is not possible to do this in the paper questionnaire mode. Another example refers to displaying instructions about how to answer questions: In mixing paper web mode, the web questionnaire cannot use pop-up windows because this is not possible in a paper questionnaire.

Another potential problem is the mode effects caused by the presence or absence of interviewers. The unimode approach may not be able to remove these effects completely. Differences may particularly remain for sensitive questions.

The second approach to diminish mode effects is designing mode-specific questionnaires. Each questionnaire should be optimal for its corresponding mode. Optimal means that the questions are defined such that the answers given are as close as possible to the “true” value.

As an example, consider answering a factual question in a CAPI survey and a web survey. In the case of a CAPI, there is always an interviewer who can assist the respondent in understanding and answering the question. In the case of a web survey, there is no interviewer assistance. The respondents are on their own. Nevertheless it is possible to develop some kind of interactive help system for web surveys. There could be help-buttons on the screen giving access to additional information about the question. It is even possible that an animated interviewer appears.

It may not be easy to design optimal mode specific questionnaires. It will require a fair amount of experimentation with different formats to obtain the best one. Moreover, the best question in one survey context may not be the best question in another survey context.

If a survey is repeated at regular intervals, as often happens at national statistical institutes, there is also a maintenance challenge. Changes in the survey must now be implemented and tested in several questionnaire forms instead of in one.

A third approach to dealing with mode effects is to identify one primary mode of data collection. This is considered the most important mode of data collection. It is the benchmark for all other modes of data collection. The

questionnaire is optimized to get the best answers in the primary mode. The questionnaires for other modes must be designed such that there will be no mode differences with the primary mode.

Pierzchala (2006) determined the main factors responsible for most mode effects. He identified three dimensions. These are:

1. Presentation—aural versus visual presentation
2. Administration—self-administered versus interviewer-assisted
3. Behavior—dynamic versus passive questionnaires.

Questionnaires for computer-assisted interviewing are usually dynamic. They have forced routing and perform consistency checks. Paper questionnaires are passive.

The first and third dimensions relate to the way in which information is transmitted. The second is related to the medium used for this. Based on these three dimensions, Pierzchala (2006) introduces *disparate modes* as modes that differ on at least one of these dimensions. Furthermore, the larger the degree of disparity is, the higher the risk of mode effects will be. See Table 7.2.

CAPI and CATI are similar in presentation, administration, and behavior of the questionnaire. The degree of disparity of the other modes of data collection is indicated with respect to both CAPI/CATI. Web surveys have a dynamic questionnaire. Presentation is visual instead of aural and web surveys are self-administered. Therefore, the degree of disparity between web surveys and CAPI/CATI surveys is 2. Mail surveys share none of the aspects with CAPI/CATI, which results in a degree of disparity equal to 3. Mail surveys and web surveys are similar in self-administration and visual presentation, but mail surveys have a passive questionnaire. Therefore, their degree of disparity is 1.

It can be concluded from Table 7.2 that the largest mode differences can be expected for a mixture of mail with CAPI or CATI, followed by a mixture of web surveys with CAPI or CATI. A combination of mail and web surveys has a reduced risk of mode effects. Combining CAPI and CATI is the safest option to avoid mode effects.

7.2.3.2 How to Mix Modes? There are no specific rules on how to mix modes. There is no best solution. Experimental research is still ongoing. The aim is to find out evidence for the advantages and disadvantages of different

TABLE 7.2 Degree of disparity among data collection modes

Mode combination	CAPI/CATI	Mail	Web
CAPI/CATI	—	3	2
Mail	3	—	1
Web	2	1	—

approaches. Design decisions should be made taking into account the available evidence. Moreover, such decisions will also depend on the specific survey context.

Some practical aspects are discussed here. Different approaches are compared on the basis of costs and response rates. The first aspect is the choice between a sequential mixed-mode survey and concurrent mixed-mode survey,

A concurrent design can be implemented in two ways. The first way is to let respondents decide which mode to use for completion of the questionnaire. The second way is that the survey researcher decides beforehand which mode is best for every respondent. Both approaches have the advantage that they reduce undercoverage errors. The second approach gives the survey researcher control over assigning modes to groups. A proper design makes it possible to distinguish mode effects from selection effects. The first approach (i.e., respondents decide) does not provide this type of control. Therefore, mode effects and selection effects may be entangled. This means that undercoverage errors are replaced by selection errors. It is up to the researcher which error to prefer in a practical situation.

A simple sequential mixed-mode design is one in which recruitment takes place in one mode and the actual data collection in another mode. For example, sampled individuals are approached by telephone and asked to participate in a web survey. This design has the advantage of the high response rates of the telephone recruitment. Use of a single mode for data collection avoids mode effects.

One step further is a design in which people are approached in a single, interviewer-assisted mode (face-to-face or telephone). At the end of the interview, the respondent is invited to participate in a panel. Such a panel may have different modes of data collection. Then there will be not only mode effects within the panel but also between the recruitment interview and the panel. The situation is complicated by the time lag between recruitment and panel. This makes it impossible to distinguish real changes over time from mode effects.

It is a common procedure to follow up with nonrespondents in a survey as this helps to increase response rates. The researcher has the choice to use the same mode for the follow-up or to use a different mode. For example, nonrespondents can be called by telephone or sent a letter in an attempt to encourage them to complete a web survey questionnaire. If the follow-up is not meant for persuading nonrespondents, but also some additional data is collected, there may be mode effects.

In recruitment, follow-up, and data collection, several modes can be used. Recruitment and follow-up focus on communication and not on real data collection. Therefore, De Leeuw (2005) uses the term *mixed-mode system* or *multi-mode system* to denote either communication with respondents (mixed-mode communication) or data collection (mixed-mode data collection).

Statistics Netherlands carried out some experiments with more complex mixed-mode designs. Some of the findings are described in Examples 7.4 and 7.5.

EXAMPLE 7.4 The ICT Survey Pilot

Statistics Netherlands carried out several pilot studies with mixed-mode surveys in the period from 2005 to 2007. The objective of these studies was to determine mixed-mode designs that, on the one hand, reduced data collection costs, and, on the other, at least preserved the quality of the collected data.

One of these pilots was conducted for the ICT survey, which collects information on the use of computers and Internet in households and by individuals. The regular ICT survey was a CATI survey. It was fairly expensive. It also suffered from undercoverage because the sample was selected from the telephone directory. Households with unlisted numbers and mobile-only households could not be selected.

One objective of this pilot was to find out what level of response could be obtained. Another objective was to establish whether people without Internet would be properly represented in a mixed-mode survey with the web as the most important mode. Therefore, respondents had the possibility of completing the questionnaire on paper. To prevent those with Internet from responding by paper, the paper questionnaire was not included in the invitation letter. People had to apply for the paper form by returning a stamped return postcard. The design of this pilot is shown in Figure 7.4.

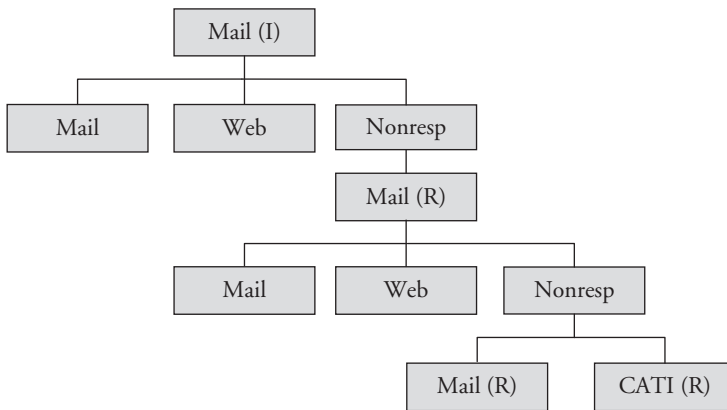


FIGURE 7.4 The mixed-mode design of the ICT survey pilot

The sample was selected from the population register. So there was no undercoverage. All individuals in the sample received an invitation letter by mail. The letter contained the Internet address of the survey and a unique log-in code.

After one week, a postcard was sent to all nonrespondents with a reminder to complete the survey questionnaire, either by web or mail.

Two weeks after receipt of the invitation letter, the remaining nonrespondents were approached again. Part of these nonrespondents received a reminder letter, and another part was called by telephone (if a telephone number was available). The telephone call was just to remind the nonrespondents and did not replace the paper/web questionnaire form.

It turned out the postcard reminders worked well. Each time they were sent, there was a substantial increase in response. The telephone reminder did not work as well as the postcard reminder. Of the people that promised by telephone to fill in the form, only 40% actually did so.

EXAMPLE 7.5 The Safety Monitor Pilot

Statistics Netherlands has conducted an experiment with its Safety Monitor to determine whether a mixed-mode survey can replace a CAPI or CATI survey without affecting the quality of the results. The Dutch Safety Monitor is an annual survey. It measures the actual and perceived safety of the people in the country.

The sample for this survey is selected from the population register. The old Safety Monitor applied two data collection modes. If a telephone number could be found, sampled persons were approached by CATI. If this was not the case, they were approached by CAPI.

The design of this pilot is shown in Figure 7.5. There were four modes of data collection in the pilot for the new Safety Monitor. All

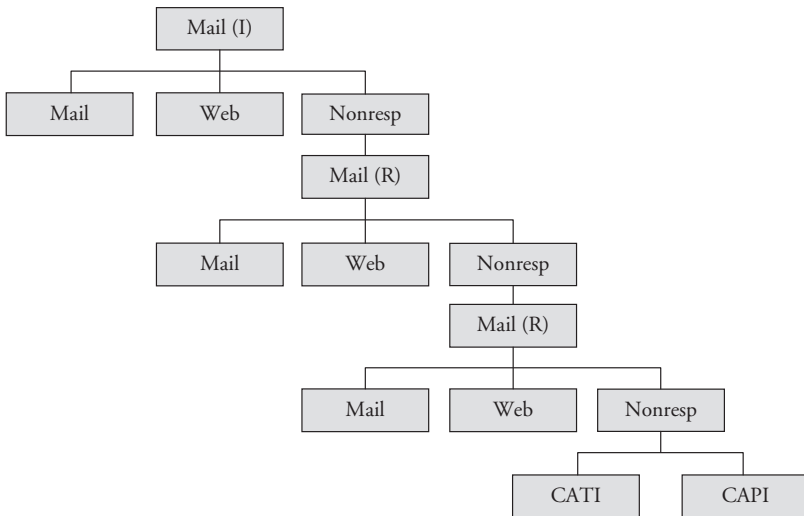


FIGURE 7.5 The mixed-mode design of the Safety Monitor

sampled individuals received a letter in which they were asked to complete the survey questionnaire on the Internet. The letter also included a postcard that could be used to request a paper questionnaire. Two reminders were sent to those that did not respond by web or mail. If still no response was obtained, nonrespondents were approached by means of CATI, if a listed telephone number was available. If not, these nonrespondents were approached by CAPI.

The response rate of the new survey turned out to be 59.7%. The response rate for the old survey was 63.5%. So including the web as one of the modes did not increase the response rate. More than half respondents (58.0%) in the new surveys completed the questionnaire with a self-administered data collection mode (web or mail). Therefore, the costs of the survey were much lower. Interviewers were involved in only 42.0% of the cases.

7.2.3.3 How to Compute Response Rates? A mixed-mode survey is sometimes designed with the objective of reducing survey costs. Another objective can be to increase response rates. Whatever the objective of the design, it is always important to compute the response rate, as it is often considered a quality indicator: the higher the response rate, the better the quality of the survey. To be able to compare response rates of modes, and to be able to compare response rates of different surveys, standardized and consistent definitions must be used.

The basic formula for the response rate (RR) is

$$(7.1) \quad RR = \frac{n_R}{n_E} = \frac{n_R}{n_R + n_{NR}},$$

where n_R is the number of (eligible) respondents, n_E is the total number eligible units (individuals, households, businesses, and so on) in the sample, and n_{NR} is the number of (eligible) nonrespondents. Eligible units are units that belong to the target population and have been selected in the sample. In practice it may be difficult to compute n_E as it is not always possible to determine whether nonrespondents are eligible.

It is important to distinguish different causes of nonresponse. Different causes may have different effects on the outcomes of the survey and may therefore require different treatment. The three basic causes of nonresponse are no contact, refusal, and not able. Taking into account the response rate definitions proposed by AAPOR (2009) and Lynn et al. (2002), expression (7.1) can be rewritten as

$$(7.2) \quad RR = \frac{n_R}{n_R + n_{IC} + n_{NC} + n_{RF} + n_{NA} + n_{OT}}.$$

Respondents may break off the completion of the questionnaire. This typically happens in web surveys. Respondents get tired or bored of filling in the form and just stop in the middle. The number of such cases is denoted by n_R . Partially completed questionnaires are treated as a nonresponse in many surveys. That is why this quantity is included in the denominator and not in the numerator of definition (7.2).

The quantities n_{NC} , n_{RF} , and n_{NA} denote the number of cases of nonresponse caused by noncontact, refusal, and not able, respectively. There is a problem with the number of noncontacts n_{NC} . This must be the number of eligible noncontacts. However, this number cannot be determined because there was no contact. It is not unlikely that some of the noncontacts may be cases of undercoverage, and those should not be included. In practice, an estimate of n_{NC} will be used.

The quantity n_{OT} denotes the number of other unprocessed cases. This could include units of which the eligibility is unknown. At Statistics Netherlands, also another cause of nonresponse is distinguished: administrative nonresponse. These cases are not processed by interviewers because their workload is too high, which may happen if they have to handle many difficult cases in a limited amount of time. Another reason can be a temporary illness of an interviewer.

In the case of interviewer-assisted surveys, it is usually possible to determine the cause of nonresponse. Interviewers observe this and record the results of their efforts as noncontact, refusal, or not able. This is much more difficult to do in self-administered surveys (web and mail). It is only observed that questionnaire forms are not returned or completed. The reason why they are not returned or compiled is not known. There can be many reasons. As an example, here are some reasons for nonresponse in web surveys:

- There are various forms of *noncontacts*. It depends on the way in which sample persons are selected. If the sampling frame is a list with e-mail addresses, noncontact occurs if the e-mail with the invitation to participate in the survey does not reach a selected individual. The e-mail address may be wrong or the e-mail may be blocked by a spam filter. If the sampling frame is a list of postal addresses and letters with an Internet address are sent to selected units, noncontact may be caused by units not receiving the letter. If recruitment for a web survey takes place by means of a face-to-face or telephone survey, noncontact can be from respondents being not at home or not answering the telephone.
- Nonresponse from *refusal* can occur after contact has been established with a sampled unit. Refusal to cooperate can have many reasons. In business surveys, a too high perceived or actual response burden, having strategies against surveys, and a lack of interest in providing data can be factors leading to refusal to cooperate. In household or individual surveys, people may not be interested, they may consider it an intrusion of their privacy, they may have no time, and so on. Sometimes a refusal can be temporary. In this case, it may be attempted to make an appointment for another day and/or time. But often a refusal is permanent. If sample individuals for a web survey are contacted by an e-mail or a letter, they may postpone and forget to

complete the questionnaire form. This can be considered a weak form of refusal. Sending a reminder helps to reduce this form of nonresponse.

- Nonresponse from *not able* is a type of nonresponse where respondents may be willing to respond but are not able to do so. Reasons for this type of nonresponse can be, for example, illness, hearing problems, or language problems. If a letter with an Internet address of a web questionnaire is sent to a sampled person, this person receives the letter, and he/she wants to participate in the web survey, but does not have access to the Internet, this can also be considered a form of nonresponse resulting from not able.

If there are self-administered-modes in a mixed-mode survey, and one wants to compare response rates over modes, only simple response rates can be computed (without taking into account the cause of nonresponse).

Suppose that the mixed-mode survey consists of M concurrent modes and that the survey researcher has preassigned modes to sampled units. Let $n_E^{(h)}$ denote the number of eligible elements in mode h , for $h = 1, 2, \dots, M$. And let $n_R^{(h)}$ be the number of respondents in mode h . Then the response rate in mode h is equal to

$$(7.3) \quad \text{RR}^{(h)} = \frac{n_R^{(h)}}{n_E^{(h)}},$$

for $h = 1, 2, \dots, M$. Consequently, the overall response rate of the survey is equal to

$$(7.4) \quad \text{RR} = \frac{n_R}{n_E} = \frac{\sum_{h=1}^M n_R^{(h)}}{\sum_{h=1}^M n_E^{(h)}} = \sum_{h=1}^M \frac{n_E^{(h)}}{n_E} \text{RR}^{(h)}.$$

So the overall response rate is the weighted mean of the mode response rates.

In the case of a concurrent mixed-mode survey where people choose their own mode of data collection, it is not possible to compute mode response rates, as the number $n_E^{(h)}$ of eligible units for each mode is unknown. Of course, the overall response rate can be computed as $\text{RR} = n_R/n_E$. Also the rate $n_R^{(h)}/n_E$ can be computed but this rate is not comparable with the response rate that would have been obtained if all respondents were approached in this mode.

In the case of a sequential mixed-mode design, the mode-specific response rates and the overall response rates can be computed. Note that for this design, the number of eligible elements in a mode is equal to the number of nonrespondents that remains after the previous mode. So

$$(7.5) \quad n_E^{(h)} = n_E^{(b-1)} - n_R^{(b-1)} = n_E - \sum_{i=1}^{b-1} n_R^{(i)}.$$

EXAMPLE 7.6 Computing response rates

The local authorities of the town of Mudwater in the country of Samplonia conduct a survey about living conditions in the town. It is a simple mixed-mode survey. All people of the age 18 years and older are sent a letter with the invitation to complete the questionnaire on the Internet. After two weeks, all nonrespondents are called by telephone in an attempt to fill in the questionnaire in a telephone interview.

The sample size is 1,600 persons. The number of persons completing the form on the Internet is equal to 672. The number of respondents in the telephone mode is equal to 278.

The overall response rate of the survey is $100 \times (672 + 278) / 1,600 = 59.4\%$. The response mode in the web mode is $100 \times 672 / 1,600 = 42.0\%$. After the web mode, there remain $1600 - 672 = 928$ persons. These are the eligible persons for the telephone mode. The response rate of the telephone mode is $100 \times 278 / 928 = 30.0\%$.

The importance of using a uniform, standardized response rate definition is stressed once more.

Only then are response rates comparable across studies and survey modes used in different analyses. See also Shih and Fan (2007).

Many survey organizations consider conducting mixed-mode surveys in which one of the data collection modes is the web. The idea is that this will substantially reduce the survey costs, as no interviewers are involved. At the same time the overall response rate must remain at an acceptable level. This is reason for concern as some experiments with web surveys show that response rates are not very high.

Some studies have looked at the response rates in mixed-mode surveys. The findings seem not to be consistent. Quigley et al. (2000) describe some studies showing increased response rates in mixed-mode surveys. Dillman, Clark, and West (1995) found that providing alternative response modes does not necessarily improve response rates. Griffin, Fischer, and Morgan (2001) reported a lower response rates when changing from mail to mixed mode. Maier (2005) found a response rate of 62% in a mail survey option, but Irani, Gregg, and Telg (2004) obtained a response rate of 78% in a web survey option.

The conclusion, at present, is that there is no clear evidence of higher response rates in mixed-mode surveys. This does not mean that this approach has no advantages, because many survey quality factors should be considered in choosing the survey mode. For example, in considering changing to a mixed-mode survey with the web as one of the modes, the response rate of the web mode is of crucial importance. The higher the response rate that can be obtained for the web mode, the more cost savings can be realized.

Shih and Fan (2007) carried out a meta-analysis of a large number of mixed-mode surveys. They observed a preference of the mail survey mode over the web

survey mode. The response rate for mail survey modes was on average 14% higher than for web survey modes. However, if respondents were offered both the mail and the web option, there were no systematic differences in response rates. They suggest offering the other mode in a follow-up of nonrespondents of the one mode.

How respondents choose the response mode in a mixed-mode survey remains unclear. Factors affecting the choice seem to be as follows:

- Technological background. However, Zhang (2000) suggests the choice not merely depend on respondents' technological backgrounds or on their access to the web. It turns out that experienced and frequent Internet users often choose to reply by mail or fax.
- The modes offered.
- Delivery format (i.e., in which format is the questionnaire initially offered).
- Mode delivery order (i.e., simultaneous mixed-mode or sequential mixed-mode).
- Type of target population.
- Use of incentives.
- Deployment of follow-up reminders.

There are other aspects also playing a role in the choice of modes to mix. These aspects are as follows:

- Budget restrictions.
- Timeliness.
- Available infrastructure for implementing mixed-mode designs.
- Data quality.
- Specific survey content (types of questions, length of survey, complexity of the questions, need for visual aids).

Roberts (2007) and Biemer and Lyberg (2003) discuss factors influencing the decision process in choosing modes. They acknowledge that choosing an optimal design is especially difficult in situations where there are many options.

The fact that modes vary with respect to factors such as the cost and speed of fieldwork, their suitability for administering different types of questionnaires and their impact on data quality means that, in principle, mixing modes allows the researcher to minimize both the costs and the errors associated with any given single-mode approach. However, although the mixed-mode designs may help in reducing survey costs, due to undercoverage, nonresponse, and specific forms of measurement error, a risk of mode effects remains.

7.2.3.4 Inference. It is also the objective of a mixed-mode survey to produce reliable and accurate estimates of population characteristics. The question is whether traditional estimation methods can be used like those described in Chapter 3, or whether new methods of inference should be developed. One of

the potential problems of a mixed-mode survey is mode effects: the same question is answered differently in a different mode. Mode effects may cause estimates to be biased.

If Dillman's unimode guidelines have been applied, there may be no mode effects. Then the mixed-mode survey can be treated as a single-mode survey. However, if there is a serious risk of mode effects, an estimation method must be used that corrects for these effects. The fear of mode bias may cause researchers to decide against using mixed-mode surveys. In principle, however, there is no reason why mixed-mode surveys, based on samples from different sampling frames, and selected with possibly different sampling designs, cannot be designed in such a way that overall sample representativity is enhanced, at the same time reducing the cost per interview and the time required to complete the fieldwork. Given the developments in society such as the decreasing response rates, increased pressure to reduce the response burden, and demands for reducing survey costs, a mixed-mode survey simply turns out to be the way to go.

EXAMPLE 7.7 Estimation effects in mixed-mode surveys

This small artificial example shows what can happen in a mixed survey if there are selection effects and mode effects.

The target population consists of two age groups: the young and the elderly. The two groups are of equal size. The objective of the survey is supposed to be estimation of the percentage of voters on the New Internet Party (NIP) at the next elections. Among the young, the population percentage is 70%, and for the elderly, it is 10%. So the overall percentage of voters for the NIP is $0.5 \times 70.0 + 0.5 \times 10.0 = 40.0\%$.

The young are more inclined to participate in a web survey. Their response probability is 0.8, whereas for the elderly, it is only 0.2. It is the opposite for a face-to-face survey: The response probability for the young is 0.2, and for the elderly, it is 0.8.

Suppose a web survey of size 1,000 is conducted. Then the expected value of the percentage of voters for the NIP in the sample would be

$$\frac{(400 \times 70) + (100 \times 10)}{500} = 58\%.$$

This percentage is much higher than the population percentage of 40%. This is not surprising as this survey design leads to an overrepresentation of young people, who typically vote for the NIP.

Suppose a face-to-face survey is conducted. Then the expected value of the percentage of voters in the sample would be

$$\frac{(100 \times 70) + (400 \times 10)}{500} = 22\%.$$

This percentage is much lower than the population percentage of 40%. Again this is not surprising as this survey design leads to an underrepresentation of young people, who typically vote for the NIP.

Now suppose a sequential mixed-mode survey is conducted. First the web is offered. Nonrespondents are visited for a face-to-face interview. It is assumed that probability to participate in the face-to-face mode after nonresponse in the web mode remains 0.2 for the young and 0.8 for the elderly. The expected value of the percentage of voters in the sample would be

$$\frac{(400 \times 70) + (20 \times 70) + (100 \times 10) + (320 \times 10)}{840} = 40\%.$$

The response rate goes up from 50% to 84%. Moreover, the estimator is now unbiased.

Up until now there were only selection effects and no mode effects. Now a mode effect is introduced. It is assumed that voting for the NIP is a sensitive topic. Therefore, 20% of the NIP voters will select a different party in the face-to-face survey. They give a truthful answer in the web survey.

The outcome for the web survey will not change. It remains 58%. The expected value of the percentage of NIP voters in the sample is now equal to

$$\frac{(100 \times 56) + (400 \times 8)}{500} = 17.6\%.$$

So the percentage drops from 22.0% to 17.6% in the face-to-face survey. In the case of a mixed-mode design, the expected value of the percentage of NIP voters becomes

$$\frac{(400 \times 70) + (20 \times 56) + (100 \times 10) + (320 \times 8)}{840} = 38.9\%.$$

The mixed-mode survey estimator has a bias now. Still the estimates are much better than would have been the case for a single-mode web survey or a face-to-face survey.

Given that different modes are likely to lead to differences in coverage and in response rates for subgroups, it is advised to attempt to correct for these differences. Weighting adjustment can do this. See Chapter 10 for details. Weighting adjustment is useful to correct for the lack of representativeness of the response. If there are only mode effects in a mixed-mode survey, weighting will not help to solve estimation problems. This is illustrated in Example 7.7 where

the composition of the response is representative with respect to age, but the mode effect still causes a bias in the estimator.

Generally speaking, there are two approaches to inference in survey methodology. They are called the design-based approach and the model-based approach. Both approaches are discussed in the context of web surveys by Couper and Miller (2008).

The *design-based approach* is the classic approach to survey sampling as described in Chapter 3. If samples are based on probability sampling, the theory of probability and statistics can be applied. This results in concepts like unbiased estimators and confidence intervals. The design-based approach also can be used for web surveys. For example, if a random sample is selected from a sampling frame and all selected elements are invited to complete the questionnaire on the web, proper inference to the population is possible.

The *model-based approach* assumes the existence of some kind of model for the relationships between the survey variables. Based on the available data, the parameters of the model are estimated. Next, the model can be used to predict the values of population parameters. The reliability and accuracy of these predictions usually depend on the validity of the model. Unfortunately it is not always possible to select the validity of the model. If the model is correct, the predictions are accurate. If the model is not correct, the predictions can be seriously biased.

A typical example of the model-based approach in a web survey is recruiting respondents by means of self-selection (see also Chapter 9). The sample is not obtained by means of probability sampling. Selection probabilities are unknown, and therefore, the design-based approach to inference cannot be applied. A way out of this problem is to model participation in the survey by means of response propensities. See Chapter 11 for a detailed description.

The situation can be even more complex for a mixed-mode survey. Alternative models have been proposed. Here is a list of some of them:

- Heckman's selection model. Cobben, Schouten, and Bethlehem (2006) propose a model for inference in mixed-mode surveys that is based on Heckman's selection model. This model simultaneously describes the relationship between target variables and auxiliary variables and between auxiliary variables and response behavior. An interesting aspect of this model is that it distinguished different causes for nonresponse, which gives it more explanatory power. Fitting this model is fairly complex.
- Regression model. Jäckle, Roberts, and Lynn (2010) use a regression model in sample characteristics and the modes are used as explanatory variables. If the mode turns out to be a significant factor once the differences in sample distribution have been accounted for, then there is a mode effect. They conclude that it is extremely difficult to distinguish mode effects from other effects like undercoverage or nonresponse. It is also difficult to find the proper model, as a different model may lead to different conclusions.
- Imputation techniques. Roberts (2005) discusses the possibility of applying imputation techniques. These techniques are mainly used to predict the

correct answer to a question in case of item nonresponse. Using all kinds of auxiliary information, it may be possible to develop imputations models that predict the answer to a question if it was asked in another mode.

- Empirical adjustments. De Leeuw (2005) suggests empirically based adjustment as a feasible approach. It means that an experiment is included in the survey design. This experiment allows for comparing various modes with each other. For example, if the first mode in a sequential mixed-mode design is the web, one might split the sample randomly into a large subsample that has to complete the questionnaire on the web and a small sample that is interviewed by telephone. If there are no nonresponse or undercoverage effects, differences can be attributed to mode effects.

7.2.4 MIXED MODE FOR BUSINESS SURVEYS

Mixed-mode designs can be applied both in business surveys and in household/individual surveys. Self-administered modes of data collection were already used in businesses early in development of information technology. This is not surprising as businesses were the first to use computers at a wide scale. Therefore, coverage was less of a problem.

EXAMPLE 7.8 Fire statistics

An early example of self-administered data collection was the production of Fire Statistics in The Netherlands in the 1980s. As all fire brigades had a microcomputer at that time, data for these statistics could be collected by means of a self-administered questionnaire. Diskettes with electronic interviewing software were sent to all fire brigades. They ran the questionnaire on their MS-DOS computers. The answers were stored on the diskette. After having completed the questionnaire, the diskette was returned to Statistics Netherlands by ordinary mail.

The rapid development of the Internet led to new modes of data collection. Already in the 1980s, prior to the widespread introduction of the World Wide Web, e-mail was explored as a new mode of survey data collection. Kiesler and Sproull (1986) describe an early experiment conducted in 1983.

In the first years of the World Wide Web, use of web surveys was limited by the low coverage of the Internet. Clayton and Werking (1998) describe a pilot carried out in 1996 for the Current Employment Statistics (CES) program of the U.S. Bureau of Labor Statistics. They expected several advantages like the lower costs of a web survey, the quick (almost immediate) response to the questions, and the greater flexibility of web survey questionnaires (they could be offered in a form layout or in a question-by-question approach). The drawback was the limited number of respondents having access to the Internet. Only 11% of CES

respondents had access to the Internet and a compatible browser. This case supports the idea that a mixed-mode approach is needed when using web as a survey tool.

Roos and Wings (2000) conducted a test with Internet data collection at Statistics Netherlands for the construction industry. In fact, this was a kind of mixed-mode survey because respondents could choose from among three modes of completing the form:

- Off-line form. The form was sent as an HTML file that was attached to an e-mail. The form was downloaded, completed off-line, and returned by e-mail.
- On-line form. The Internet address of an on-line web form was sent by e-mail. The form was completed on-line.
- E-mail form. An e-mail was sent containing the questionnaire in plain text. Respondents clicked the reply button, answered the questions, and sent the e-mail back.

A sample of 1,500 companies was invited to participate in the experiment. Overall, 188 companies were willing and able to participate. Of those, 149 could surf the Internet and 39 only had e-mail. Questionnaire completion times of all three modes were similar to that of a paper form. Respondents preferred the form-based layout over the question-by-question layout. The conclusion of the experiment was that web surveys worked well.

The web is increasingly used as one of the modes in mixed-mode business surveys. Meckel, Walters, and Baugh (2005) give an example of such a survey with a concurrent design. The target population consisted of small- and medium-sized companies. Sample companies were sent both a paper questionnaire form and a link to a website. So they could choose between web and paper.

Using the web for self-administered methods of data collected seems an obvious way to go for business surveys. For many surveys, all companies have experience with self-completion of forms. Moreover, companies in specific target populations have access to the Internet. So undercoverage is not a problem. There are additional advantages. One is that companies are asked in many government surveys to provide administrative and financial information. Copying and pasting this information from their systems to a web questionnaire can be a lot easier than having to write it down on a paper form. Another is that checks can be built into the web questionnaire forms. This makes it possible to detect and correct errors while filling in the form. Thus, data quality is improved. It should also be noted that many of the questions asked in business surveys are factual questions. This reduces the risk of mode effects in mixed-mode surveys.

It should be noted that the response to a survey may depend on the organization conducting it (a national statistical institute, a private market research company, or an academic researcher). Statistical organizations of the government usually have a sampling frame that is available (a business register). Moreover, their surveys are often compulsory. So it is easy to select a sample and contact the

businesses in the sample. This simplifies the decision to introduce the web as a data collection mode.

Private market research companies and academic researchers may have different objectives for their surveys, like market and product analysis, customer satisfaction, and so on. A sampling frame is hardly ever available. And even if a proper sample can be selected, then still the information may be lacking to contact the selected businesses. This is the reason why often the studies are conducted on a small target population (where the necessary information can be managed) and why often private companies relate to panel construction for web data collection, instead of sample surveys using a mixed mode. The use of web panels is described in Chapter 12. For this reason, the discussion and examples in this chapter refer to government organization surveys.

The increasing demand for information and the continuing pressure to reduce the administrative burden of companies have encouraged survey organizations all over the world to implement mixed-mode surveys more and more. In the United States, the Bureau of Labor Statistics as well as many federal statistical agencies are conducting mixed-mode surveys where the web is one of the data collection modes. In Europe, several national statistical institutes also have introduced web data collection as a mode. In general, in each country, there is an ongoing process of moving respondents from paper to the electronic/web surveys. Even for the 2010 U. S. Census, there are examples of mixed-mode data collection, including the web.

New ways of data collection are usually welcomed by survey managers, especially when they offer the opportunity to reduce costs or to improve the timeliness of data. The Internet is such an opportunity. Besides the advantages already mentioned, there are other interesting aspects for government survey organizations that conduct large-scale data collection operations, particularly if companies have to participate regularly in several surveys. For example, the U.S. Bureau of Labor Statistics offers companies a common portal or gateway for its web surveys. It is called the “Internet Data Collection Facility” or IDCF. In addition to providing a secure common gateway, the IDCF requires that all survey applications meet internal standards for graphical user interfaces so that on-line questionnaires have the same look and feel. Analogously, Italy and many European countries have constructed or have under construction such portals for business web survey management.

The literature shows that mixing mail and web data collection does not increase response rates, but it only causes a shift from using the paper version of the questionnaire to the web version of the questionnaire.

■ EXAMPLE 7.9 Mixed-mode in the Italian SCI survey

Biffignandi and Zeli (2008) investigated the response rates in an Italian business survey. This was the SCI survey. SCI stands for “Sistema dei Conti delle Imprese.” It is a annual compulsory survey among large

enterprises. The survey is conducted by the Italian Statistical Institute (ISTAT), and it collects information about the economic and financial status of companies. Companies have a choice to respond by paper or use the Internet. Starting in 2003, a electronic questionnaire was delivered to businesses and they had the choice to send their answer on a paper questionnaire or via the Internet. The survey plans to move completely toward a web questionnaire.

An analysis of the response rate by mode, in the first period of introduction of Internet data collection (2003–2006), is interesting. The trend of the response is shown in Figure 7.6. The overall response rate did not change much over the year, but there are considerable changes in the composition of the response by mode.

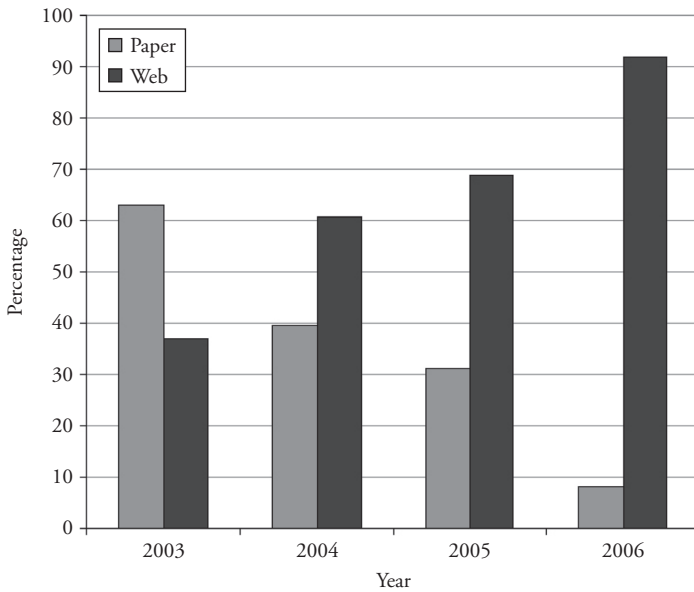


FIGURE 7.6 Response by mode in the SCI survey

The share of web mode responses has been constantly growing since 2003. This share increased from a share of 35.7% in 2003 to 92.0% in 2006. This result confirms that companies appreciate this response mode.

The electronic questionnaire form contained consistency checks. There were also facilities to compute the amount totals. It was expected that this would help to improve the quality of the data. Table 7.3 shows the average number of various types of problems by mode. Outliers denote values that differ substantially from the anticipated values. These values are not by definition incorrect. A careful evaluation must show whether they are correct and can be left as is, or whether they are wrong and must

be corrected. Substitutions denote incorrect values. As they are within a predefined threshold from the anticipated value, these values are automatically replaced by the anticipated value. It can be observed in the table that changing to the electronic questionnaire as a mode of data collection leads to a reduction of the number of problems.

TABLE 7.3 Detected mean errors in the SCI survey

Year	Response mode	Outliers	Substitutions	Corrected errors
2003	Paper	4.40	2.71	4.42
	Web	4.03	2.40	3.68
	Both modes	4.26	2.60	4.15
2004	Paper	0.79	4.48	4.13
	Web	1.03	3.30	3.15
	Both modes	0.93	4.05	3.54
2005	Paper	0.89	4.90	4.74
	Web	0.85	3.39	2.88
	Both modes	0.86	4.48	3.46
2006	Paper	0.84	4.30	3.56
	Web	0.85	3.31	3.05
	Both modes	0.85	4.01	3.09

Rosen and Gomes (2004) report on a test conducted in April 2004 by the U. S. Bureau of Labor Statistics. The Current Employment Statistics (CES) program involves monthly surveys among business establishments. Data are collected on employment, payroll, and working hours. The common way to collect these data is Touchtone Data Entry (TDE). The test investigated whether it was possible to convert TDE respondents to the web respondents. A sample of 3,000 TDE respondents was contacted by telephone, fax, and mail (1,000 for each contact method). They were all invited to change to web reporting. The response rate was 74%. All those who agreed to report by web received their initial web account information by mail.

It is worth noting that at the time of this study, 71% of the TDE respondents met the criteria imposed for reporting via the web (having access to the Internet, having e-mail at their desk, and using Internet Explorer 6.0 or higher). Of those meeting the eligibility criteria for the web, 89% reported that they wanted to switch to web reporting.

An important finding was that offering the web mode had a negative effect on response rates. Initially, the response rate dropped by 8%. Extensive follow-up procedures were needed to ensure the respondents activated their web accounts. As it turned out, fax was the most cost-effective contact method when converting respondents from TDE to web reporting.

7.2.5 MIXED MODE FOR SURVEYS AMONG HOUSEHOLDS AND INDIVIDUALS

One factor standing in the way of large-scale application of the Internet for conducting surveys among households and individuals is the penetration of the Internet in the population. Even though the number of persons with Internet access is rapidly growing, population coverage is far from 100%. Moreover, people with Internet access may only have it at their workplace and not at home. This may also prevent participation in web surveys.

Table 7.4 presents an overview of Internet penetration in the world. There is a large variation across the various regions. Penetration is highest in North America, where 77% of the people have access to the Internet. Penetration is also high in Australia/Oceania (61%) and Europe (58%). Penetration is lowest in Asia and Africa. Note that in some regions with a low penetration (Africa, Middle East, Latin America), Internet access is increasing very rapidly.

Internet penetration has a large variation within Europe. See Table 7.5 for details. In some countries, like United Kingdom, Luxembourg, Finland, Denmark and the Netherlands, Internet penetration is very high. There is almost full population coverage. Internet access is very poor in several Balkan countries, like Bosnia-Herzegovina, Romania, Albania, Montenegro, Bulgaria, and Croatia. Nevertheless, in most countries, more than half of the people have Internet access.

It should be noted that even with (almost) complete Internet coverage there are still sampling problems. As there is no sampling frame for an e-mail address, other modes have to be used to select a sample. Indeed, Couper and Miller (2008) state that “frames of internet users in a form suitable for sampling do not – and likely will not – exist.” Couper (2007 page 832) adds that only for specialized populations list may exist, and that there is no analogue for random digit dialing in telephone surveys. There is no algorithm to generate existing e-mail addresses randomly. Currently, even if probability sampling of Internet users was technically possible, the sample would not be representative because of coverage problems (Couper, 2007; Bethlehem, 2009). However, within the framework of probability sampling, there are possibilities of conducting mixed-mode surveys. For example, a letter could be sent with a request to go to a survey website, or contact could be made by telephone and an e-mail address could be asked for to send a link to the survey questionnaire.

Couper and Miller (2008) conclude that there is not much evidence that mixed-mode surveys will increase the overall response rates. However, there is clear evidence that including self-administered modes (web and mail) will decrease the costs of the survey.

Beukenhorst (2008) reports on an ongoing program of Statistics Netherlands to move all single-mode surveys to mixed-mode surveys. He states that this is not leading to higher response rates. Respondents who would have responded anyway, have done so again but in a different mode. Face-to-face survey non-respondents did not change their minds when approached in a follow-up in a different mode. There has also been little success in contacting face-to-face noncontacts by telephone. From field experiments, it was concluded that it is

TABLE 7.4 Internet access in the world

World region	Population in 2010 (millions)	Internet users in 2000 (millions)	Internet users in 2010 (millions)	Internet penetration in 2010	Penetration growth 2000–2010	Share of users
Africa	1,014	5	111	11%	2,357%	6%
Asia	3,835	114	825	22%	622%	42%
Europe	813	105	475	58%	352%	24%
Middle East	212	3	63	30%	1,825%	3%
North America	344	108	266	77%	146%	14%
Latin America and Caribbean	593	18	205	35%	1,033%	10%
Oceania and Australia	35	8	21	61%	179%	1%
World	6,846	361	1,966	29%	445%	100%

(Source: www.internetworldstats.com)

TABLE 7.5 Internet access in European countries

Country	Internet penetration
Isle of Man, Svalbard, Vatican City	30% or less
Bosnia-Herzegovina, Jersey, Ukraine, Romania, Cyprus	30–40%
Russia, Albania, Montenegro, Turkey, Belarus, Greece, Bulgaria, Portugal, Croatia	40–50%
Macedonia, Italy, San Marino, Serbia, Poland, Malta, Lithuania	50–60%
Hungary, Spain, Slovenia, Czech Republic, Lichtenstein, Ireland, Latvia, France, Gibraltar	60–70%
Slovakia, Guernsey & Alderney, Austria, Estonia, Monaco, Switzerland, Faroer Islands, Belgium, Germany, Andorra	70–80%
United Kingdom, Luxembourg, Finland, Denmark, Netherlands	80–90%

(Source: www.internetworldstats.com)

possible to use a mixed mode including Internet where the sample consists of individuals, but it is not yet evident whether the same applies for household surveys in which all household members have to complete a questionnaire form (for example, the Labor Force Survey). Beukenhorst (2008) concludes that the results of a mixed-mode test with the 2008 Travel Survey (with a household-level questionnaire and individual-level questionnaires) are “relatively disappointing.”

Statistics Netherlands conducts many general-population surveys. Until a few years ago, these were all single-mode surveys. Some like (the first round of) the Labor Force Survey were CAPI surveys, whereas others were CATI surveys. The choice for interviewer-assisted surveys was made because high data quality was required. Most questions in these types of surveys are factual questions, so that there are no problems with socially desirable answers. The sampling frame for both CAPI and CATI surveys is the population register of the Netherlands. Both CAPI and CATI surveys suffer from nonresponse. Moreover, CATI surveys suffer from increasing undercoverage problems because for less and less people it is possible to link a telephone number to their addresses. At least 30% of the telephone numbers are unlisted.

These problems and the rapidly increasing costs of interviewer-assisted surveys caused Statistics Netherlands to look for new ways of collecting data. Mixed-mode surveys seem to be a promising alternative as they may keep response rates at the same level, they can reduce the coverage problem, and are cheaper than single-mode CAPI or CATI surveys.

EXAMPLE 7.10 Measurement errors in the ESS mixed-mode experiment

The European Social Survey (ESS) has two main objectives. The first one is to explore and explain the interaction among Europe’s changing institutions, its political and economic structures, and the attitudes,

beliefs, and behaviors of its people. The second objective is to show that it is possible to conduct a high-quality, cross-national social survey.

The ESS is funded jointly by the European Commission, the European Science Foundation, and scientific funding bodies in 30 European countries. The survey is conducted biannually. In each participating country, approximately 2,000 respondents have to answer a one-hour questionnaire in a face-to-face interview and to complete a short supplementary questionnaire either face-to-face or on paper.

A major challenge of the ESS is collecting data that allow for comparing countries. This can only be accomplished by applying the fundamental methodological principles of survey sampling. The ESS attempts to achieve this by enforcing consistent, standardized procedures and protocols in all participating countries. These include sampling design, questionnaire definition, and fieldwork procedures.

The ESS is likely to change in the near future from single-mode face-to-face surveys to mixed-mode surveys. It is expected that costs can be reduced and response can be kept at an acceptable level. An experiment will be conducted to test these assumptions. The modes that will be considered for the mixed-mode survey, are web, telephone, and face-to-face. The sampling procedures will be kept the same.

The experiment will be a multitrade–multimethod (MTMM) experiment. A limited number of questions is repeated in a different format at the end of the questionnaire in each mode. This allows for investigating the possible mode effects and question effects.

EXAMPLE 7.11 Response rates in mixed-mode surveys

Statistics Netherlands carried out some experiments with mixed-mode surveys. One experiment was conducted for the ICT survey. This survey collects information on the use of computers and Internet in households and by individuals. The regular ICT survey was a CATI survey. The survey was fairly expensive. It also suffered from undercoverage because the sample was selected from the telephone directory. Households with unlisted numbers and mobile-only households could not be selected.

The main objectives of this pilot were as follows:

- To find out what level of response could be obtained.
- To establish whether people without Internet would be properly represented in a mixed-mode survey with the web as the most important mode.

Respondents had the possibility of completing the questionnaire on paper. To prevent those with Internet from responding by paper, the paper questionnaire was not included in the invitation letter. People had to apply for the paper form by returning a stamped return postcard.

The sample was selected from the population register. So there was no undercoverage. All persons in the sample received an invitation letter by mail. The letter contained the Internet address of the survey and a unique log-in code.

After one week, a postcard was sent to all nonrespondents with a reminder to complete the survey questionnaire, by either web or mail.

Two weeks after receipt of the invitation letter, the nonrespondents were approached again. They were randomly split into two subgroups. All people in the first subgroup with a known telephone number were approached by telephone (CATI). The remaining people in the first subgroup received a reminder letter by mail. All people in the other subgroup were reminded by mail. Note that the CATI approach was just to remind people to complete the questionnaire on the Internet, not to conduct a telephone interview.

Table 7.6 contains the response rates by age class and mode. The response rates were substantially lower in the mixed-mode survey than in the single-mode survey. The differences were more extreme for the elderly.

It is also clear from the table that the web response rate decreases with an increasing age. The opposite effect can be observed for the mail response rate. These trends can be explained by the lower Internet penetration among the elderly. And even if they have Internet, they may find it easier to complete the paper questionnaire.

TABLE 7.6 Response rates (%) by age and mode in the ICT experiment

Age class	Single-mode		Mixed-mode		Difference
	CATI	Web	Mail	Total	
0–25	54	36	6	42	–12
26–35	42	26	7	33	–9
36–45	52	29	7	36	–16
46–55	53	29	9	38	–15
56–65	52	23	16	39	–13
66 and older	59	13	24	37	–22

The analysis of the results also showed low response rates for single persons, single parents, divorced people, widowed people, and ethnic minorities from non-Western countries and people with a low income.

Another conclusion of this experiment was that reminding people by telephone works better than reminding them by mail. The answers to the

questions of those reminded by telephone were more accurate than the answers of those reminded by mail. This is probably because the personal attention of an interviewer motivates the respondent more. More details can be found in the study by Janssen (2006).

7.3 Application

The ideal situation for a web survey is to have a sampling frame containing the e-mail addresses of all members of the target population. This makes it easy to inform sample persons they have been selected in the survey. Moreover, it is also very easy for the sample persons to access the survey questionnaire. It is just a matter of clicking on the link included in the e-mail. Often such a sampling frame for e-mail addresses does not exist. Then some other means has to be deployed to get into contact with those selected for the survey. The least expensive way to do this is by mail.

Schonlau, Asch, and Du (2003) conducted a mixed-mode experiment in which they tested this. A sample of high-school students was approached by mail. They were invited to complete a survey questionnaire on the web. The objective of this experiment was to obtain answers to three questions:

- Can the majority of respondents be convinced by mail to participate in a web survey?
- Are reminders by telephone effective for increasing response rates?
- Are reminders by mail effective for increasing response rates?

The target population consisted of high-school students graduating in 2001. There was a sampling frame with addresses. A sample of 1,750 students was selected.

The fieldwork of the survey consisted of two stages. In the first stage, potential respondents were asked by mail to complete the questionnaire on the Internet. There was a possibility of filling in a paper questionnaire. It was not included in the invitation letter, but it could be requested. The objective of this design was to stimulate the students to respond by web as much as possible, because conducting a web survey is cheaper than conducting a mail survey.

After some time, reminders (without a paper copy of the questionnaire) were sent to all students in the sample. Moreover, a random sample of nonrespondents was contacted and reminded by telephone.

In the second stage of the experiment, all remaining nonrespondents were sent a paper questionnaire form. So they had a choice to complete either the mail or the web questionnaire. In addition, an incentive was sent to a random sample of nonrespondents. The incentive was a McDonald gift certificate of \$3.

TABLE 7.7 Modes, reminders, and incentives in different stages of the survey

Stage	Mode	Reminders/incentives
1	Invitation letter sent by mail. No paper questionnaire was included.	All nonrespondents were reminded by mail (no paper questionnaire was included).
	Request to complete the web questionnaire. A paper questionnaire could be requested.	A random subsample of nonrespondents was reminded by telephone.
2	Letter to stage 1 nonrespondents with a paper copy of the questionnaire. Choice of web or mail response.	First reminder: postcard to all remaining nonrespondents.
		Second reminder: letter with paper questionnaire to all remaining nonrespondents.
		A random subsample of stage 1 nonrespondents received a \$3 incentive by mail. The paper questionnaire was included in the letter.

Two reminders were sent. The first one was just a postcard, and the second one was a letter that included a paper copy of the questionnaire. The experimental design is summarized in Table 7.7.

It turned out that of the respondents, only 35% of those did so by the web. Although they were not encouraged to do so, 65% responded by mail. Quigley et al. (2000) report on a similar experiment where 73% of the respondents preferred the web. However, there were differences between the experiments. One is that different target populations were approached (students versus the military), and the second is that Quigley et al. offered the paper questionnaires only very late in the fieldwork period.

The fact that only about one third of the students preferred the web was considered disappointing. It was expected that the students were intensive computer and Internet users. Therefore, it could have been attractive for them to do the survey on the Internet.

The use of incentives increased the response rate. However, this increase only occurred for the mail mode. There was no effect for the web mode. Two factors may explain this. One, by the time the mail response and the incentives were introduced, the survey had been in the field already for a considerable period of time. This may have made the students less aware of the fact that they were expected to complete a web survey questionnaire. Second, the mode used for sending the incentives (mail) may have affected the respondents' choice of the response mode (mail).

Reminding the students by telephone was very effective. There was a substantial increase in the response rate in the web mode. For those receiving the telephone call, the subsequent response rate was 30%. The subsequent response rate of the other students was only 18%.

7.4 Summary

There is a growing interest in conducting mixed-mode surveys where the web is one of the data collection modes. It is considered a potential means for increasing response rates, reducing undercoverage problems, and decreasing survey costs.

There are two approaches for implementing mixed-mode designs: a sequential design and a concurrent design. In a sequential mixed-mode survey, the sample persons are approached in one mode. Nonrespondents are followed up in a different mode. This process can be repeated for a series of modes. For the concurrent (parallel) approach, the sample is divided into groups and each group is approached with a different mode. Many different modes can be combined in a mixed-mode survey: face-to-face, telephone, mail, fax, IVR, TDE, and so on. Therefore, many different designs are possible. Sometimes sequential modes and concurrent modes also are combined.

No best mixed-mode design exists. It depends on the specific research context, the topics of the surveys, and the type of questions asked (factual or attitudinal). Different phenomena may have different effects in different modes. For example, interviewer-assisted surveys may suffer from socially desirable answers.

Mixed-mode surveys seem to be not very successful at increasing response rates. In particular, the response rates for the web mode are disappointingly low. Also providing sampled units with a choice of mode does not seem to increase the overall response rates. Therefore, mixed-mode surveys are not a solution for declining response rates.

Mixed-mode surveys seem to be very successful at reducing survey costs. This is a consequence of replacing (part of) interviewer-assisted interviewing with self-administered completion of questionnaires.

Mixed-mode surveys may suffer from mode effects (i.e., the same question is answered differently in a different mode). It is not easy to correct for mode effects because they are confounded with other effects such as selection effects.

The mixed-mode approach is promising for both household/individual surveys and business surveys. Although it may be hard to disentangle some of the methodological problems, it is important to continue research in both application areas. Cross-fertilization of the results of experiments in business and household/individual surveys will help to improve the quality of future mixed-mode surveys.

In summary, there is an increasing interest in using mixed-mode surveys for data collection. Although these surveys may help to reduce survey costs, there are still some estimation problems that have not yet been solved. Continued research is required in this area.

KEY TERMS

Concurrent mixed mode: A mixed-mode design in which respondents can choose among alternative modes of data collection, or in which different groups are approached by different modes.

Mixed-mode survey: A survey in which various modes of data collection are combined. Modes can be used concurrently (different groups are approached by different modes) or sequentially (nonrespondents of a mode are reapproached by a different mode).

Sequential mixed mode: A mixed-mode design in which nonrespondents from one mode are reapproached in another mode.

Mixed-mode systems: A system of data collection that involves different modes of data collection and different modes of communication with the respondents.

Mode effect: The phenomenon that a question is answered differently when asked in a different mode. Sometimes this term is used in a wider context, in which it denotes the combined differences of the modes. This includes, for example, differences in coverage and differences in response rates.

Socially desirable answer: The tendency that respondents give answers that will be viewed as more favorable by others. This particularly happens for sensitive questions.

Unimode design: A mixed-mode survey design in which the same questionnaire is used in each mode. Following the guidelines for unimode questionnaires, the questions are defined in such a way that mode effects are minimized.

EXERCISES

Exercise 7.1. What is a unimode survey?

- a. A unimode survey is a special type of mixed-mode survey. Questions are defined in exactly the same way in each mode.
- b. A unimode survey is a special type of mixed-mode survey. Questions may be defined differently in each mode, but they measure the same concept.
- c. It is a sequential mixed-mode survey in which the first mode produces the highest quality data.
- d. It is a different term for a single mode survey (i.e., a survey with only one mode of data collection).

Exercise 7.2. Can the bias of estimators from a mixed-mode approach be corrected by means of a weighting adjustment?

- a. Yes, weighting adjustment can be successful if the bias is caused by undercoverage and/or nonresponse.
- b. Yes, weighting adjustment can be successful if the bias is caused by mode effects.
- c. Yes, weighting adjustment will always be able to reduce a mixed-mode bias.
- d. No, weighting adjustment will never be able to reduce a mixed-mode bias.

Exercise 7.3. A survey was conducted using web and mail for data collection concurrently. An invitation letter was sent by ordinary mail to a sample of 2,000

potential respondents. They were offered a choice of completing a paper questionnaire form or a web form. Overall, 1,200 completed questionnaires were returned by mail, and 800 forms were completed on the web. Calculate the overall response rate and the response rate for each mode.

Exercise 7.4. A survey was conducted using web and CATI for data collection sequentially. An invitation letter was sent by ordinary mail to a sample of 2,000 potential respondents. They were asked to complete the questionnaire on the web. A total of 420 persons did so. After two weeks, all remaining respondents were called by telephone. Overall, 640 persons completed the questionnaire by telephone. Calculate the overall response rate and the response rate by mode.

Exercise 7.5. According to the literature, what can be said if the response rates of CATI, web, and Interactive Voice Response (IVR) are compared?

- a. CATI is lower than the web.
- b. IVR is higher than CATI.
- c. CATI is higher than the web.
- d. IVR is lower than the web.

Exercise 7.6. According to the literature, what can be said if the item non-response rates of CATI, web, and Interactive Voice Response (IVR) are compared?

- a. CATI is higher than the web.
- b. IVR is higher than CATI.
- c. CATI is lower than the web.
- d. IVR is lower than the web.

Exercise 7.7. According to the literature, what can be said if the accuracy of the answers of CATI, web, and Interactive Voice Response (IVR) are compared?

- a. CATI is higher than the web.
- b. IVR is higher than CATI.
- c. CATI is lower than the web.
- d. IVR is lower than the web.

Exercise 7.8. If a CAPI survey is replaced by a sequential mixed-mode survey, with the web as the first mode and CAPI as the second mode, what can be said about the costs and the response rate of the new survey?

- a. Both the response rate and the costs will go up.
- b. Both the response rate and the costs will go down.

- c. The response rate will go up and the costs will go down.
- d. The response rate will go down and the costs will go up.

Exercise 7.9. Which of the following phenomena cannot cause a mode effect in a mixed-mode survey?

- a. Straight-lining in matrix questions.
- b. Memory effects in recall questions.
- c. Response order effects.
- d. Socially desirable answers to sensitive questions.

Exercise 7.10. Which of the following statements describe the advantages of web surveys over mail surveys?

- a. Checks can be included in the questionnaire.
- b. The questionnaire can be completed quickly.
- c. Dynamic routing can be implemented in the questionnaire.
- d. There are no undercoverage problems.

REFERENCES

- AAPOR. (2009), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, 6th edition*. The American Association for Public Opinion Research, Deerfield, IL.
- Ariel, A., Giesen, D., Kerssemakers, F., & Vis-Visschers, R. (2008), *Literature Review on Mixed Mode Studies*. Internal report DMH-2008-04-16-RVCS, Statistics Netherlands, Heerlen, the Netherlands.
- Bethlehem, J. (2009), *Applied Survey Methods—A Statistical Perspective*. John Wiley & Sons, Hoboken, NJ.
- Bethlehem, J. G., Cobben, F., & Schouten, B. (2011), *Handbook on Nonresponse in Household Surveys*. John Wiley & Sons, Hoboken, NJ.
- Beukenhorst, D. (2008), *Mixed Mode: Update on Past Research and Plans for the Near Future*. Internal Report, Statistics Netherlands, Heerlen, the Netherlands.
- Beukenhorst, D. & Wetzels, W. (2009), *A Comparison of Two Mixed Mode Designs of the Dutch Safety Monitor: Mode Effects, Costs, Logistics*. Paper presented at the European Survey Research Association Conference, Warsaw, Poland.
- Biemer, P. & Lyberg, L. (2003), *Introduction to Survey Quality*. John Wiley & Sons, Hoboken, NJ.
- Biffignandi, S. & Fabrizi, E. (2006), Mixing Web and Paper Based Data Collection Modes in a Survey on Manufacturing Firms. *Proceedings of the European Conference on Quality Survey Statistic—Q2006*, Cardiff, U. K.
- Biffignandi, S. & Zeli, A. (2008), *Statistical Quality Analysis in a Mixed-mode Survey*. Paper presented at the Business Data Collection Workshop, Ottawa, Canada.

- Biffignandi, S., Pratesi, M., Lozar Manfreda, K., & Vehovar, V. (2004), List Assisted Web Surveys: Quality, Timeliness and Nonresponse in the Steps of the Participation Flow. *Journal of Official Statistics*, 20, pp. 451–465.
- Bowling, A. (2005), Mode of Questionnaire Administration Can Have Serious Effects on Data Quality. *Journal of Public Health*, 27, pp. 281–291.
- Christian, L. M., Dillman, D., & Smyth, J. D. (2005), Instructing Web and Telephone Respondents to Report Date Answers in Format Desired by the Surveyor. *Social and Economic Sciences Research Center Technical Report 05–067*, Washington State University, Pullman, WA.
- Clayton, R. & Werking, G. (1998), Business Surveys of the Future: The World Wide Web as a Data Collection Methodology. In: Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls II, W. L., & O'Reilly, J. M. (eds.), *Computer Assisted Survey Information Collection*, John Wiley & Sons, New York, pp. 543–562.
- Cobben, F., Schouten, B., & Bethlehem, J. G. (2006), *A Model for Statistical Inference based on Mixed Mode Interviewing*. Proceedings of Q2006 European Conference on Quality in Survey Statistics - Q2006, Cardiff, U. K.
- Couper, M. (2007), *Web Surveys in a Mixed Mode World*. Paper presented at the ESRC Conference “Survey Research in the 21st Century: Challenges and Opportunities,” London, U. K.
- Couper, M. & Miller, P. (2008), Web Survey Methods Introduction. *Public Opinion Quarterly*, 72, pp. 831–835.
- De Leeuw, E. D. (2005), To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21, pp. 233–255.
- De Leeuw, E. D. (1992), *Data Quality in Mail, Telephone, and Face-to-Face Surveys*. TT-Publications, Amsterdam, the Netherlands.
- De Leeuw, E. D. & De Heer, W. (2002), Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In: Groves, R. M., Dillman, D. A., Eltinghe, J. I., & Little, R. J. A. (eds.), *Survey Nonresponse*, John Wiley & Sons, New York.
- De Leeuw, E. D. (2008), Choosing the Method of Data Collection. In: De Leeuw, E. D., Hox, J. J. & Dillman, D. A. (eds.), *International Handbook of Survey Methodology*, Lawrence Erlbaum Associates, New York, pp. 113–135.
- Dillman, D. A. (2000), *Mail and Internet Surveys: The Tailored Design Method*. John Wiley & Sons, Hoboken, NJ.
- Dillman, D. A. (2007), *Mail and Internet Surveys: The Tailored Design Method*. John Wiley & Sons, Hoboken, NJ.
- Dillman, D. A., Clark, J. R., & West, K. K. (1995), Influence of an Invitation to Answer by Telephone on Response to Census Questionnaire. *Public Opinion Quarterly*, 58, pp. 557–568.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009), *Internet, Mail and Mixed-mode Surveys: The Tailored Design Method*. John Wiley & Sons, Hoboken, NJ.
- Griffin, D., Fischer, D., & Morgan, M. (2001), *Testing an Internet Response Option for the American Community Survey*. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Montreal, Canada.
- Groves, R., Fowler, F. J., Couper, M. D., Lepkowski, P., Singer, J. M., & Tourangeau, R. (2004), *Survey Methodology*. John Wiley & Sons, New York.

- Hochstim, J. R. (1967), A Critical Comparison of Three Strategies of Collecting Data From Households. *Journal of American Statistical Association*, 62, pp. 976–989.
- Irani, T. A., Gregg, J. A., & Telg, R. (2004), Choosing to Use the Web: Comparing Early and Late Respondents to an Online Web-based Survey Designed to Assess IT Computer Skills Perceptions of County Extension Agents. *Journal of Southern Agricultural Education Research*, 54, pp. 168–179.
- Jäckle, A., Roberts, C., & Lynn, P. (2010), Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, 78, pp. 3–20.
- Janssen, B. (2006), *Web Data Collection in a Mixed Mode Approach: An Experiment*. Paper presented at the European Conference on Quality in Official Statistics (Q2006), Cardiff, U.K.
- Knapp, H. & Kirk, S. A. (2003), Using Pencil and Paper, Internet and Touch-Tone Phones for Self-Administered Surveys: Does Methodology Matter? *Computers in Human Behavior*, 19, pp. 117–134.
- Kiesler, S. & Sproull, L. S. (1986), Response Effects in the Electronic Survey. *Public Opinion Quarterly*, 50, pp. 402–413.
- Kraan, T., Van den Brakel, Buelens, B., & Huys, H. (2010), *Social Desirability Bias, Response Order Effect and Selection Effects in the New Dutch Safety Monitor*. Discussion Paper 10004, Statistics Netherlands, The Hague/Heerlen, the Netherlands.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008), Social Desirability Bias in CATI, IVR, and Web Surveys, the Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, 72, pp. 847–865.
- Lynn, P., Beerten, R., Laiho, J., & Martin, J. (2002), Towards Standardisation of Survey Outcome Categories and Response Rate Calculations. *Research in Official Statistics*, 1, pp. 63–86.
- Maier, S. R. (2005), Comparing Internet vs. Paper in Newspaper Source Surveys. *Newspaper Research Journal*, 26, pp. 57–71.
- Meckel, M., Walters, D., & Baugh, P. (2005), Mixed-Mode Surveys Using Mail and Web Questionnaires. *The Electronic Journal of Business Research Methodology*, 3, pp. 69–80.
- Mooney, G., Giesbrecht, L., & Shettle, C. (1993), *To Pay or not to Pay: That is the Question*. Paper presented at the Meeting of the American Association for Public Opinion Research, St. Charles, IL.
- Pierzchala, M. (2006), *Disparate Modes and Their Effect on Instrument Design*. Proceedings of the 10th International Blaise Users Conference, Arnhem, the Netherlands. <http://www.blaiseusers.org/2006/Papers/207.pdf>.
- Quigley, B., Riemer, R. A., Cruzen, D. E., & Rosen, S. (2000), *Internet Versus Paper Survey Administration: Preliminary Finding on Response Rates*. Paper presented at the 42nd Annual Conference of the International Military Testing Association, Edinburgh, Scotland.
- Roberts, C. (2005), *Expert Workshop on Mixed Mode Data Collection in Comparative Social Surveys*. Report, Economic and Social Research Council, National Centre for Research Methods, City University, London, U.K.
- Roberts, C. (2007), *Mixing Modes of Data Collection in Surveys: A Methodological Review*. Report, Economic and Social Research Council, National Centre for Research Methods, City University, London, U.K.

- Roos, M. & Wings, H. (2000), *Blaise Internet Services Put to the Test: Web-surveying the Construction Industry*. Proceedings of the 6th International Blaise Users Conference, Kinsale, Ireland.
- Rogelberg, S. G. & Stanton, J. M. (2007), Understanding and Dealing With Organization Survey Nonresponse. *Organization Research Methods*, 10, pp. 195–209.
- Rosen, R. & Gomes, T. (2004), *Converting CES Reporters from TDE to Web Data Collection*. Paper presented at the Joint Statistical Meeting, Toronto, Canada.
- Schonlau, M., Asch, B. J., & Du, C. (2003), Web Surveys as Part of a Mixed Mode Strategy for Populations than Cannot Be Contacted by E-mail. *Social Science Computer Review*, 21, pp. 218–222.
- Sheehan, K. (2001), E-mail Survey Response Rates: A review. *Journal of Computer-Mediated Communication*, 6. <http://jcmc.indiana.edu/vol6/issue2/sheehan.html>.
- Shih, T. H. & Fan, X. (2007), Response Rates and Mode Preferences in Web-Mail Mixed-Mode Surveys: A Meta-Analysis. *International Journal of Internet Science*, 2, pp. 59–82.
- Tourangeau, R. & Yan, T. (2007), Sensitive questions in surveys. *Psychological Bulletin*, 133, pp. 859–883.
- Tourangeau, R. (1984), Cognitive Sciences and Survey Methods. In: Janine, T. T. Janine, Lofts, G., Strafe, M., Tanner, J., & Tourangeau, R. (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, National Academy of Science, Washington, DC, pp. 73–100.
- Voogt, R. J. & Saris, W. (2005), Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects. *Journal of Official Statistics*, 21, pp. 367–387.
- Zhang, Y. (2000), Using the Internet for Survey Research: A Case Study. *Journal of the American Society for Information Science*, 51, pp. 57–68.

The Problem of Undercoverage

8.1 Introduction

Collecting data with a survey is often a complex, costly, and time-consuming process. Not surprisingly, continuous attempts have been made all through the history of survey research to improve timeliness and reduce costs, while maintaining a high level of data quality. Developments in information technology in the last decades of the previous century made it possible to use microcomputers for data collecting. This led to the introduction of computer-assisted interviewing (CAI). Replacing the paper questionnaire by an electronic one turned out to have many advantages, among which are considerably shorter survey processing times and higher data quality. More on the benefits of CAI can be found in Couper et al. (1998). The next important development in the area of survey research was the fast rise of the Internet in the 1990s. This made it possible to conduct surveys on-line. Web surveys seem to have some attractive advantages in terms of costs and timeliness:

- Nowadays many people are connected to the Internet. Therefore, a web survey is a simple means to getting access to a large group of potential respondents.
- Web survey questionnaires can be distributed at very low cost. No interviewers are needed. This makes it cheaper than face-to-face or telephone surveys. A web survey can also be cheaper in terms of mailing and printing costs.

- Web surveys can be set up very quickly. Fieldwork can start immediately after the questionnaire has been installed on the Internet.

So a web survey is a fast and cheap means of collecting large amounts of data. Not surprisingly, many survey organizations (and other organizations) started conducting such surveys. However, costs and timeliness are not the only aspects. More important is the question of whether web surveys can produce reliable and precise estimates of population characteristics.

When conducting a survey, a survey researcher is confronted with all kinds of phenomena that may have a negative impact on the quality, and therefore the reliability, of the outcomes. Some of these disturbances are almost impossible to prevent. So, efforts will have to be aimed at reducing their impact as much as possible. Nevertheless, notwithstanding all these efforts, the final estimates of population characteristics may be affected. One of these phenomena is undercoverage. This is the topic of this chapter.

To be able to select a sample from a target population, a sampling frame is required. A *sampling frame* is a list of all elements in the target population. For every element in the list there must be information on how to contact that element. Such contact information can include, for example, name and address, telephone number, or e-mail address. Such lists can exist on paper (a card-index box for the members of a club or a telephone directory), or in a computer (a database containing a register of all companies). If such lists are not available, detailed geographical maps are sometimes used.

For a face-to-face survey among persons, the sampling frame could consist of names and addresses. Some countries (the Netherlands and the Scandinavian countries) have population registers. These contain the names and addresses of all permanent residents in the country.

If a population register is not available, an address register could be an alternative. For example, TNT Post, the postal service company in the Netherlands, has a Postal Delivery Points file. This is a computer file containing all addresses (of both private houses and companies) where mail can be delivered. Typically, this file can be used to draw a sample of households. If required, a person can be randomly drawn from each selected address.

For a telephone survey, the sampling frame could be a telephone directory. A disadvantage of this sampling frame is that many people have unlisted telephone numbers. An alternative could be to apply random digit dialing (RDD), where valid telephone numbers are generated by some computer algorithm.

The obvious sampling frame for a web survey would be a list of e-mail addresses. Sometimes such a sampling frame exists. For example, all employees of a large company may have a company e-mail address. Similarly, all students of university usually have an e-mail address. The situation is more complicated for a general-population survey. Unfortunately, there does not exist (yet) a list of e-mail addresses of everybody in the country.

The sampling frame should be an accurate representation of the population. There is a risk of drawing the wrong conclusion from the survey if the sample has

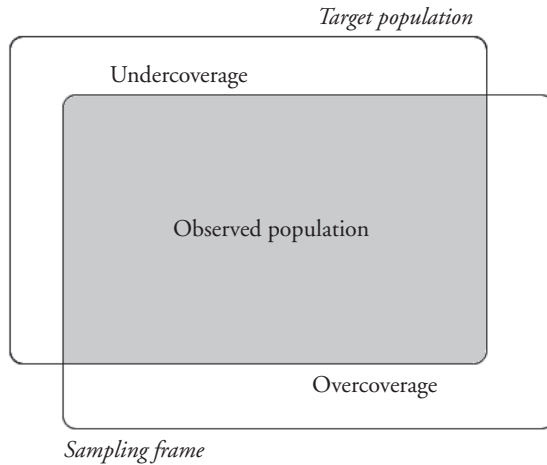


FIGURE 8.1 The target population and the sampling frame

been selected from a sampling frame that differs from the population. Figure 8.1 shows what can go wrong.

The first problem is *undercoverage*. This occurs if the target population contains elements that do not have a counterpart in the sampling frame. Such elements can never be selected in the sample. An example of undercoverage is a survey where the sample is selected from a population register. Illegal immigrants may be considered part of the population, but they will never be encountered in the sampling frame. Another example is a web survey, where respondents are selected via the Internet. Then there will be undercoverage resulting from people without Internet access. Undercoverage can have serious consequences. If the elements outside the sampling frame systematically differ from the elements in the sampling frame, the estimates of population parameters may be seriously biased. A complicating factor is that it is often not very easy to detect the existence of undercoverage.

The second sampling frame problem is *overcoverage*. This refers to the situation where the sampling frame contains elements that do not belong to the target population. If such elements end up in the sample and their data are used in the analysis, the estimates of population parameters may be affected. It should be simple to detect overcoverage in the field. This should become clear from the answers to the questions.

EXAMPLE 8.1 A web survey with telephone recruitment

The local authorities of a town intend to set up a web panel of citizens. To recruit people for this panel, a simple random sample is selected from the telephone directory of the town. People are called and invited to join the panel. If they agree, they are asked for their e-mail address. A link to

the questionnaire on the Internet and a unique access code is sent to each respondent by e-mail.

At first sight, it might be a good idea to use the telephone directory of the town as a sampling frame. However, it suffers from serious coverage problems. Undercoverage occurs because many people have unlisted numbers, and some will have no telephone at all. Moreover, there is a rapid increase in the use of mobile phones. In many countries, mobile phone numbers are not listed in directories.

The telephone directory also suffers from overcoverage because it contains the telephone numbers of shops, companies, and so on. Hence, it may happen that people are contacted that do not belong to the target population. Moreover, some people may have a higher than assumed contact probability because they can be contacted both at home and in the office.

Even if people agree to participate, there may still be problems because these people do not necessarily have Internet at home. This would be a case of nonresponse. This problem could be solved by advising them to go to a public library with Internet access, to complete the questionnaire at work, or by simply giving people Internet access. Another solution could be to use a different mode of data collection, for example, a paper questionnaire that is sent by ordinary mail. This would be an example of a mixed-mode survey.

The set of all elements that can be contacted through the sampling frame is called the *frame population*. The sample is always selected from the sampling frame. Consequently, the conclusions drawn from the survey will apply to the frame population. Only if the frame population coincides with the target population will the results also apply to the target population.

The frame population for a web survey is by definition restricted to those having Internet. This can be a problem if the target population is the general population. Not everyone has access to the Internet. Figure 8.2 gives an overview of Internet access by households in the European Union in 2007. See also Eurostat (2007). It is clear there are large differences. On the one hand, Internet access is more than 80% in the Netherlands. In Sweden and Denmark, more than 70% of the households have access to the Internet. On the other hand, there are 13 countries where less than 50% of the household have access. Internet access is lowest in Romania and Bulgaria, with a percentage around 20%.

Figure 8.2 also shows how many Internet connections are based on broadband. It is clear that there are still many Internet connections that use slow modems. This imposes restrictions on which advanced features one can use (for example, animation and video) in a web questionnaire. Some questionnaires may simply not work for some people.

Undercoverage in web surveys would not be a problem if those with Internet access did not differ systematically from those without it. If people with Internet access could be considered a random sample from the population, valid conclusions

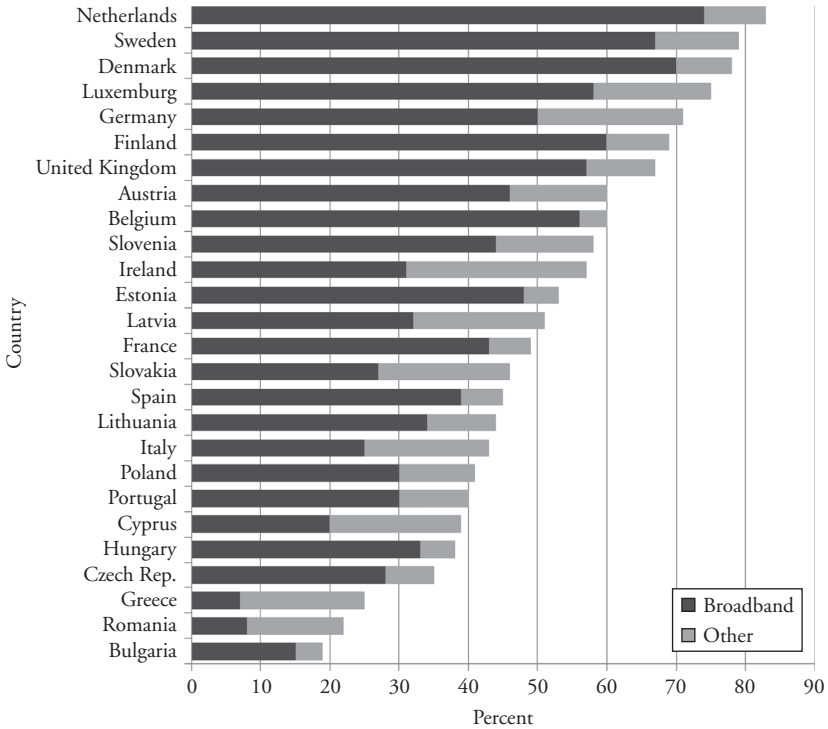


FIGURE 8.2 Internet access by households in the European Union in 2006

could be drawn from web surveys. Unfortunately, this is not the case. There are substantial differences between those with and without Internet access.

Analysis of data about Internet access in the Netherlands in 2005 (source: www.cbs.nl) shows some patterns that also are found in other countries. Figure 8.3 shows the distribution of Internet access by gender. Clearly, more males than females have access to the Internet.

Figure 8.4 contains the percentage of people having Internet access by age group. Clearly, Internet access decreases with age. In particular, people age 55 and older will be very much underrepresented when the Internet is used as a selection mechanism.

Figure 8.5 contains the percentage of people using the Internet by level of education. It is clear that people with a higher level of education more frequently have Internet access than people with a lower level of education.

According to De Haan and Van 't Hof (2006), Internet access among non-native young people in the Netherlands is much lower than among native young people: 91% of the young natives have access to the Internet. This is 80% for young people from Surinam and Antilles, 68% for young people from Turkey, and only 64% for young people from Morocco. These results are in line with the findings of authors in other countries. See, e.g., Couper (2000), and Dillman and Bowker (2001).

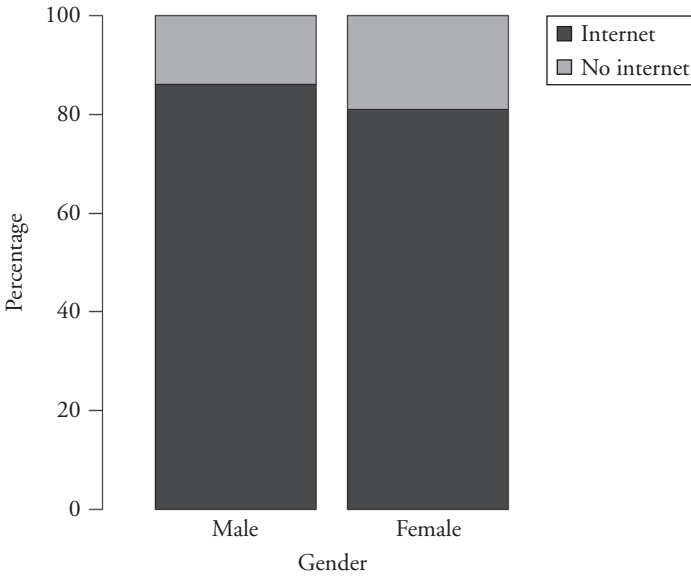


FIGURE 8.3 Internet access by gender

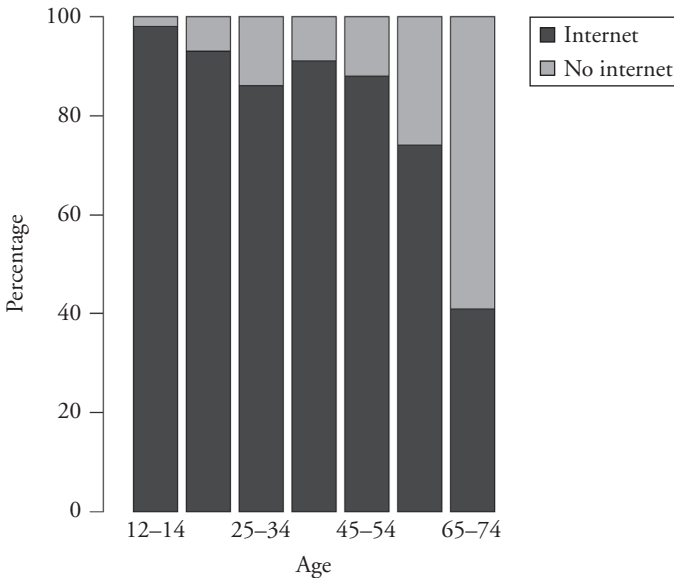


FIGURE 8.4 Internet access by age

It is clear that use of the Internet as a sampling frame will cause problems because specific groups are substantially under represented. Specific groups in the target population will not be able to fill in the (electronic) questionnaire form.

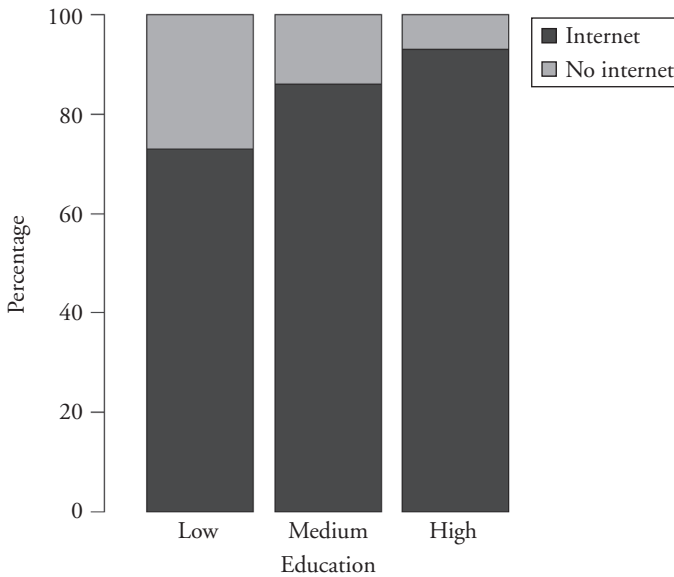


FIGURE 8.5 Internet access by level of education

8.2 Theory

8.2.1 THE INTERNET POPULATION

Let the target population U of the survey consist of N identifiable elements, which are labeled $1, 2, \dots, N$. Therefore, the target population can be denoted by

$$(8.1) \quad U = \{1, 2, \dots, N\}.$$

Associated with each element k is a value Y_k of the target variable Y . The aim of the web survey is assumed to be estimation of the population mean

$$(8.2) \quad \bar{Y} = \frac{1}{N} \sum_{k=1}^N Y_k$$

of the target variable Y .

The population U is divided into two subpopulations. There is a subpopulation U_I of elements having access to the Internet. There is also a subpopulation U_{NI} of elements not having access to the Internet. Associated with each element k is an indicator I_k , where $I_k=1$ if element k has access to the Internet (and thus is an element of subpopulation U_I), and $I_k=0$ otherwise. The subpopulation U_I will be called the *Internet population*, and U_{NI} is the *non-Internet population*. The number of elements in the Internet population U_I is equal to

$$(8.3) \quad N_I = \sum_{k=1}^N I_k.$$

Likewise,

$$(8.4) \quad N_{NI} = \sum_{k=1}^N (1 - I_k)$$

denotes the size of the non-Internet population U_{NI} , where $N_I + N_{NI} = N$.

The mean of the target variable for the elements in the Internet population is equal to

$$(8.5) \quad \bar{Y}_I = \frac{1}{N_I} \sum_{k=1}^N I_k Y_k.$$

Likewise, the mean of the target variable for the non-Internet population is denoted by

$$(8.6) \quad \bar{Y}_{NI} = \frac{1}{N_{NI}} \sum_{k=1}^N (1 - I_k) Y_k.$$

8.2.2 A RANDOM SAMPLE FROM THE INTERNET POPULATION

The more or less ideal case is considered now in which it is possible to select a random sample without replacement from the Internet population. This would require a sampling frame listing all elements having access to the Internet. Often there is no such list. A solution could be to select a random sample from a larger sampling frame. Examples include a population or an address list. People selected from such a list are recruited for the web survey by sending them a letter, by calling them on the telephone, or by visiting them at home. Only those with access to the Internet are selected for the web surveys. These persons are provided with a link to the electronic questionnaire, and possibly a unique entry code.

A random sample selected without replacement from the Internet population is denoted by a series

$$(8.7) \quad a_1, a_2, \dots, a_N$$

of N indicators. The k th indicator a_k assumes the value 1 if element k is selected, and otherwise it assumes the value 0, for $k = 1, 2, \dots, N$. Note that always $a_k = 0$ for elements k in the non-Internet population. The sample size is denoted by

$$(8.8) \quad n_I = a_1 + a_2 + \dots + a_N.$$

Horvitz and Thompson (1952) have shown that an unbiased estimator of a population mean always can be defined if all elements in the population have a known, positive probability of being selected. The Horvitz–Thompson estimator for the mean of the Internet population is defined by

$$(8.9) \quad \bar{y}_{HT} = \frac{1}{N_I} \sum_{k=1}^N a_k I_k \frac{Y_k}{\pi_k}.$$

The quantity π_k is called the *first-order inclusion probability* of element k . It is defined as the expected value

$$(8.10) \quad \pi_k = E(a_k)$$

if the indicator a_k . Note that by definition, $Y_k / \pi_k = 0$ for all elements outside the Internet population. The values of the π_k are determined by the sampling design. For a simple random sample of size n (with equal probabilities and without replacement) from a population of size N , all π_k are equal to n / N .

In the case of a simple random sample from the Internet population, all first-order inclusion probabilities are equal to n / N_I . Therefore, expression (8.9) reduces to

$$(8.11) \quad \bar{y}_I = \frac{1}{n} \sum_{k=1}^N a_k I_k Y_k.$$

This estimator is an unbiased estimator of the mean \bar{Y}_I of the Internet population but not necessarily of the mean \bar{Y} of the target population. The bias can be written as

$$(8.12) \quad B(\bar{y}_{HT}) = E(\bar{y}_{HT}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N} (\bar{Y}_I - \bar{Y}_{NI}).$$

The magnitude of this bias is determined by two factors. The first factor is the relative size N_{NI} / N of the non-Internet population. The bias will increase as a larger proportion of the population does not have access to Internet. The second factor is the *contrast* $\bar{Y}_I - \bar{Y}_{NI}$ between the Internet population and the non-Internet population. It is the difference between the population means of the two subpopulations. The more the mean of the target variable differs for these two subpopulations, the larger the bias will be.

The relative size of the non-Internet population cannot be neglected in many countries. Furthermore, there are substantial differences between those with and without Internet. Specific groups are under represented in the Internet population, for example, the elderly, those with a low level of education, and ethnic minority groups. So, the conclusion is that generally a random sample from an Internet population will lead to biased estimates for the parameters of the target population.

It is to be expected that Internet coverage will increase over time. The factor N_{NI} / N will become smaller, and this will reduce the bias. It is unclear, however, whether the contrast will also become smaller over time. It is even possible that it increases, as the remaining group of people without Internet access may be differing more and more from the Internet users. So the combined effect of a smaller non-Internet population and a larger contrast need not lead necessarily to a smaller bias.

It is important to note that the value of expression (8.12) does not depend on the sample size. Increasing the sample size will not reduce the bias. So the

problem of undercoverage in web surveys does not diminish by collecting a large number of observations.

The precision of an estimator is often quantified by a 95% confidence interval. Suppose that a simple random sample is selected from the target population. Then the sample mean \bar{y} can be computed. This is an unbiased estimator for the population mean \bar{Y} . As the sample mean is (approximately) normally distributed, the 95% confidence interval for the population mean is equal to

$$(8.13) \quad I = (\bar{y} - 1.96 \times S(\bar{y}); \bar{y} + 1.96 \times S(\bar{y})),$$

where $S(\bar{y})$ is the standard error of the sample mean. The probability that this interval contains the true value, which is by definition (approximately) equal to

$$(8.14) \quad P(\bar{Y} \in I) = 0.95.$$

The standard error will decrease if the sample size increases. This will lead to a smaller confidence interval. The estimator is more precise. If a simple random sample is selected from the Internet population, the sample mean \bar{y}_I is used to estimate the population mean \bar{Y} . Analogous to expression (8.13), the confidence will be computed as

$$(8.15) \quad I_I = (\bar{y}_I - 1.96 \times S(\bar{y}_I); \bar{y}_I + 1.96 \times S(\bar{y}_I)).$$

The confidence level of this interval is not by definition equal to 0.95. It can be shown that

$$(8.16) \quad P(\bar{Y} \in I_I) = \Phi\left(1.96 - \frac{B(\bar{y}_I)}{S(\bar{y}_I)}\right) - \Phi\left(-1.96 - \frac{B(\bar{y}_I)}{S(\bar{y}_I)}\right),$$

in which Φ is the standard normal distribution function. The quantity $B(\bar{y}_I)/S(\bar{y}_I)$ is called the *relative bias*. Apparently, the confidence level depends on the value of this relative bias. Figure 8.6 contains a plot of the confidence level as a function of the relative bias.

It is clear that the confidence level can be much lower than expected. A larger sample size will lead to a smaller standard error, but the bias will remain the same. So the relative bias increases. If the bias is equal to the standard error (i.e., the relative bias is 1) the confidence level is only 0.83. As the relative bias increases, the situation becomes worse. The confidence level is even less than 0.5 for a relative bias of 2. The conclusion is that undercoverage may lead to an incorrect confidence interval.

8.2.3 REDUCING THE NONCOVERAGE BIAS

There are several ways in which the negative effects of undercoverage can be reduced. Three approaches are discussed here.

The first approach is to give Internet access to persons in the sample without it. The Dutch Longitudinal Internet Studies for the Social Sciences (LISS) panel

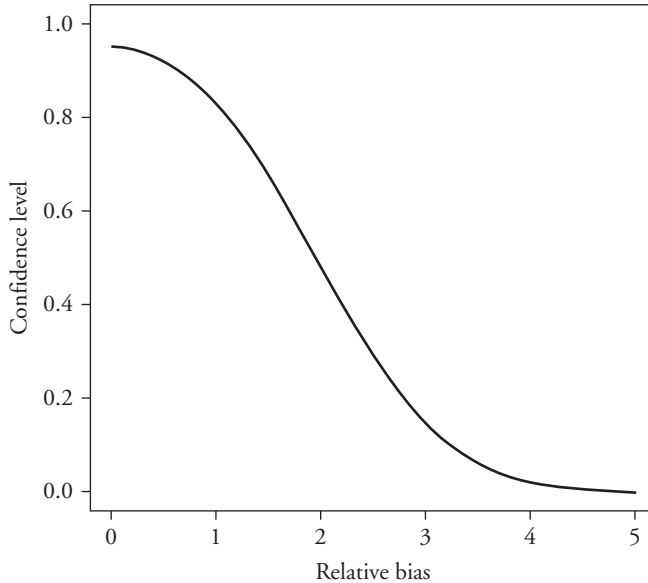


FIGURE 8.6 The confidence level of the 95% confidence interval as a function of the relative bias

is the result of such an attempt. See the study by Scherpenzeel (2008) for more information. This web panel has been constructed by selecting a random sample of households from the population register of the Netherlands. Selected households were recruited for this panel by means of a face-to-face interview (CAPI) or a telephone interview (CATI). Cooperative households without Internet access were provided with equipment giving them access to Internet. It should be noted, however, that there is always a (small) group that refuses to use the Internet. This group usually comprises the elderly. Analysis by Scherpenzeel and Bethlehem (2011) showed nevertheless that this approach reduces the undercoverage bias.

A second approach is to conduct a mixed-mode survey. In this survey different modes of data collection are used. Possible modes are face-to-face interviewing (CAPI), telephone interviewing (CATI), and self-administered modes like mail and web. De Leeuw (2005) describes two-mixed mode designs. The first design is use of different modes *concurrently*. The sample is divided into groups, and each group is approached by a different mode. The other approach is to use different modes *sequentially*. All sample persons are approached by one mode. The nonrespondents are then followed up by a different mode than the one used in the first approach. This process can be repeated for several modes.

The undercoverage problem of web surveys could be addressed by a sequential mixed-mode design. First, a sample is selected. If it turns out that selected persons are willing to participate in a survey but do not have access to

the Internet; they are, for the time being, considered to be nonrespondents. In the second phase of the fieldwork, these nonrespondents are approached with a different mode of data collection. The cheapest one would be a mail survey, but if quality is vital, CATI or CAPI should be preferred. The collected data of the two phases of the fieldwork are combined into an estimate for the target population.

A third approach is to apply adjustment weighting. This family of techniques attempts to reduce the bias of survey estimates by assigning weights to responding elements. These weights correct for the over- or underrepresentation of specific groups in the response. Adjustment weighting is treated in detail in Chapter 10. Here one such technique is summarized. It is called poststratification.

Poststratification requires one or more auxiliary variables. An auxiliary variable is a variable that has been measured in the survey, and for which the distribution in the target population is available. Typical target variables are gender, age, marital status, and region. By comparing the response distribution of an auxiliary variable with its population distribution, it can be assessed whether the survey response is representative for the population (with respect to this variable). If these distributions differ considerably, one must conclude that the response is selective. To correct this, adjustment weights are computed. Weights are assigned to all records of observed elements. Estimates of population characteristics can now be obtained by using the weighted values instead of the unweighted values.

To carry out poststratification, one or more categorical auxiliary variables are needed. Here, only one such variable is considered. The situation for more variables is not essentially different. Suppose there is an auxiliary variable X having L categories. So it divides the target population into L strata. The strata are denoted by the subsets U_1, U_2, \dots, U_L of the population U . The number of target population elements in stratum U_b is denoted by N_b , for $b = 1, 2, \dots, L$. The population size N is equal to $N = N_1 + N_2 + \dots + N_L$. This population information is assumed to be available.

Suppose a simple random sample of size n is selected from the Internet population. If n_b denotes the number of sample elements in stratum b , then $n = n_1 + n_2 + \dots + n_L$. The values of the n_b are the result of a random selection process, so they are random variables. Note that because the sample is selected from the Internet population, only elements in the substrata $U_I \cap U_b$ are observed (for $b = 1, 2, \dots, L$).

Poststratification assigns identical adjustment weights to all elements in the same stratum. The weight w_k for an element k in stratum b is equal to

$$(8.17) \quad w_k = \frac{N_b/N}{n_b/n}.$$

Poststratification comes down to replacing the simple sample mean

$$(8.18) \quad \bar{y}_I = \frac{1}{n} \sum_{k=1}^N a_k I_k Y_k$$

by the weighted sample mean

$$\bar{y}_{I,PS} = \frac{1}{n} \sum_{k=1}^N a_k w_k I_k Y_k.$$

Substituting the weights and working out this expression leads to the post-stratification estimator

$$(8.20) \quad \bar{y}_{I,PS} = \frac{1}{N} \sum_{b=1}^L N_b \bar{y}_I^{(b)} = \sum_{b=1}^L W_b \bar{y}_I^{(b)},$$

where $\bar{y}_I^{(b)}$ is the sample mean in stratum b and $W_b = N_b / N$ is the relative size of stratum b . The expected value of this poststratification estimator is equal to

$$(8.21) \quad E(\bar{y}_{I,PS}) = \frac{1}{N} \sum_{b=1}^L N_b E(\bar{y}_I^{(b)}) = \sum_{b=1}^L W_b \bar{Y}_I^{(b)} = \tilde{Y}_I,$$

where $\bar{Y}_I^{(b)}$ is the mean of the target variable in stratum b of the Internet population. Generally, this mean will not be equal to the mean $\bar{Y}^{(b)}$ of the target variable in stratum b of the target population. The bias of this estimator is equal to

$$(8.22) \quad \begin{aligned} B(\bar{y}_{I,PS}) &= E(\bar{y}_{I,PS}) - \bar{Y} = \tilde{Y}_I - \bar{Y} = \sum_{b=1}^L W_b (\bar{Y}_I^{(b)} - \bar{Y}^{(b)}) \\ &= \sum_{b=1}^L W_b \frac{N_{NI,b}}{N_b} (\bar{Y}_I^{(b)} - \bar{Y}_{NI}^{(b)}), \end{aligned}$$

where $N_{NI,b}$ is the number of elements in stratum b of the non-Internet population.

The bias will be small if there is (on average) no difference between elements with and without Internet within the strata. This is the case if there is a strong relationship between the target variable Y and the stratification variable X . The variation in the values of Y will manifest itself in this case between strata but not within strata. In other words, the strata are homogeneous with respect to the target variable. In nonresponse correction terminology, this situation comes down to missing at random (MAR).

Application of poststratification will successfully reduce the bias of the estimator if proper auxiliary variables can be found. Such variables should satisfy three conditions:

- They have to be measured in the survey (or complete sample).
- Their population distribution (N_1, N_2, \dots, N_L) must be known.
- They must be strongly correlated with all target variables.

Unfortunately, such variables are not available very often, or there is only a weak correlation.

8.2.4 MIXED-MODE DATA COLLECTION

The fundamental problem of a web survey is that persons without Internet access are excluded from the survey. This problem could be solved by selecting a stratified sample. The target population is assumed to consist of two strata: the Internet population U_I of size N_I and the non-Internet population U_{NI} of size N_{NI} .

To be able to compute an unbiased estimate, a simple random sample must be selected from both strata. The web survey provides the data about the Internet stratum. If this is a random sample with equal probabilities, the sample mean

$$(8.23) \quad \bar{y}_I = \frac{1}{n} \sum_{k=1}^N a_k I_k Y_k$$

is an unbiased estimator of the mean of the Internet population.

Now suppose a random sample (with equal probabilities) of size m is selected from the non-Internet stratum. Of course, there is no sampling frame for this population. This problem could be avoided by selecting a sample from the complete target population (a reference survey) and only using people without Internet access. Selected people with Internet access can be added to the large online sample, but this will have no substantial effect on estimators. The sample mean of the non-Internet sample is denoted by

$$(8.24) \quad \bar{y}_{NI} = \frac{1}{m} \sum_{k=1}^N b_k (1 - I_k) Y_k,$$

where the indicator b_k denotes whether element k is selected in the non-Internet survey, and

$$(8.25) \quad m = \sum_{k=1}^N b_k (1 - I_k).$$

The stratification estimator is now defined by

$$(8.26) \quad \bar{y}_{ST} = \frac{N_I}{N} \bar{y}_I + \frac{N_{NI}}{N} \bar{y}_{NI}.$$

This is an unbiased estimator for the mean of the target population. Application of this estimator assumes the size N_I of the Internet population and the size N_{NI} of the non-Internet population to be known. The variance of the estimator is equal to

$$(8.27) \quad V(\bar{y}_{ST}) = \left(\frac{N_I}{N}\right)^2 V(\bar{y}_I) + \left(\frac{N_{NI}}{N}\right)^2 V(\bar{y}_{NI}).$$

The variance of the sample mean in the Internet stratum is of order $1/n$, and the variance in the non-Internet stratum is of order $1/m$. As m will be much smaller than n in practical situations, and the relative sizes of the Internet-population and

the non-Internet-population do not differ that much, the second term will determine the magnitude of the variance. So the advantages of the large sample size of the web survey are for a great part lost by the bias correction.

Note that the sizes of the Internet and the non-Internet population are usually unknown. In this case they have to be estimated. This can, for example, be done using data from the non-Internet survey.

8.3 Application

The possible consequences of undercoverage and the effectiveness of correction techniques are now illustrated using a simulation experiment. A fictitious population was constructed. For this population, reported voting behavior in an election survey was simulated and analyzed.

The relationship between variables involved was such that it could resemble more or less a real-life situation. This relationship is shown graphically in Figure 8.7.

With respect to the Internet population, both *missing at random* (MAR) and *not missing at random* (NMAR) were introduced. The characteristics of estimators (before and after correction) were computed based on a large number of simulations.

First, the distribution of the estimator was determined in the ideal situation of a simple random sample from the target population. Then, it was explored how the characteristics of the estimator change if a simple random sample is selected just from the Internet population. Finally, the affects of weighting (poststratification) were analyzed.

A fictitious population of 30,000 individuals was constructed. There were five variables:

- Age in three categories: Young (with probability 0.40), Middle aged (with probability 0.35), and Old (with probability 0.25).
- Ethnic origin in two categories: Native (with probability 0.85) and Non-native (with probability 0.15).
- Having access to the Internet with two categories Yes and No. The probability of having access to the Internet depended on the two variables Age and Ethnic origin. For natives, the probabilities were 0.90 (for Young), 0.70 (for

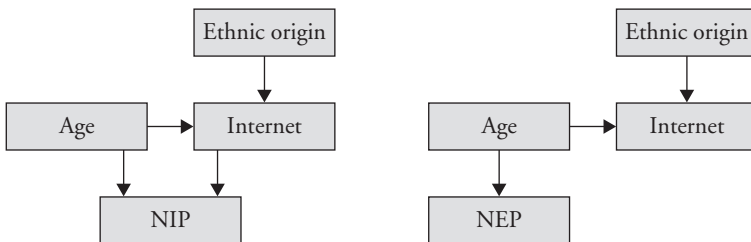


FIGURE 8.7 Relationships between variables

Middle aged), and 0.50 (for Old). So, Internet access decreased with age. For non-natives, these probabilities were 0.20 (for Young), 0.10 (for Middle aged), and 0.00 (for Old). These probabilities reflected the much lower Internet access among non-natives.

- Voted for the National Elderly Party (NEP). The probability to vote for this party depended on age only. Probabilities were 0.00 (for Young), 0.40 (for Middle aged), and 0.60 (for Old).
- Voted for the New Internet Party (NIP). The probability to vote for this party depended on both age and having Internet access. For people with Internet access, the probabilities were 0.80 (for Young), 0.40 (for Middle aged), and 0.20 (for Old). For people without Internet access, all probabilities were equal to 0.10. So, for people with Internet access, voting decreased with age. Voting probability was low for people without Internet.

In the experiment, the variable NEP (National Elderly Party) suffered from missingness because of MAR. There was a direct relationship between voting for this party and age, and there was a direct relationship between age and having Internet access. This will cause estimates to be biased. It should be possible to correct for this bias by weighting using the variable age.

The variable NIP (New Internet Party) suffered from NMAR. There existed (among other relationships) a direct relationship between voting for this party and having Internet access. As a result, estimates will be biased, and no correction is possible.

The distribution of estimators for the percentage of voters for both parties was determined in various situations by repeating the selection of the sample 1,000 times. In all cases, the sample size was $n=2,000$.

Figure 8.8 contains the results for the variable NEP (voted for the National Elderly Party). The distributions of the estimator are displayed by means of box plots. The upper box plot shows the distribution of the estimator for a simple random sample from the complete target population. The vertical line denotes the population value to be estimated (25.4%). The estimator has a symmetric distribution around this value. The estimator is clearly unbiased.

The middle box plot shows the distribution of the estimator if samples are not selected from the complete target population but just from the Internet population. The shape of the distribution remains the same, but the distribution as a whole has shifted to the left. All values of the estimator are systematically lower. The expected value of the estimator is only 20.3%. The estimator is biased. The explanation of this bias is simple: Relatively few elderly have Internet access. Therefore, they are underrepresented in samples selected from the Internet. These persons typically vote for the NEP.

The lower box plot shows the distribution of the estimator in the case of poststratification by age. The bias disappears. This was possible because this is a case of MAR.

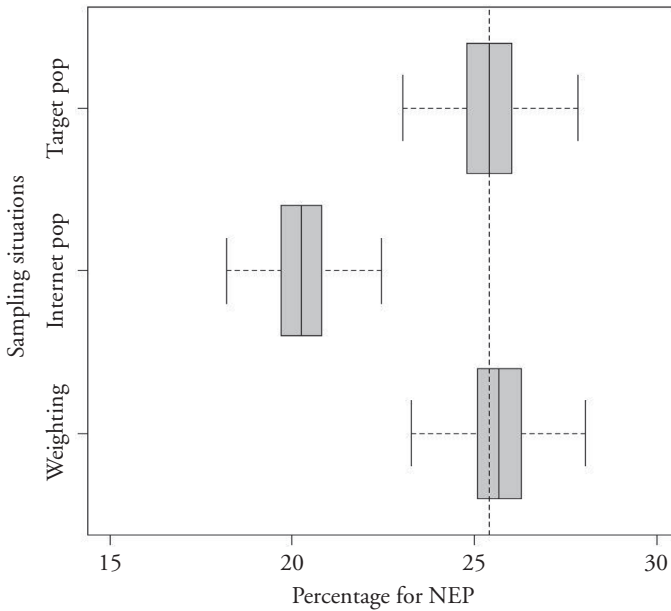


FIGURE 8.8 Results of the simulations for variable NEP (National Elderly Party)

TABLE 8.1 Summary of simulation results for the variable NEP

Simulation	Mean	Standard error	Bias	Relative bias
Samples from the target population	25.4	0.9	0.0	0.0
Samples from the Internet population	20.3	0.8	-5.1	-6.4
Samples from the Internet population, with weighting adjustment	25.4	0.8	0.0	0.0

The simulation results are summarized in Table 8.1. Sampling from the Internet results in a large relative bias of -6.4 . Weighting eliminates the bias, and therefore, the relative bias is 0. In the case of sampling from the Internet, the confidence level of the 95% confidence interval is close to 0. This implies that almost certainly a wrong conclusion will be drawn from this web survey.

Figure 8.9 contains the results for the variable NIP (voted for the New Internet Party). The upper box plot shows the distribution of the estimator for simple random samples from the complete target population. The vertical line denotes the population value to be estimated (39.5%). As the estimator has a symmetric distribution around this value, it is clear that the estimator is unbiased.

The middle box plot shows what happens if samples are not selected from the complete target population but just from the Internet population. The

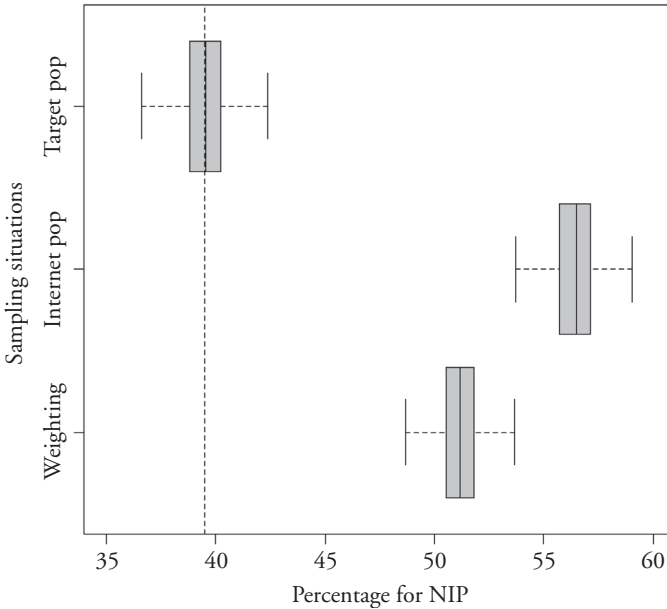


FIGURE 8.9 Results of the simulations for variable NIP (New Internet Party)

TABLE 8.2 Summary of simulation results for the variable NIP

Simulation	Mean	Standard error	Bias	Relative bias
Samples from the target population	39.5	1.1	0.0	0.0
Samples from the Internet population	56.5	1.1	17.0	15.5
Samples from the Internet population, with weighting adjustment	51.1	1.0	11.6	11.6

distribution has shifted to the right considerably. All values of the estimator are systematically too high. The expected value of the estimator is now 56.5%. The estimator is severely biased. The explanation of this bias is straightforward: Voters for the NIP are overrepresented in Internet samples.

The lower box plot in Figure 8.9 shows the effect of poststratification by age. Only a small part of the bias is removed. This is not surprising as there is a direct relationship between voting for the NIP and having access to the Internet. This is a case of NMAR.

The simulation results are summarized in Table 8.2. Sampling from the Internet results in a very large relative bias of 15.5. Weighting only can reduce this relative bias to 11.6. In both cases, the confidence level of the 95% confidence interval is close to 0. This implies that almost certainly a wrong conclusion will be drawn from this web survey.

8.4 Summary

A web survey may suffer from undercoverage. This happens if not all elements in the population have access to the Internet. Elements without Internet access will never be selected in the sample.

Undercoverage may cause estimates of population characteristics to be biased. The magnitude of the bias is determined by two factors:

1. The relative size of the group without Internet access. The larger the group, the larger the bias.
2. The difference between those with and without Internet access. The larger the difference (on average). The larger the difference, the larger the bias.

Several approaches may help to reduce bias from undercoverage. A first approach is to provide Internet access to those in the sample without it. This may not completely solve the problem as there may still be persons refusing to work with the Internet.

A second approach is to conduct a mixed-mode survey. A different mode of data collection (face-to-face, telephone, or mail) can be used for those without Internet access.

A third approach is to carry out some kind of adjustment weighting. By assigning weights, the response is corrected for under- or overrepresented groups. There is no guarantee that weighting will completely remove the bias.

KEY TERMS

Contrast: The difference between the average of the target variable in the Internet population and the average of this variable in the non-Internet population.

First-order inclusion probability: The probability that a population element is selected in the sample. The first-order inclusion probability is determined by the sampling design.

Frame population: All elements that are represented in the sampling frame.

Internet population: The subpopulation of the target population consisting of elements that have access to the Internet.

Missing at Random (MAR): Nonresponse depends on auxiliary variables only. Estimators will be biased, but a correction is possible if some technique is used that takes advantage of this auxiliary information.

Missing Completely at Random (MCAR): Nonresponse happens completely independent of all survey variables. The estimators will not be biased.

Mixed-mode survey: A survey in which various modes of data collection are combined. Modes can be used concurrently (different groups are approached by

different modes) or sequentially (nonrespondents of a mode are reapproached in a different mode).

Non-Internet population: The subpopulation of the target population consisting of elements that do not have access to the Internet.

Not Missing at Random (NMAR): Nonresponse depends directly on the target variables of the survey. The estimators will be biased, and the correction techniques will not be successful.

Overcoverage: The phenomenon that the sampling frame contains elements that do not belong to the target population.

Poststratification: A weighting adjustment technique that divides the population in strata and subsequently assigns the same weight to all observed elements within a stratum.

Sampling frame: A list (electronic or on paper) of all elements in the target population. Using the information in the sampling frame, it must be possible to contact each element.

Undercoverage: The sampling frame does not cover completely the target population of the survey. There are persons in the population who do not appear in the sampling frame. They will never be selected in the sample.

EXERCISES

Exercise 8.1. What is the difference between undercoverage and nonresponse?

- a. In the case of nonresponse, persons are selected in the sample, and in the case of undercoverage, they are never selected.
- b. In the case of undercoverage, persons are selected in the sample, and in the case of nonresponse, they are never selected.
- c. Nonresponse is caused by the persons selected in the sample, and undercoverage by the researcher.
- d. There is no difference between undercoverage and nonresponse;

Exercise 8.2. Can the bias from undercoverage always be corrected by means of a weighting adjustment?

- a. Yes, weighting adjustment will always remove the bias.
- b. No, weighting adjustment will only work if persons are missing because of MAR.
- c. No, weighting adjustment will only work if persons are missing because of MAR, and the proper auxiliary variables are included in the weighting model.
- d. No, weighting adjustment will never remove or reduce such a bias.

Exercise 8.3. What happens to the bias from undercoverage in web surveys if Internet access increases in the target population?

- a. The bias will increase.
- b. The bias will decrease.
- c. The bias will not change.
- d. It depends on the average difference between persons with and without Internet access.

Exercise 8.4. What happens to the bias from undercoverage in web surveys if the sample size is increased?

- a. The bias will increase.
- b. The bias will decrease.
- c. The bias will not change.
- d. It depends on the average difference between persons with and without Internet access.

Exercise 8.5. A researcher wants to estimate the average number of hours per week the adult inhabitants of Samplonia spend on the Internet. He draws a simple random sample of Internet users. There is no nonresponse. The sample mean turns out to be 5 hours.

- a. Given that only three out of five inhabitants have access to the Internet, compute an estimate of the bias of the sample mean.
- b. Compute a better estimate for the average number of hours an inhabitant spends on the Internet.

Exercise 8.6. A town council wants to know what percentage of the population is engaged in some form of voluntary work. As only a limited budget is available, it is decided to conduct an on-line survey. The target population consists of 1,000,000 persons. Only 70% of these persons have access to the Internet. It turns out that 10,000 persons participate in the survey. Of these respondents, 70% does some voluntary work.

- a. Assuming that the 10,000 respondents are a simple random sample without replacement from the target population, compute the 95% confidence interval of the percentage of persons in the population doing voluntary work. There is a strong suspicion that the survey estimates may be biased because only people with Internet access can participate. Therefore, a follow-up survey is conducted among people without Internet access. It turns out to be possible to draw a simple random sample of size 100 from this non-Internet population. The result is that 40% of the respondents in the follow-up survey do voluntary work.

- b. Compute an improved estimate for the population percentage of people involved in voluntary work.
- c. Compute a new 95% confidence interval of the percentage of persons in the population doing voluntary work.
- d. Compare both confidence intervals and explain any differences.

REFERENCES

- Bethlehem, J. G. & Hofman, L. P. M. B. (2006), *Blaise–Alive and Kicking for 20 Years*. Proceedings of the 10-th Blaise, Users Meeting, Statistics Netherlands, Voorburg/Heerlen, the Netherlands, pp. 61–88.
- Couper, M. P. (2000), Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, pp. 464–494.
- Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls II, W. L., & O'Reilly, J. M. (eds.) (1998), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.
- De Haan, J. & Van 't Hof, C. (2006), *Jaarboek ICT en Samenleving, De Digitale Generatie*. Netherlands Institute for Social Research/SCP, The Hague, the Netherlands.
- De Leeuw, E. D. (2005), To Mix or Not To Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21, pp. 233–255.
- Dillman, D. A. & Bowker, D. (2001), The Web Questionnaire Challenge to Survey Methodologists. In: Reips, U. D. and Bosnjak, M. (eds.), *Dimensions of Internet Science*, Pabst Science Publishers, Lengerich, Germany.
- Eurostat. (2007), *More Than 40% of Households have Broadband Internet Access*. Eurostat News Release 166/2007. Eurostat, Luxembourg.
- Horvitz, D. G. & Thompson, D. J. (1952), A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, pp. 663–685.
- Scherpenzeel, A. (2008), An Online Panel as a Platform for Multi-disciplinary Research. In: Stoop, I. & Wittenberg, M. (eds.), *Access Panels and Online Research, Panacea or Pitfall?* Aksant, Amsterdam, pp. 101–106.
- Scherpenzeel, A. & Bethlehem, J. (2011), How Representative are Online-panels? Problems of Coverage and Selection and Possible Solutions. In: Das, M., Ester, P., Kaczmirek, L., & Mohler, P. (eds.), *Social Research and the Internet: Advances in Applied Methods and New Research Strategies*. Routledge Academic, New York, pp. 105–132.

The Problem of Self-Selection

9.1 Introduction

Web surveys are a fast, cheap, and attractive means of collecting large amounts of data. Not surprisingly, many survey organizations are conducting such surveys. The question is, however, whether a web survey is also attractive from a quality point of view because there are methodological problems. One of these problems is *self-selection*. With this phenomenon, the sample is not a probability sample. Instead, it is left to the Internet users themselves to participate in a web survey. Estimation problems caused by self-selection are the topic of this chapter. After an introduction, some theory is described. It is also explored whether weighting adjustment techniques can help to solve the problem. Practical implications are shown using simulated samples from a fictitious population.

The objective of a survey is to collect information about a well-defined target population. To this end, a sample is selected from this population. The methodology of survey sampling has been developed over a period of more than 100 years. It is based on the fundamental principle of probability sampling. Selecting random samples makes it possible to apply probability theory. Unbiased estimators can be defined, and the accuracy of these estimators can be quantified and controlled. The probability sampling principle has been successfully applied in official and academic statistics since the 1940 and, to a lesser extent, in more commercial market research. See Chapter 1 for a historical overview of the development of survey sampling.

Horvitz and Thompson (1952) show in their seminal paper that unbiased estimates of population characteristics can be computed only if a real probability sample has been selected, every element in the population has a nonzero

probability of selection, and all these probabilities are known to the researcher. Furthermore, the accuracy of estimates can be computed only under these conditions.

At first sight, web surveys seem to have much in common with other types of surveys. It is just another mode of data collection. Questions are not asked face-to-face, by telephone, or on paper but over the Internet. What is different, however, is that many web surveys are self-selection surveys. The principles of probability sampling have not been applied. Samples are not constructed by means of probability sampling but instead rely on self-selection of respondents. This can have a major impact on survey results.

Web surveys appear in many different forms, from simple e-mail surveys to professionally designed interactive forms. Of course, web surveys can be based on probability sampling. An example is a survey among students of a university, where every student has an e-mail address. So a random sample can be selected from the list of all e-mail addresses. Unfortunately, many web surveys, particularly those conducted by market research organizations, are not based on probability sampling. The survey questionnaire is simply put on the web. Respondents are those people who happen to have Internet access, visit the website, and decide to participate in the survey. The survey researcher is not in control over the selection process. Therefore, no unbiased estimates can be computed nor can the accuracy of estimates be determined.

■ EXAMPLE 9.1 Opinion Polls in the Netherlands

All major opinion polls in the Netherlands use web panels that have been set up by means of self-selection. Examples are the *Politieke Barometer* and *Peil.nl*. The values of some demographic variables are recorded during the recruitment phase. Therefore, the distribution of these variables in a poll can be compared with their distribution in the population. Weighting adjustment techniques can be applied in an attempt to correct for over- or underrepresentation of specific groups.

Another example of a large self-selection web survey in the Netherlands was *21minuten.nl*, a survey that was supposed to supply answers to questions about important problems in Dutch society. The first edition of this survey was conducted in 2006. Within a period of six weeks, approximately 170,000 people completed the on-line questionnaire. A similar survey was conducted in Germany (*Perspektive Deutschland*).

Vonk, Van Ossenbruggen, and Willems (2006) describe a study across 19 on-line panels of Dutch market research organizations. It shows that most of them use self-selection.

There was an intensive political discussion in the Netherlands in January 2010 about the introduction of a system of road pricing. An important participant in this discussion was the Dutch Automobile Association (ANWB). This organization conducted a poll on its website.

It was a self-selection survey. Everyone could participate. Everyone could participate even more than once. There was no check on this. Within a period of a few weeks, the questionnaire was completed more than 400,000 times. In the same period, the Dutch newspaper *De Telegraaf* conducted a simple self-selection web survey on the same topic on its website. In one weekend, the questionnaire was completed approximately 196,000 times. As this newspaper is known to support the interests of car owners, it was not surprising that the great majority (89%) turned out to be against road pricing.

Self-selection web survey results are sometimes claimed to be “representative” in the media because of the large number of respondents or as a result of advanced adjustment weighting procedures. This claim was typically made for some of the web surveys mentioned in Example 9.1. Unfortunately, such claims are not based on methodological knowledge.

The term *representative* is confusing. Kruskal and Mosteller (1979a, 1979b, 1979c) show that it can have many meanings and that it is often used in a very loose sense to convey a vague idea of good quality. It is even sometimes claimed that a large number of respondents ensures validity and reliability. Unfortunately, it is a well-known fact in the survey methodology literature that this is not the case. It is shown again in this chapter.

The essential problem of self-selection is that the selection probabilities are unknown. Some of these probabilities may even be equal to 0. This makes it impossible to construct unbiased estimators using the theory of Horvitz and Thompson (1952).

The problem of self-selection is illustrated using survey results related to the general elections in the Netherlands in 2006. Various market research organizations carried out opinion polls in an attempt to predict the outcome of these elections. The results of the three major polls are summarized in Table 9.1. The polls were conducted one day before the elections. The table contains the numbers of seats in parliament. The number of seats is directly related to the percentages of votes.

Politieke Barometer, *Peil.nl*, and *De Stemming* are opinion polls based on samples from self-selection web panels. To reduce a possible bias, adjustment weighting has been carried out. DPES is the Dutch Parliamentary Election Study. The fieldwork for this comprehensive survey was carried out by Statistics Netherlands. The sample was a true probability sample. It was selected from the population register. The mode of data collection for DPES was CAPI (face-to-face interviewing with laptops).

Bold numbers in Table 9.1 denote predictions differing three seats or more from the true result. This happened only for the web panels, not for the election survey. These predictions were even considered unsatisfactory by the organizations that produced them. It is clear that in this example the probability-sampling-based DPES outperformed the self-selection surveys.

TABLE 9.1 The Dutch Parliamentary elections 2006: Comparison of various opinion surveys with the official results

	Election result	Politieke Barometer	Peil.nl	De Stemming	DPES 2006
Sample size		1,000	2,500	2,000	2,600
Seats in parliament:					
CDA (christian democrats)	41	41	42	41	41
PvdA (social democrats)	33	37	38	31	32
VVD (liberals)	22	23	22	21	22
SP (socialists)	25	23	23	32	26
GL (green party)	7	7	8	5	7
D66 (liberal democrats)	3	3	2	1	3
ChristenUnie (christian)	6	6	6	8	6
SGP (christian)	2	2	2	1	2
PvdD (Animal party)	2	2	1	2	2
PvdV (Conservative)	9	4	5	6	8
Other parties	0	2	1	2	1
Mean Absolute Difference		1.27	1.45	2.00	0.36

Probability sampling has the additional advantage that it provides protection against certain groups in the population attempting to manipulate the outcomes of the survey. This may typically play a role in opinion polls. Self-selection does not have this safeguard. There are several examples where organizations attempted to influence the outcomes of surveys by advising their members to participate in it.

An example of this effect could be observed in the election of the 2005 Book of the Year Award (Dutch: NS Publieksprijs), a high-profile literary prize in the Netherlands. The winning book was determined by means of a poll on a website. People could vote for one of the nominated books or mention another book of their choice. More than 90,000 people participated in the survey. The winner turned out to be the new interconfessional Bible translation published by the Netherlands and Flanders Bible societies. This book was not nominated, but nevertheless an overwhelming majority (72%) voted for it. This was a result of a campaign launched by (among others) Bible societies, a Christian broadcaster, and a Christian newspaper. Although this was all completely within the rules of the poll, the group of voters was not representative of the Dutch population as a whole.

9.2 Theory

9.2.1 BASIC SAMPLING THEORY

Let the target population U of the survey consist of N identifiable elements, which are labeled $1, 2, \dots, N$. Therefore, the target population can be denoted by

$$U = \{1, 2, \dots, N\}. \quad (9.1)$$

Associated with each element k is a value Y_k of the target variable Y . The aim of the web survey is assumed to be estimation of the population mean

$$\bar{Y} = \frac{1}{N} \sum_{k=1}^N Y_k \quad (9.2)$$

of the target variable Y .

Suppose a probability sample is selected without replacement. It means that each element can appear at most once in the sample. Therefore, the sample can be represented by a set of indicators.

$$a = a_1, a_2, \dots, a_N. \quad (9.3)$$

The k th indicator a_k assumes the value 1 if element k is selected in the sample, and otherwise it assumes the value 0. The expected value (i.e., the mean value over all possible samples) of a_k is denoted by

$$\pi_k = E(a_k). \quad (9.4)$$

The quantity π_k is called the *first-order inclusion probability* of element k (for $k = 1, 2, \dots, N$). For deriving variance formulas, second-order inclusion probabilities also are required. The *second-order inclusion probability* of elements k and l (with $k \neq l$) is equal to

$$\pi_{kl} = E(a_k a_l) \quad (9.5)$$

and, by definition, $\pi_{kk} = \pi_k$. The *sample size*, (i.e., the number of selected elements) is denoted by n . Because the indicators a_k have the value 1 for all elements in the sample, and the value 0 for all other elements, the sample size can be written as the sum of the values of the indicators:

$$n = \sum_{k=1}^N a_k. \quad (9.6)$$

The Horvitz–Thompson estimator is defined by

$$\bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^N a_k \frac{Y_k}{\pi_k}. \quad (9.7)$$

The indicators a_k filter out the sample values of the target variable. Note that each value Y_k is weighted with its inverse selection probability π_k . Thus, the

estimator is corrected for the that elements with a large inclusion probability are overrepresented in the sample.

The Horvitz–Thompson estimator is an unbiased estimator of the population mean. The variance of this estimator is equal to

$$V(\bar{y}_{HT}) = \frac{1}{N^2} \sum_{k=1}^N \sum_{l=1}^N (\pi_{kl} - \pi_k \pi_l) \frac{Y_k Y_l}{\pi_k \pi_l}. \quad (9.8)$$

For without replacement samples of fixed size n , the variance can be rewritten in the form

$$V(\bar{y}_{HT}) = \frac{1}{2N^2} \sum_{k=1}^N \sum_{l=1}^N (\pi_k \pi_l - \pi_{kl}) \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l} \right)^2. \quad (9.9)$$

This expression shows that the variance can be reduced by taking the first-order inclusion probabilities as much as possible proportional to the values of the target variable.

The *variance of the estimator* is just one way to quantify the precision of an estimator. A small variance means a high precision and a large variance a small precision. Another way to quantify the precision is the *standard error*. It is defined by

$$S(\bar{y}_{HT}) = \sqrt{V(\bar{y}_{HT})}. \quad (9.10)$$

The standard error is required to compute the confidence interval. The *confidence interval* is a range of possible values of the population mean. The interval encompasses the true value of the population mean with a high probability if an estimator is unbiased. This probability is called the *confidence level*. It is denoted by $(1 - \alpha)$, where α is a small probability. Often the value $\alpha=0.05$ is used, corresponding to a confidence level of 95%.

The distribution of many estimators (including the sample mean) can for large (probability) sample sizes be approximated by a normal distribution. This makes it easier to compute confidence intervals. Only the standard error of the estimator is required. The 95% confidence interval of the Horvitz–Thompson estimator is equal to

$$(\bar{y}_{HT} - 1.96 \times S(\bar{y}_{HT}); \bar{y}_{HT} + 1.96 \times S(\bar{y}_{HT})). \quad (9.11)$$

The best known and probably most often used type of probability sample is a *simple random sample without replacement*. First-order inclusion probabilities of all elements are equal for this type of sample. It can be shown that all first-order

inclusion probabilities are equal to n/N . Furthermore, all second-order inclusion probabilities are equal to $n(n-1)/N(N-1)$. Substitution of these values of the inclusion probabilities in expression (9.7) results in a simple estimator, the *sample mean*

$$\bar{y} = \frac{1}{n} \sum_{k=1}^N a_k Y_k = \frac{1}{n} \sum_{i=1}^n y_i, \quad (9.12)$$

where y_1, y_2, \dots, y_n denote the n observations that have become available in the sample. This is an unbiased estimator with variance

$$V(\bar{y}) = \frac{1-f}{n} S^2, \quad (9.13)$$

where $f = n/N$ is the *sampling fraction* and S^2 is the population variance, defined by

$$S^2 = \frac{1}{N-1} \sum_{k=1}^N (Y_k - \bar{Y})^2. \quad (9.14)$$

From expression (9.13) it is clear that an increased sample size produces more precise estimators.

9.2.2 A SELF-SELECTION SAMPLE FROM THE INTERNET POPULATION

The population U is divided into two subpopulations: a subpopulation U_I of elements having access to the Internet and a subpopulation U_{NI} of elements not having access to the Internet. Associated with each element k is an indicator I_k , where $I_k = 1$ if element k has access to the Internet ($k \in U_I$), and $I_k = 0$ otherwise ($k \in U_{NI}$). The subpopulation U_I is called the *Internet population* and the subpopulation U_{NI} is called the *non-Internet population*. Let

$$N_I = \sum_{k=1}^N I_k \quad (9.15)$$

denote the size of the Internet population U_I . The mean of the values of the target variable in the Internet population is defined by

$$\bar{Y}_I = \frac{1}{N_I} \sum_{k=1}^N I_k Y_k. \quad (9.16)$$

Likewise, $N_{NI} = N - N_I$ denotes the size of the subpopulation without the Internet, where $N_I + N_{NI} = N$. The mean of the values of the target variable in the non-Internet population is defined by

$$\bar{Y}_{NI} = \frac{1}{N_{NI}} \sum_{k=1}^N (1 - I_k) Y_k. \quad (9.17)$$

What happens if a self-selection sample is selected from the Internet? This section shows that estimators can be substantially biased and that this bias can be larger than the bias caused by nonresponse in surveys based on probability samples.

Participation in a self-selection requires that respondents are aware of the existence of a survey (they have to visit the website accidentally, or they have to follow up a banner or an e-mail message). They also have to decide to fill in the questionnaire on the Internet. This means that each element k in the Internet population has unknown probability ρ_k of participating in the survey, for $k = 1, 2, \dots, N_I$.

The responding elements are denoted by a set of indicators

$$R_1, R_2, \dots, R_N, \quad (9.18)$$

where the k th indicator R_k assumes the value 1 if element k participates, and otherwise it assumes the value 0, for $k = 1, 2, \dots, N$. The expected value $\rho_k = E(R_k)$ is called the *response probability* of element k . For sake of convenience, response probabilities also are introduced for elements in the non-Internet population. By definition, the values of all these probabilities are 0. The realized sample size is denoted by

$$n_S = \sum_{k=1}^N R_k. \quad (9.19)$$

Lacking any knowledge about the values of the response probabilities, researchers usually implicitly assume all these probabilities to be equal. In other words, simple random sampling is assumed. Consequently, the sample mean

$$\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^N R_k Y_k \quad (9.20)$$

is used as an estimator for the population mean. The expected value of this estimator is approximately equal to

$$E(\bar{y}_S) \approx \tilde{Y} = \frac{1}{N_I \bar{\rho}} \sum_{k=1}^N \rho_k I_k Y_k, \quad (9.21)$$

where $\bar{\rho}$ is the mean of all response propensities in the Internet population. This expression was derived by Bethlehem (1988).

Using an approach similar to that of Cochran (1977, p. 31), it can be shown that the variance of estimator (9.20) is approximately equal to

$$V(\bar{y}) \approx \frac{1}{(N\bar{\rho})^2} \sum_{k=1}^N \rho_k(1 - \rho_k)(Y_k - \bar{Y})^2. \quad (9.22)$$

Note that this expression for the variance does not contain the sample size n (because no fixed size sample was drawn) but the expected sample size $N\bar{\rho}$. Not surprisingly, the variance decreases as the expected sample size increases.

It is clear from expression (9.21) that, generally, the expected value of the sample mean is not equal to the population mean of the Internet population. One situation in which the bias vanishes is that in which all response probabilities in the Internet population are equal. In terms of nonresponse correction theory, this comes down to missing completely at random (MCAR). This is the situation in which the cause of missing data is completely independent of all variables measured in the survey. For more information on MCAR and other missing data mechanisms, see Little and Rubin (2002). Indeed, in the case of MCAR, self-selection does not lead to an unrepresentative sample because all elements have the same selection probability.

Bethlehem (2002) shows that the bias of the sample mean (9.20) can be written as

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y}_I \approx \bar{Y} - \bar{Y}_I = \frac{C_{\rho Y}}{\bar{\rho}} = \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}}, \quad (9.23)$$

in which

$$C_{\rho Y} = \frac{1}{N_I} \sum_{k=1}^N I_k(\rho_k - \bar{\rho})(Y_k - \bar{Y}_I) \quad (9.24)$$

is the covariance between the values of target variable and the response probabilities in the Internet population and $\bar{\rho}$ is the average response probability. Furthermore, $R_{\rho Y}$ is the correlation coefficient between the target variable and the response behavior, S_{ρ} is the standard deviation of the response probabilities, and S_Y is the standard deviation of the target variable. The bias of the sample mean (as an estimator of the mean of the Internet population) is determined by three factors:

- The average response probability. If people are more likely to participate in the survey, the average response probability will be higher, and thus, the bias will be smaller.
- The relationship between the target variable and the response behavior. A strong correlation between the values of the target variable and the response probabilities will lead to a large bias.
- The variation in the response probabilities. The more these values vary, the larger the bias will be.

There are three situations in which this bias vanishes:

1. All response propensities are equal. Again, this is the case in which the self-selection process can be compared with a simple random sample.
2. All values of the target variable are equal. This situation is very unlikely to occur in practice. No survey would be necessary in this case. One observation would be sufficient.
3. There is no relationship between the target variable and the response behavior. It means participation does not depend on the value of the target variable.

Expression (9.23) for the bias of the estimator can be used to compute an upper bound for the bias. Given the mean response probability $\bar{\rho}$, there is a maximum value the standard deviation S_ρ of the response probabilities cannot exceed:

$$S(\rho) \leq \sqrt{\bar{\rho}(1 - \bar{\rho})}. \quad (9.25)$$

This implies that in the worst case, S_ρ assumes its maximum value and the correlation coefficient $R_{\rho Y}$ is equal to either $+1$ or -1 . Then the absolute value of the bias will be equal to

$$|B_{max}| = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1}. \quad (9.26)$$

This worst-case expression of the value of the bias also applies to the situation in which a probability sample has been drawn and subsequently non-response occurs in the fieldwork.

EXAMPLE 9.2 The worst-case bias in Dutch surveys

General-population surveys of Statistics Netherlands have response rates of around 70%. This means the absolute maximum bias is equal to

$$0.65 \times S_Y.$$

A large web survey in the Netherlands was *21minuten.nl*. This survey was supposed to provide answers to questions about important problems in the Dutch society. Within a period of six weeks in 2006, approximately 170,000 people completed the questionnaire (which took on average minutes). As everyone could participate in the survey, the target population was not defined properly. If it is assumed the target population consists of all Dutch citizens from the age of 18, the average response probability was $170,000/12,800,000 = 0.0133$. Hence, the absolute maximum bias is equal to

$$8.61 \times S_Y.$$

The conclusion is that the bias of the large web survey can be a factor 13 larger than the bias of the small probability survey.

Therefore, expression (9.26) provides a means to compare potential biases in various surveys.

It is important to note that the value of expression (9.22) does not depend on the sample size. Increasing the sample size will not reduce the bias. So the problem of self-selection bias in web surveys does not diminish by having more people completing the survey questionnaire.

The precision of an estimator often is quantified by a 95% confidence interval. Suppose a simple random sample is selected from a target population. Then the sample mean \bar{y} can be computed. This is an unbiased estimator for the population mean \bar{Y} . As the sample mean is (approximately) normally distributed, the 95% confidence interval for the population mean is equal to

$$I = (\bar{y} - 1.96 \times S(\bar{y}); \bar{y} + 1.96 \times S(\bar{y})), \quad (9.27)$$

where $S(\bar{y})$ is the standard error of the sample mean. The probability that this interval contains the true value is by definition (approximately) equal to

$$P(\bar{Y} \in I) = 0.95. \quad (9.28)$$

The standard error decreases with an increasing sample size. Therefore, the width of interval (9.27) is smaller for a larger sample.

If a self-selection sample is selected from the Internet population, the sample mean \bar{y}_S is used to estimate the population mean \bar{Y} . Analogous to expression (9.27), the confidence will be computed as

$$I_S = (\bar{y}_S - 1.96 \times S(\bar{y}_S); \bar{y}_S + 1.96 \times S(\bar{y}_S)), \quad (9.29)$$

The confidence level of this interval is not by definition equal to 0.95. It can be shown that

$$P(\bar{Y} \in I_S) = \Phi\left(1.96 - \frac{B(\bar{y}_S)}{S(\bar{y}_S)}\right) - \Phi\left(-1.96 - \frac{B(\bar{y}_S)}{S(\bar{y}_S)}\right), \quad (9.30)$$

in which Φ is the standard normal distribution function. The quantity $B(\bar{y}_S)/S(\bar{y}_S)$ is called the *relative bias*. Apparently, the confidence level depends on the value of

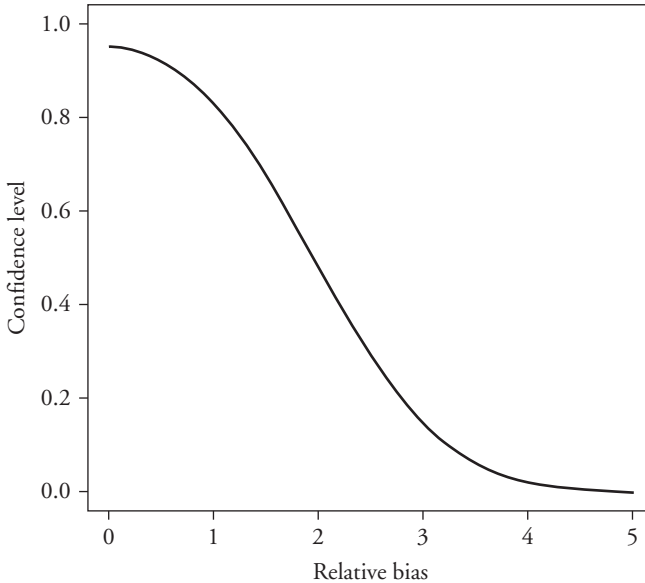


FIGURE 9.1 The confidence level of the 95% confidence interval as a function of the relative bias.

this relative bias. Figure 9.1 contains a plot of the confidence level as a function of the relative bias.

An increased sample size will reduce the standard error, but the bias remains the same. Hence, the relative bias will increase. It is clear that the confidence level can be much lower than expected. If the bias is equal to the standard error (i.e., the relative bias is 1), the confidence level is only 0.83. As the relative bias increases, the situation becomes worse. The confidence level is even less than 0.5 for a relative bias of 2. The conclusion is that self-selection may lead to an incorrect confidence interval.

9.2.3 REDUCING THE SELF-SELECTION BIAS

There are several ways in which the negative effects of self-selection can be reduced. Three approaches are discussed here.

The first approach is to avoid self-selection by selecting a proper probability sample. It is possible to conduct a web survey that is based on probability sampling. This requires a sampling frame. Sometimes such sampling frames are available. An example is a survey among employees of a company, where every employee has a company-assigned e-mail address. The sampling frame for this example consists of the list of e-mail addresses. The situation is not so straightforward for a general-population survey. Unfortunately, population registers do not contain e-mail addresses. A solution can be to approach sampled persons by some other mode. One option is to send them a letter with the request

to go to a specific website, where they can complete the online questionnaire form. Such a letter also should contain a unique identification code that has to be entered. Use of such identifying codes guarantees that only sampled persons respond, and that they respond only once. Another option is to approach sampled persons face-to-face (CAPI) or by telephone (CATI) and to ask them for their e-mail address (if they want to participate). If they cooperate, they are sent a link to the on-line questionnaire form.

■ EXAMPLE 9.3 The LISS panel

The Dutch Longitudinal Internet Study for Social Sciences (LISS) panel is the result of an attempt to set up a web panel where the panel members are recruited by means of probability sampling. See Scherpenzeel (2008) for a detailed description of this panel.

The panel has been constructed by selecting a random sample of households from the population register of the Netherlands. Selected households were recruited for this panel by means of a face-to-face interview (CAPI) or a telephone interview (CATI). Cooperative households without Internet access were provided with equipment giving them access to the Internet. Analysis by Scherpenzeel and Bethlehem (2011) showed that this panel produced better estimates than panels based on self-selection.

A second approach to reduce the negative effects of self-selection is applying some form of adjustment weighting. Adjustment weighting is a family of techniques that attempts to reduce the bias of survey estimates by assigning weights to responding elements. These weights correct for the over- or underrepresentation of specific groups in the response. Adjustment weighting is treated in detail in Chapter 10. Here one such technique is summarized. It is called poststratification.

Poststratification requires one or more auxiliary variables. An auxiliary variable is a variable that has been measured in the survey, and for which the distribution in the target population is available. Typical target variables are gender, age, marital status, and region. By comparing the response distribution of an auxiliary variable with its population distribution, it can be assessed whether the survey response is representative for the population (with respect to this variable). If these distributions differ considerably, one must conclude that the response is selective. To correct this, adjustment weights are computed. Weights are assigned to all records of observed elements. Estimates of population characteristics now can be obtained by using the weighted values instead of the unweighted values.

To carry out poststratification, one or more qualitative auxiliary variables are needed. Here, only one such variable is considered. The situation for more variables is not essentially different. Suppose there is an auxiliary variable X having L categories. So it divides the target population into L strata. The strata are denoted by the subsets U_1, U_2, \dots, U_L of the population U . The number of

target population elements in stratum U_b is denoted by N_b , for $b = 1, 2, \dots, L$. The population size N is equal to $N = N_1 + N_2 + \dots + N_L$. This is the population information assumed to be available.

Suppose a sample of size n is selected from the Internet population. If n_b denotes the number of sample elements in stratum b , then $n = n_1 + n_2 + \dots + n_L$. Note that because the sample is selected from the Internet population U_I , only elements in the substrata $U_I \cap U_b$ are observed (for $b = 1, 2, \dots, L$).

Poststratification assigns identical adjustment weights to all elements in the same stratum. The weight w_k for an element k in stratum b is equal to

$$w_k = \frac{N_b/N}{n_b/n}. \quad (9.31)$$

Poststratification comes down to replacing the simple sample mean

$$\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^N R_k Y_k \quad (9.32)$$

with the weighted sample mean

$$\bar{y}_{S,PS} = \frac{1}{n_S} \sum_{k=1}^N w_k R_k Y_k. \quad (9.33)$$

Substituting the weights and working out this expression leads to the poststratification estimator

$$\bar{y}_{S,PS} = \frac{1}{N} \sum_{b=1}^L N_b \bar{y}_S^{(b)} = \sum_{b=1}^L W_b \bar{y}_S^{(b)}, \quad (9.34)$$

where $\bar{y}_S^{(b)}$ is the sample mean in stratum b and $W_b = N_b/N$ is the relative size of stratum b . The expected value of this poststratification estimator is equal to

$$E(\bar{y}_{S,PS}) = \frac{1}{N} \sum_{b=1}^L N_b E(\bar{y}_S^{(b)}) = \sum_{b=1}^L W_b \tilde{Y}^{(b)} = \tilde{Y}^*, \quad (9.35)$$

where

$$\tilde{Y}^* = \frac{1}{N_b} \sum_{k=1}^{N_b} \frac{\rho_{k,b}}{\bar{\rho}_b} Y_{k,b} \quad (9.36)$$

is the weighted mean of the target variable in stratum b . The subscript k, b denotes the k th element in stratum b , and $\bar{\rho}_b$ is the average response probability in stratum b .

Generally, this mean will not be equal to the mean \bar{Y}_b of the target variable in stratum b of the target population. The bias of this estimator is equal to

$$\begin{aligned} B(\bar{y}_{S,PS}) &= E(\bar{y}_{S,PS}) - \bar{Y} = \tilde{Y}^* - \bar{Y} = \sum_{b=1}^L W_b (\tilde{Y}^{(b)} - \bar{Y}^{(b)}) \\ &= \sum_{b=1}^L W_b \frac{R_{\rho Y}^{(b)} S_{\rho}^{(b)} S_Y^{(b)}}{\bar{\rho}^{(b)}}, \end{aligned} \quad (9.37)$$

where the subscript b indicates that the respective quantities are computed just for stratum b and not for the complete population. The bias (9.37) will be small if

- The response propensities are similar within strata.
- The values of the target variable are similar within strata.
- There is no correlation between the response behavior and the target variable within strata.

These conditions can be realized if there is a strong relationship between the target variable Y and the stratification variable X . Then the variation in the values of Y manifests itself between strata and not within strata. In other words, the strata are homogeneous with respect to the target variable. Also, if the strata are homogeneous with respect to the response propensities, the bias will be reduced. In terms of missing data terminology, this situation comes down to missing at random (MAR).

It can be shown that, in general, the variance of the poststratification estimator is approximately equal to

$$V(\bar{y}_{PS}) = \sum_{b=1}^L W_b^2 V(\bar{y}_b). \quad (9.38)$$

In the case of a self-selection web survey, the variance $V(\bar{y}_b)$ of the sample mean in a stratum is the analogue of variance (9.22) but restricted to observations in that stratum. Therefore, the variance of the poststratification estimator is approximately equal to

$$V(\bar{y}_{S,PS}) = \sum_{b=1}^L W_b^2 \frac{1}{(N_b \bar{\rho}_b)^2} \sum_{k \in U_b}^N \rho_k (1 - \rho_k) (Y_k - \tilde{Y}^{(b)})^2. \quad (9.39)$$

This variance is small if the strata are homogeneous with respect to the target variable. So, a strong correlation between the target variable Y and the stratification variable X will reduce both the bias and the variance of the estimator.

The conclusion can be that application of poststratification will successfully reduce the bias of the estimator if proper auxiliary variables can be found. Such variables should satisfy three conditions:

- They have to be measured in the survey.
- Their population distribution (N_1, N_2, \dots, N_L) must be known.
- They must produce homogeneous strata.

Unfortunately, such variables are rarely available, or there is only a weak correlation. One way to solve this problem is to carry out a *reference survey*. The objective of such a survey is to measure just the auxiliary variables required for weighting purposes. To obtain unbiased estimates of the population distributions of these variables, data preferably should be collected with CAPI or CATI. The reference survey is discussed in more detail in Chapter 10.

A third approach to reduce a self-selection bias is to apply *propensity weighting*. This technique particularly is used by several market research organizations. See also the studies by Börsch-Supan et al. (2004) and Duffy et al. (2005). The original idea behind propensity weighting goes back to Rosenbaum and Rubin (1983, 1984).

Propensity scores are obtained by modeling a variable that indicates whether someone participates in the survey. Usually a logistic regression model is used where the indicator variable is the dependent variable and the attitudinal variables are the explanatory variables. These attitudinal variables are assumed to explain why someone participates or not. Fitting the logistic regression model comes down to estimating the probability (propensity score) of participating, given the values of the explanatory variables.

Each person k in the population is assumed to have a certain, unknown probability ρ_k of participating in the survey, for $k = 1, 2, \dots, N$. Let R_1, R_2, \dots, R_N denote indicator variables, where $R_k = 1$ if person k participates in the survey, and $R_k = 0$ otherwise. Consequently, $P(R_k = 1) = \rho_k$.

The *propensity score* $\rho(X)$ is the conditional probability that a person with observed characteristics X participates; i.e.,

$$\rho(X) = P(R = 1|X). \quad (9.40)$$

It is assumed that within the strata defined by the values of the observed characteristics X , all persons have the same participation propensity. This is the MAR assumption. The propensity score is often modeled using a logit model:

$$\log\left(\frac{\rho(X_k)}{1 - \rho(X_k)}\right) = \alpha + \beta'X_k. \quad (9.41)$$

The model is fitted using maximum likelihood estimation. Once propensity scores have been estimated, they are used to stratify the population. Each stratum consists of elements with (approximately) the same propensity scores. If indeed all elements within a stratum have the same response propensity, there will be no bias if just the elements in the Internet population are used for estimation

purposes. Cochran (1968) claims that five strata are usually sufficient to remove a large part of the bias.

From a theoretical point of view, propensity weighting should be sufficient to remove the bias. However, in practice, the propensity score variable often will be combined with other (demographic) variables in a more extended weighting procedure; see, e.g., Schonlau et al. (2004). The use of propensity scores is described in more detail in Chapter 11.

9.3 Application

The possible consequences of self-selection and the effectiveness of correction techniques are illustrated using a simulation experiment. A fictitious population was constructed. For this population, reported voting behavior in an election survey was simulated and analyzed. The relationships between variables involved were modeled somewhat stronger than they probably would be in a real-life situation. Effects are therefore more pronounced, making it clearer what the pitfalls are.

The characteristics of estimators (before and after correction) were computed based on a large number of simulations. First, the distribution of the estimator was determined in the ideal situation of a simple random sample from the target population. Then, it was explored how the characteristics of the estimator changed if self-selection is applied. Finally, the effect of weighting (poststratification) was analyzed.

A fictitious population of 100,000 individuals was constructed. There were five variables:

- The variable *Internet* indicated how active a person was on the Internet. There were two categories: very active users and passive users. The population consisted of 1% of very active users and of 99% of passive users. Active users had a response propensity of 0.99, and passive users had a response propensity of 0.01.
- The variable *Age* in three categories: young, middle aged, and old. The active Internet users consisted of 60% of young people, 30% of middle-aged people, and 10% of old people. The age distribution for passive Internet users was 40% young, 35% middle aged, and 25% old. Typically younger people were more active Internet users.
- Voted for the National Elderly Party (NEP). The probability to vote for this party only depended on age. Probabilities were 0.00 (for Young), 0.30 (for Middle aged), and 0.60 (for Old).
- Voted for the New Internet Party (NIP). The probability to vote for this party depended on both age and use of the Internet. For active Internet users, the probabilities were 0.80 (for young), 0.40 (for middle aged), and 0.20 (for old). For passive Internet users, all probabilities were equal to 0.10. So, for active users, voting for the NIP decreased with age. Voting probability was always low for passive users.

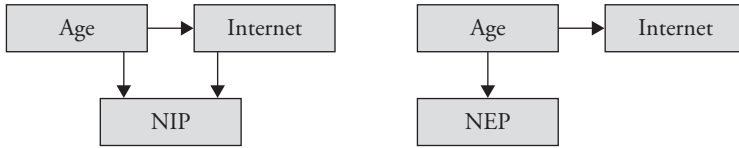


FIGURE 9.2 Relationships among variables

Figure 9.2 shows the relationships among the variables in a graphical way. The decision not to participate in a self-selection survey can be considered a form of nonresponse. Nonresponse theory distinguishes three nonresponse generating mechanisms:

- *Missing Completely At Random* (MCAR). There is no relationship at all between the mechanism causing data to be missing and the target variables of the survey. This situation causes no problems. The mechanism only leads to a reduced number of observations. Estimators will not be biased.
- *Missing At Random* (MAR). There is an indirect relationship between the mechanism causing data to be missing and the target variables of the survey. The relationship runs through a third variable, and this variable is measured in the survey as an auxiliary variable. Estimates are biased in this case, but it is possible to correct for this bias. For example, if the auxiliary variable is used to construct strata, there will be no bias within strata, and the poststratification will remove the bias.
- *Not Missing At Random* (NMAR). There is a direct relationship between the mechanism causing data to be missing and the target variables of the survey. This is the worst case. Estimators will be biased, and it is not possible to remove this bias.

The variable NEP (National Elderly Party) suffers from MAR. There is a direct relationship between voting for this party and age, and there is a direct relationship between age and the propensity to participate in the survey. This will cause estimates to be biased. It should be possible to correct for this bias by weighting using the variable age.

The variable NIP (National Internet Party) suffers from NMAR. There is a direct relationship between voting for this party and the response propensity. Estimates will be biased, and there is no correction possible.

The distribution of estimators for the percentage of votes for both parties was determined in various situations by repeating the selection of the sample 1,000 times. The average response propensity in the population is 0.01971. Therefore, the expected sample size in a self-selection survey was equal to $100,000 \times 0.01971 = 1,971$.

Figure 9.3 contains the results for the variable NEP (voted for National Elderly Party). The upper box plot shows the distribution of the estimator for simple random samples of size $n = 1,971$ from the target population. The

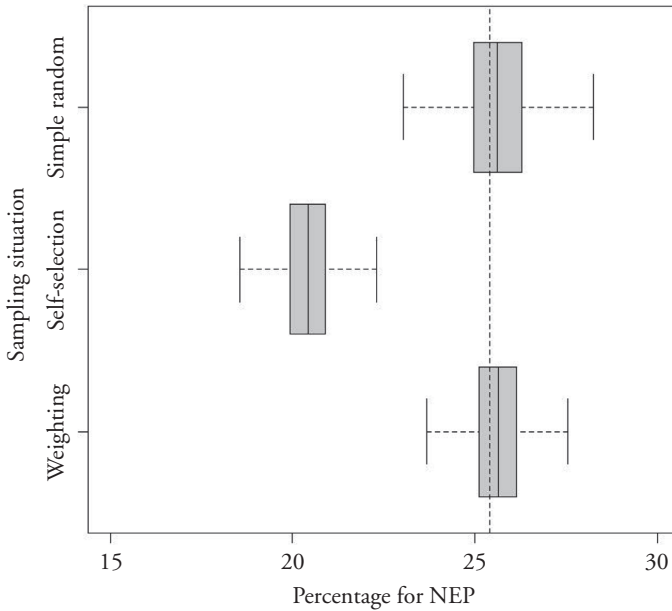


FIGURE 9.3 Results of the simulations for variable NEP (National Elderly Party)

vertical line denotes the population value to be estimated (25.6%). The estimator has a symmetric distribution around this value. This is a clear indication that the estimator is unbiased.

The middle box plot shows what happens if samples are selected by means of self-selection. The shape of the distribution remains more or less the same, but the distribution as a whole has shifted to the left. All values of the estimator are systematically too low. The expected value of the estimator is only 20.4%. The estimator is biased. The explanation of this bias is simple: Relative few elderly are active Internet users. Therefore, they are underrepresented in the samples. These are typically people who will vote for the NEP.

The lower box plot shows the distribution of the estimator in case of poststratification by age. The bias is removed. This was possible because this is a case of MAR.

The simulation results are summarized in Table 9.2. Self-selection results in a large relative bias of -5.2 . Weighting eliminates the bias, and therefore, the

TABLE 9.2 Summary of simulation results for the variable NEP

Simulation	Mean	Standard error	Bias	Relative bias
Random	25.6	1.0	0.0	0.0
Self-selection	20.4	0.7	-5.2	-7.4
Self-selection, with weighting adjustment	25.6	0.7	0.0	0.0

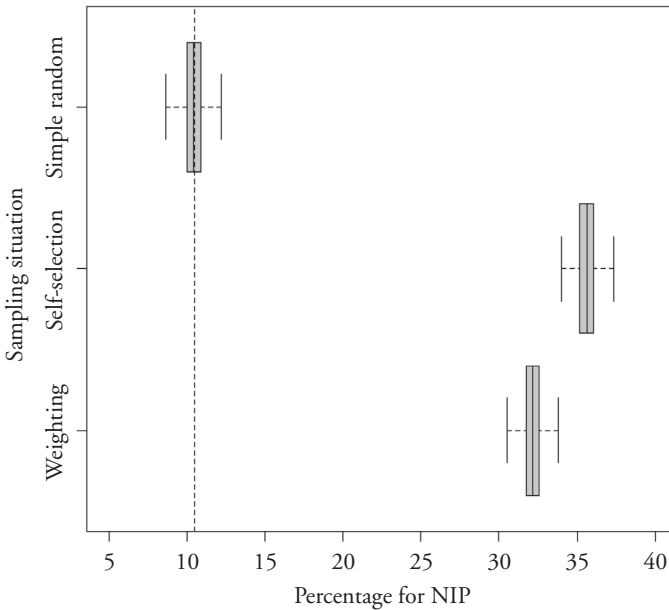


FIGURE 9.4 Results of the simulations for variable NIP (*New Internet Party*)

relative bias is 0. In the case of self-selection, the confidence level of the 95% confidence interval is close to 0. This implies that almost certainly a wrong conclusion will be drawn from this web survey.

Figure 9.4 contains the results for the variable NIP (voted for the New Internet Party). The upper box plot shows the distribution of the estimator for simple random samples of size 1,971 from the target population. The vertical line denotes the population value to be estimated (10.4%). As the estimator has a symmetric distribution around this value, it is clear that the estimator is unbiased.

The middle box plot shows what happens if samples are obtained by means of self-selection. The distribution has shifted to the right considerably. All values of the estimator are systematically too high. The expected value of the estimator is now 35.65%. The estimator is severely biased. The explanation of this bias is straightforward: Voters for the NIP are overrepresented in Internet samples.

The lower box plot in Figure 9.4 shows the effect of poststratification by age. Only a small part of the bias is removed. This is not surprising as there is a direct relationship between voting for the NIP and having access to the Internet. This is a case of NMAR.

The simulation results are summarized in Table 9.3. Self-selection results in a very large relative bias of 42.0. Weighting can reduce this relative bias only to 36.2. In both cases, the confidence level of the 95% confidence interval is close to 0. This implies that almost certainly wrong conclusions will be drawn from this web survey.

TABLE 9.3 Summary of simulation results for the variable NIP

Simulation	Mean	Standard error	Bias	Relative bias
Random samples	10.4	0.7	0.0	0.0
Self-selection	35.6	0.6	25.2	42.0
Self-selection, with weighting adjustment	32.1	0.6	21.7	36.2

9.4 Summary

A web survey may suffer from self-selection. The survey questionnaire is simply put on the web. Respondents are those people who happen to have Internet access, visit, the website, and decide to participate in the survey. The survey researcher is not in control over the selection process.

Because selection probabilities are unknown, it is not possible to compute unbiased estimates. It is also impossible to compute the precision of estimates.

It is usually assumed that a self-selection sample can be treated as a simple random sample. Hence, sample means are used as estimates of population means. Such an estimator can be seriously biased. The magnitude of the bias depends on:

1. The average response probability. If people are more likely to participate in the survey, the average response probability will be higher, causing the bias to be smaller.
2. The relationship between the target variable and the response behavior. A strong correlation between the values of the target variable and the response probabilities, will lead to a large bias.
3. The variation in the response probabilities. The more these values vary, the larger the bias will be.

The bias of the estimator is independent of the sample size. The bias does not go away for large samples. Particularly for large samples, confidence intervals give a wrong picture. The confidence level of this interval is often closer to 0% than to 95%

Several approaches may help to reduce a bias from self-selection. A first approach is to select a proper probability sample from a sampling frame. A second approach is to carry out some kind of adjustment weighting. By assigning weights, the response is corrected for under- or overrepresented groups. There is no guarantee that weighting will completely remove the bias.

KEY TERMS

First-order inclusion probability: The probability that a population element is selected in the sample. The first-order inclusion probability is determined by the sampling design.

Internet population: The subpopulation of the target population consisting of elements that have access to the Internet.

Missing at Random (MAR): Nonresponse depends on auxiliary variables only. Estimators will be biased, but correction is possible if some technique is used that takes advantage of this auxiliary information.

Missing Completely at Random (MCAR): Nonresponse happens completely independent of all survey variables. The estimators will not be biased.

Mixed-mode survey: A survey in which various modes of data collection are combined. Modes can be used concurrently (different groups are approached by different modes) or sequentially (nonrespondents of a mode are reapproached in a different mode).

Non-Internet population: The subpopulation of the target population consisting of elements that do not have access to the Internet.

Not Missing at Random (NMAR): Nonresponse depends directly on the target variables of the survey. The estimators will not be biased, and the correction techniques will not be successful.

Poststratification: A weighting adjustment technique that divides the population in strata and subsequently assigns the same weight to all observed elements within a stratum.

Self-selection survey: A survey for which the sample has been recruited by means of self-selection. It is left to the persons themselves to decide to participate in a survey. The selection probabilities are unknown.

EXERCISES

Exercise 9.1. Which of the following reasons may cause a researcher to use self-selection for a web survey instead of a probability sample?

- a. A self-selection survey is cheaper.
- b. The sample will be larger.
- c. There is no nonresponse.
- d. No sampling frame is needed.

Exercise 9.2. Can the bias from self-selection always be corrected by means of a weighting adjustment?

- a. Yes, weighting adjustment will always remove the bias.
- b. No, weighting adjustment will only work if persons are missing as a result of MAR.
- c. No, weighting adjustment will only work if persons are missing as a result of MAR, and the proper auxiliary variables are included in the weighting model.
- d. No, weighting adjustment will never remove or reduce such a bias.

Exercise 9.3. What happens to the bias from self-selection in web surveys if Internet access increases in the target population?

- The bias will increase.
- The bias will decrease.
- The bias will not change.
- It depends on the response probabilities of the people getting Internet access.

Exercise 9.4. Suppose all response probabilities in a specific population are less than 0.5. What would happen to the bias from self-selection if all response probabilities were twice as large?

- The bias will be twice as large.
- The bias will be halved.
- The bias will not change.
- The bias will vanish.

Exercise 9.5. The target variable Y of a survey indicates whether persons have ($Y = 1$) or have not ($Y = 0$) a specific property. There is also an auxiliary variable X with two categories ($X = 1$ and $X = 2$). The population consists of 2,400 people. The distribution of these people in the table obtained by crossing X and Y is given below. Within each cell of the table, the response probabilities are the same. The values of these probabilities also are given in the table below.

	$Y = 0$	$Y = 1$
$X = 1$	$N = 1,000$ $\rho = 0.4$	$N = 200$ $\rho = 0.4$
$X = 2$	$N = 200$ $\rho = 0.8$	$N = 1,000$ $\rho = 0.8$

- Suppose a simple random sample is selected. Assume that all selected persons will respond. What will be expected value of the percentage of people having the specific property?
- Suppose sampling relies on self-selection and the response probabilities are as indicated in the table. What will be the expected value of the estimated percentage?
- Suppose sampling relies on self-selection and the response probabilities are as indicated in the table. If the variable X is used for adjustment weighting (poststratification), what will be the expected value of the estimated percentage?
- Explain why, or why not, the self-selection bias is removed by weighting by X .

Exercise 9.6. A researcher wants to investigate the transport behavior of commuters in a town. He intends to conduct a survey in which he asks commuters what means of transport they use to go to their work. The researcher thinks it is a good idea to have information about many commuters. He does not have the money and time to draw a proper random sample. He decides to go to the main railway station during rush hour. He succeeds in interviewing a lot of people in the station. Not surprisingly, it turns out that many people use the train to commute.

a. Explain what is wrong with this sampling design?

The researcher observes that he has many young people in his survey and only a very few elderly. He decides to carry out a weighting adjustment by post-stratifying by the variable age.

b. Explain why or why not this will help to improve his estimators.

REFERENCES

- Bethlehem, J. G. (1988), Reduction of the Nonresponse Bias Through Regression Estimation. *Journal of Official Statistics*, 4, pp. 251–260.
- Bethlehem, J. G. (2002), Weighting Nonresponse Adjustments Based on Auxiliary Information. In: Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (eds.), *Survey Nonresponse*. John Wiley & Sons, New York.
- Börsch-Supan, A., Elsner, D., Faßbender, H., Kiefer, R., McFadden, D., & Winter, J. (2004), *Correcting the Participation Bias in an Online Survey*. Report, University of Munich, Germany.
- Cochran, W. G. (1997), *Sampling Techniques*, 3rd ed. John Wiley & Sons, New York.
- Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005), Comparing Data from Online and Face-to-face Surveys. *International Journal of Market Research*, 47, pp. 615–639.
- Horvitz, D. G. & Thompson, D. J. (1952), A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, pp. 663–685.
- Kruskal, W. & Mosteller, F. (1979a), Representative Sampling, I: Non-scientific Literature, *International Statistical Review*, 47, pp. 13–24.
- Kruskal, W. & Mosteller, F. (1979b), Representative Sampling, II: Scientific Literature, Excluding Statistics, *International Statistical Review*, 47, pp. 111–127.
- Kruskal, W. & Mosteller, F. (1979c), Representative Sampling, III: The Current Statistical Literature, *International Statistical Review*, 47, pp. 245–265.
- Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 3rd ed. John Wiley & Sons, New York.
- Rosenbaum, P. R. & Rubin, D. B. (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, pp. 41–55.
- Rosenbaum, P. R. & Rubin, D. B. (1984), Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79, pp. 516–524.

- Scherpenzeel, A. (2008), An Online Panel as a Platform for Multi-disciplinary Research. In: Stoop, I. & Wittenberg, M. (eds.), *Access Panels and Online Research, Panacea or Pitfall?* Aksant, Amsterdam, pp. 101–106.
- Scherpenzeel, A. & Bethlehem, J. (2011), How Representative are Online-panels? Problems of Coverage and Selection and Possible Solutions. In: Das, M., Ester, P., Kaczmirek, L. & Mohler, P. (eds.), *Social Research and the Internet: Advances in applied Methods and New Research Strategies*. Routledge Academic, New York, pp. 105–132.
- Schonlau, M., Zapert, K., Payne Simon, L., Haynes Sanstad, K., Marcus, S., Adams, J. Kan, H., Turber, R., & Berry, S. (2004), A Comparison Between Responses from Propensity-weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review*, 22, pp. 128–138.
- Vonk, T., Van Ossenbruggen, R., & Willems, P. (2006), The Effects of Panel Recruitment and Management on Research Results, A Study Among 19 Online Panels. *Panel Research 2006, ESOMAR World Research*, ESOMAR Publication Services, Vol. 317, pp. 79–99.

Weighting Adjustment Techniques

10.1 Introduction

It is the basic idea of survey sampling that observations on only a part of the elements in a population allow for drawing a valid and accurate conclusion about the population as a whole. Horvitz and Thompson (1952) have shown in their seminal paper that this possible provided a probability sample has been selected and that each element in the population has a positive probability of selection in the sample. If these conditions are satisfied, unbiased estimates of population characteristics can be computed. Moreover, the accuracy of these estimates can be computed. Chapter 3 describes this in more detail.

Let $U = \{1, 2, \dots, N\}$ denote the population to be surveyed. Let Y denote a target variable of the survey. The value of Y for element k is denoted by Y_k , for $k = 1, 2, \dots, N$. Let the aim of the survey be estimation of the population mean

$$(10.1) \quad \bar{Y} = \frac{1}{N} \sum_{k=1}^N Y_k.$$

A sample design is chosen to select a sample from this population. Only sampling designs are considered here that draw a sample without replacement. This implies a sample can be represented by a series of indicators

$$(10.2) \quad a_1, a_2, \dots, a_N,$$

where the indicator assumes the value 1 if element k is selected in the sample, and otherwise it assumes the value 0. The expected value of a_k is denoted by

$$(10.3) \quad \pi_k = E(a_k).$$

This quantity π_k is called the *first-order inclusion probability* of element k . It is equal to the probability that this element is selected in the sample. The second-order inclusion probability of two elements k and l (with $k \neq l$) is defined as

$$(10.4) \quad \pi_{kl} = E(a_k a_l).$$

It is the probability that elements k and l are selected together in the sample.

The most common sampling design is a *simple random sample* (without replacement). This sampling design assigns the same probability of selection to each element in the population. This implies that $\pi_k = n/N$ for all k . Furthermore, all second-order inclusion probabilities are equal to $\pi_{kl} = n(n-1)/N(N-1)$.

Horvitz and Thompson (1952) show that an unbiased estimator always can be constructed. Their estimator can be written as

$$(10.5) \quad \bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^N a_k \frac{Y_k}{\pi_k}.$$

The Horvitz–Thompson estimator (10.5) is an unbiased estimator of the population mean \bar{Y} provided that $\pi_k > 0$ for all k . The variance of this estimator is equal to

$$(10.6) \quad V(\bar{y}_{HT}) = \frac{1}{2N^2} \sum_{k=1}^N \sum_{l=1}^N (\pi_k \pi_l - \pi_{kl}) \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l} \right)^2.$$

A closer look at estimator (10.5) makes it clear that proper estimation requires the sample elements to be weighted. Elements with a larger inclusion probability will be overrepresented in the sample. This is corrected in the estimator by dividing by the inclusion probability.

If the *design weight* d_k for element k is defined as $d_k = 1/\pi_k$ (for $k = 1, 2, \dots, N$), then the Horvitz–Thompson estimator can be written as

$$(10.7) \quad \bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^N a_k d_k Y_k.$$

If a sample of size n is selected, the values of Y for the selected elements are denoted by y_1, y_2, \dots, y_n , and d_1, d_2, \dots, d_n are the corresponding design weights, then Horvitz–Thompson estimator also can be written as

$$(10.8) \quad \bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^n d_i y_i.$$

In the case of simple random sampling, all design weights are equal to N/n . Consequently, the Horvitz–Thompson estimator reduces to the simple sample mean

$$(10.9) \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

EXAMPLE 10.1 A web survey based on an address sample

Suppose a web survey is conducted among all adult inhabitants of a town. To select a random sample of persons, addresses are randomly selected from a list of all addresses in a town. One person is randomly drawn at each selected address by determining the adult who is the first to have his/her birthday. This person is invited to complete the on-line questionnaire. This is not an equal probability sample, and therefore, design weights have to be included in the estimator.

Let the size of the adult population be denoted by N . Suppose this population is distributed over M addresses, where the number of adults at address b is denoted by N_b , for $b = 1, 2, \dots, M$. The consequence of this sampling design is that the inclusion probability for an element k at address b is equal to

$$\pi_k = \frac{n}{M} \frac{1}{N_b}.$$

Let the indicators a_1, a_2, \dots, a_M denote the selected addresses. Furthermore, let the indicators b_{bk} denote which persons are selected at address b (for $b = 1, 2, \dots, M$ and $k = 1, 2, \dots, N_b$). By substituting all these quantities in expression (10.5), the Horvitz–Thompson estimator becomes

$$\bar{y}_{HT} = \frac{M}{N} \frac{1}{n} \sum_{b=1}^M a_b N_b \sum_{k=1}^{N_b} b_{bk} Y_{bk}.$$

If the measured value of Y for selected address i is denoted by y_i , the estimator can be rewritten as

$$\bar{y}_{HT} = \frac{M}{N} \frac{1}{n} \sum_{i=1}^n N_i y_i.$$

It is clear that this expression is not equal to the sample mean of the y_i .

It is convenient to assume that a web survey sample is a random sample that has been selected with equal probabilities. Under this assumption, the sample mean is an unbiased estimator of the population mean and a population percentage is an unbiased estimator of a population percentage.

However, such an assumption can lead to a serious bias. Suppose the sample is in fact selected with first-order inclusion probabilities $\pi_1, \pi_2, \dots, \pi_N$. Then the expected value of the sample mean is equal to

$$(10.10) \quad E(\bar{y}) = \frac{1}{n} \sum_{k=1}^N E(a_k) Y_k = \frac{1}{n} \sum_{k=1}^N \pi_k Y_k.$$

The bias of this estimator turns out to be equal to

$$(10.11) \quad B(\bar{y}) = E(\bar{y}) - \bar{Y} = \frac{1}{n} \sum_{k=1}^N (\pi_k - \bar{\pi})(Y_k - \bar{Y}) = \frac{N}{n} C_{\pi Y},$$

where $C_{\pi Y}$ is the covariance between the inclusion probabilities. The stronger the (linear) relationship between inclusion probabilities and the target variable, the larger the bias will be.

This theory shows that if a sample is not selected with equal probabilities, weighting is always required to obtain unbiased estimates of population characteristics. The design weights have to be computed for this.

A second type of weighting may be applied to improve the precision of estimators. The aim is not reducing or removing a bias but reducing the variance of the estimator. An additional advantage of such weighting techniques is that the weighted sample becomes representative with respect to some auxiliary variables. The weighting techniques described in Sections 10.2 (poststratification), 10.3 (the generalized regression estimator), and 10.4 (raking ratio estimation) all can do this if the proper auxiliary information is available.

A third type of weighting often is used to correct for bias caused by non-response. According to the *Random Response Model* (see, e.g., Bethlehem, 2009), each element k in the population has an (unknown) response probability ρ_k . If element k is selected in the sample, a random mechanism is activated that results with probability ρ_k in response and with probability $1 - \rho_k$ in nonresponse. If nonresponse occurs in a simple random sample, the response mean \bar{y}_R is not unbiased any more. Bethlehem (2009) shows that the bias is equal to

$$(10.12) \quad B(\bar{y}_R) = \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}},$$

where $R_{\rho Y}$ is the correlation between the values of the target variable and the response probabilities, S_{ρ} is the standard deviation of the response probabilities, S_Y

is the standard deviation of the variable Y , and $\bar{\rho}$ is the population mean of the response probabilities. The bias will be large if

- The relationship between the target variable and the response behavior is strong
- The variation in the response probabilities is large
- The average response probability is low

The weighting techniques described in the following sections can reduce the nonresponse bias provided proper auxiliary information is available.

The three reasons for weighting described above apply to any survey, whatever the mode of data collection. There are two more reasons for weighting that are particularly important for many web surveys. These reasons are undercoverage and self-selecting.

Undercoverage problems are described in detail in Chapter 8. *Undercoverage* occurs if the target population contains elements that do not have a counterpart in the sampling frame. Such elements can never be selected in the sample. Undercoverage occurs in a web survey if the target population is wider than just persons with Internet access and respondents are selected via the Internet. There is undercoverage because it is impossible for people without Internet to participate in the survey. This type of undercoverage can have serious consequences. If people without Internet access systematically differ from persons with access to the Internet, the estimates of population parameters may be seriously biased. It is shown in Chapter 8 that the bias of the sample mean \bar{y}_I is equal to

$$(10.13) \quad B(\bar{y}_I) = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N} (\bar{Y}_I - \bar{Y}_{NI}).$$

The magnitude of this bias is determined by two factors. The first factor is the relative size N_{NI}/N of the non-Internet population. The bias will increase as a larger proportion of the population does not have access to the Internet. The second factor is the *contrast* $\bar{Y}_I - \bar{Y}_{NI}$ between the Internet population and the non-Internet population. It is the difference between the population means of the two subpopulations. The more the mean of the target variable differs for these two subpopulations, the larger the bias will be.

The weighting techniques in the following sections can be attempted to reduce the bias resulting from undercoverage. There is no guarantee that this will be successful, as will be shown in the subsequent section.

Self-selection problems are described in detail in Chapter 9. *Self-selection* is the phenomenon that the sample is not selected by means of a probability sample. Instead, it is left to the Internet users themselves to participate in a web survey. The survey questionnaire is simply put on the web. Respondents are those people who happen to have Internet access, visit the website, and decide to participate in the survey. The survey researcher is not in control of the selection process.

Participation in a self-selection requires that respondents are aware of the existence of a survey. They have to visit the website accidentally, or they have to follow up a banner or an e-mail message. They also have to decide to fill in the questionnaire on the Internet. This means that each element k in the Internet population has unknown probability ρ_k of participating in the survey, for $k = 1, 2, \dots, N_I$, where N_I is the size of the population of persons having access to the Internet.

Assuming it is the objective of the web survey to estimate the mean \bar{Y}_I of the Internet population, Bethlehem (2002) shows that the bias of the sample mean \bar{y}_S can be written as

$$(10.14) \quad B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y}_I = \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}},$$

where $R_{\rho Y}$ is the correlation coefficient between the target variable and the response behavior, S_ρ is the standard deviation of the response probabilities, and S_Y is the standard deviation of the target variable. This bias is large if

- The relationship between the target variable and the participation probabilities is strong
- The variation in the participation probabilities is large
- The average participation probability is low

The weighting techniques in the following sections also can be attempted to reduce the bias resulting from self-selection. Again, there is no guarantee that this will be successful, as will be shown in this chapter.

Three types of weighting adjustment techniques will be described: post-stratification, generalized regression estimation, and raking ratio estimation. Note that Chapter 11 is devoted to the use of so-called propensity scores. These propensity scores also can be used for weighting.

It will be made clear in this chapter that an effective weighting adjustment procedure requires proper auxiliary information.

10.2 Theory

10.2.1 THE CONCEPT OF REPRESENTATIVITY

The principles of weighting adjustment are closely related to the concept of *representativity*. This concept often is used in survey research, but usually it is not clear what it means. Kruskal and Mosteller (1979a, 1979b, 1979c) present an extensive overview of what *representative* is supposed to mean in nonscientific literature, scientific literature excluding statistics, and the statistical literature. They found the following meanings for “representative sampling”:

- General acclaim for data
- Absence of selective forces
- Miniature of the population
- Typical or ideal case(s)
- Coverage of the population
- A vague term, to be made precise
- Representative sampling as a specific sampling method
- As permitting good estimation
- Good enough for a particular purpose

To avoid confusion, Kruskal and Mosteller recommended not using the word *representative* but instead to specify what one means. In this chapter, the concept of representativity with respect to a variable is used. A survey data set is defined to be *representative with respect to a variable X* if the distribution of X in the data set is equal to the distribution of this variable in the population. For example, if a sample is representative with respect to a variable, then the sample mean \bar{x} is equal to the population mean \bar{X} .

Weighting adjustment is based on the use of *auxiliary information*. Auxiliary information is defined here as a set of variables that have been measured in the survey, and for which the distribution in the population is available. By comparing the population distribution of an auxiliary variable with its response distribution, it can be assessed whether or not the response is representative for the population (with respect to this variable). If the distributions differ considerably, one must conclude that the survey response is not representative.

The next step is to use the auxiliary information to compute *adjustment weights*. Weights are assigned to all observed elements. Estimates of population characteristics can now be obtained by using weighted values instead of the unweighted values. The weights are defined in such a way that population characteristics for the auxiliary variables can be computed without error. So the weighted sample is forced to be *representative with respect to the auxiliary variables* used.

Recall that, whatever sampling design is used, always an unbiased estimator can be constructed. This is the Horvitz–Thompson estimator. It can be written as

$$(10.15) \quad \bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^n d_i y_i,$$

where $d_i = 1 / \pi_i$ is the *design weight* and y_i is observed value of sample element i , for $i = 1, 2, \dots, n$.

Adjustment weighting replaces this estimator by a new estimator

$$(10.16) \quad \bar{y}_W = \frac{1}{N} \sum_{i=1}^n w_i y_i,$$

where the weight w_i is equal to

$$(10.17) \quad w_i = c_i \times d_i$$

and c_i is a *correction weight* produced by a weighting adjustment technique.

Weighting adjustment techniques impose the condition of representativity with respect to one or more selected auxiliary variables. Suppose X is such an auxiliary variable. Representativity with respect to X implies that the weights w_i have to be such that

$$(10.18) \quad \frac{1}{N} \sum_{i=1}^n w_i x_i = \bar{X}.$$

This means that if the weights are used to estimate the population mean of the auxiliary variable, the estimate is exactly equal to the population mean.

If the response can be made representative with respect to several auxiliary variables, and if all these variables have a strong relationship with the phenomena to be investigated, then the weighted sample also will be (approximately) representative with respect to these phenomena, and hence, estimates of population characteristics will be more accurate.

Several weighting techniques will be described in this section. It starts with the simplest and most commonly used one: *poststratification*. Next, *generalized regression estimation* is described, which is more general than poststratification. This technique can be applied in situations where the auxiliary information is inadequate for poststratification. Furthermore, *raking ratio estimation* is discussed as an alternative for generalized regression estimation. Also, an introduction into *calibration* is given. This can be seen as an even more general-theoretical framework for adjustment weighting that includes generalized regression estimation and raking ratio estimation as special cases.

10.2.2 POSTSTRATIFICATION

The concept of *stratification* has a long history in survey methodology. Stratification means that the target population of the survey is divided into several groups. These groups are called *strata*. A sample is selected from each group so that estimates can be computed for each group separately. The next step is to combine the group estimates into an estimate for the whole population.

Stratification played already a role in the first ideas that emerged about sampling. It was Anders Kiaer, director of the Norwegian statistical institute, who proposed at a meeting of the ISI (International Statistical Institute) in Bern in 1895 to use sampling instead of complete enumeration. He argued that good results could be obtained with his *Representative Method*. His idea was to select a sample that should reflect all aspects of the population as much as possible. One way to realize such a sample was the “balanced sample”. He divided the population into groups using variables like gender, age, and region. The sizes of the groups were

supposed to be known. The same percentage of persons was taken in each group. Selection of samples took place in some haphazard way (probability sampling had not yet been invented). As a result, the sample distribution of variables like gender, age, and region was similar to the distribution in the population. Hence, the sample was representative with respect to these variables.

Holt and Smith (1979) noted the wide use of stratification in the 1970s as it has two attractive properties: (1) it leads to representative samples and (2) it improves the precision of estimators. There is, however, also a drawback. To be able to draw a stratified sample, a sampling frame is required for each group separately. This is not always the case. There are many situations in which membership of a group can be established only after inspection of the sampled data. For example, to obtain a sample that is representative with respect to age groups, a sample would have to be drawn from each age group separately. Usually, there is no sampling frame for each age group.

Poststratification is an estimation technique that attempts to make the sample representative after the data has been collected. It comes down to assigning stratum weights. Respondents in underrepresented groups get a weight larger than 1, and respondents in overrepresented groups get a weight smaller than 1. By using weighted values as in expression (10.16), properties of estimators will be improved.

Strata can be obtained by using a single auxiliary variable or by crossing several auxiliary variables. Poststratification is particularly effective if the strata are *homogeneous*. This means the people within strata resemble each other. If this is the case, poststratification will not only improve the precision (as measured by the variance or the standard error of estimators) but also reduce a possible bias.

First, the theory of poststratification is described in the ideal case of a simple random sample without nonresponse, undercoverage, and self-selection problems. Then it is explored if and when poststratification can reduce a bias caused by these problems.

To be able to carry out poststratification, one or more qualitative auxiliary variables are needed. The theory is described for one such variable, but the case of more variables is not essentially different. Suppose there is an auxiliary variable X having L categories. So it divides the population U into L strata U_1, U_2, \dots, U_L . The number of population elements in stratum U_b is denoted by N_b , for $b = 1, 2, \dots, L$. So $N = N_1 + N_2 + \dots + N_L$. These stratum sizes are supposed to be known. This is the population distribution of the variable X .

Assume a simple random sample of size n is selected without replacement from the population. If n_b denotes the number of sample elements in stratum U_b (for $b = 1, 2, \dots, L$), then $n = n_1 + n_2 + \dots + n_L$. Note that the values of the n_b are the result of a random selection process. So, they are random variables.

To get a sample that is representative with respect to the variable X , the proportion of elements in stratum b should be equal to N_b / N , for $b = 1, 2, \dots, L$. However, the proportion of sample elements in stratum b is equal to n_b / n . To correct for this, each observed element i in stratum U_b is assigned a correction weight equal to

$$(10.19) \quad c_i = \frac{N_b/N}{n_b/n}.$$

If the values of the inclusion weights $d_i = n / N$ and correction weights (10.19) are substituted in expression (10.16), the result is the *poststratification estimator*

$$(10.20) \quad \bar{y}_{PS} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}^{(h)},$$

where $\bar{y}^{(h)}$ is the mean of the observed elements in stratum h . So, the poststratification estimator is equal to a weighted sum of sample stratum means.

EXAMPLE 10.2 Computing weights by means of poststratification

To show the effects of poststratification, a fictitious population is used. It is the population described in Section 8.3 in Chapter 8. The population consists of all eligible voters of a town. The size of the target population is 30,000 persons.

An election survey is conducted among these voters. A simple random sample of size 1,000 is drawn from this population. Age (in three classes) is used as an auxiliary variable. Table 10.1 contains the population and sample frequencies of this variable.

TABLE 10.1 Computing poststratification weights

Age	Population		Sample		Weight
	Frequency	Percentage	Frequency	Percentage	
Young	11,949	39.8300	402	40.2000	0.990796
Middle	10,582	35.2733	342	34.2000	1.031384
Elderly	7,469	24.8967	256	25.6000	0.972526
Total	30,000	100.0000	1,000	100.0000	

Young people are slightly overrepresented. The sample percentage (40.2%) is larger than the population percentage (39.8%). Therefore, poststratification assigns a weight smaller than 1. The weight for this group is obtained by dividing the population percentage by the sample percentage. The result is 0.990796.

Likewise, middle-aged persons are underrepresented. The sample percentage (34.2%) is smaller than the population percentage (35.3%). Therefore the weight is larger than 1. It is obtained by dividing 35.2733 by 34.2 resulting in a weight of 1.031383.

The adjustment weights w_i are obtained by multiplying the correction weights c_i by the inclusion weights d_i . Here all inclusion weights are equal to $N / n = 30$. Suppose the weights are used to estimate the number of old

persons in the population. The weighted estimate would be $0.972526 \times 30 \times 256 = 7,469$, and this is exactly the population frequency. Thus, application of weights to the auxiliary variables results in perfect estimates. If there is a strong relationship between the auxiliary variable and the target variable, also estimates for the target variable will be improved if these weights are used.

There is no simple exact analytical expression for the variance of poststratification estimator as defined by (10.20). There is, however, a large sample approximation:

$$(10.21) \quad V(\bar{y}_{PS}) = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{b=1}^L W_b S_b^2 + \frac{1}{n^2} \sum_{b=1}^L (1 - W_b) S_b^2,$$

where $W_b = N_b / N$ is the relative size of stratum b and S_b^2 is the (adjusted) population variance of the target variable in stratum b . The poststratification estimator is precise if the strata are homogeneous with respect to the target variable. This implies that variation in the values of the target variable is typically caused by differences in means between strata and not by variation within strata.

EXAMPLE 10.3 The variance of poststratification estimator

Suppose one of the aims of the election survey in example 10.1 is to estimate the percentage of people voting for the National Elderly Party (NEP). To that end, a simple random sample of size 1,000 is drawn from the population of size 30,000.

The variance of the sample percentage turns out to be equal to 1.832. If poststratification is carried out with age (in three categories) as the auxiliary variable, the variance of the estimator is equal to 1.294. So poststratification reduces the variance of the estimator. Apparently the age strata are more homogeneous with respect to voting behavior than the population as a whole.

The *effective sample size* n_{eff} is sometimes used as an indicator of how effective a sampling design or estimation procedure is. It is the sample size needed to obtain the same level of precision with the sample mean in simple random sampling. In the case of poststratification, it is defined as

$$n_{eff} = n \frac{V(\bar{y})}{V(\bar{y}_{PS})}.$$

The effective sample size for the election survey is $1,000 \times 1.832 / 1.294 = 1,416$. So, the sample mean requires 416 more sample elements to obtain the same precision as the poststratification estimator.

Now suppose the sample is affected by *nonresponse*. Then the poststratification estimator takes the form

$$(10.22) \quad \bar{y}_{R,PS} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_R^{(h)},$$

where $\bar{y}_R^{(h)}$ denotes the mean of the responding elements in stratum h . The bias of this estimator is equal to

$$(10.23) \quad B(\bar{y}_{R,PS}) = \frac{1}{N} \sum_{h=1}^L N_h \frac{R_{\rho Y}^{(h)} S_{\rho}^{(h)} S_Y^{(h)}}{\bar{\rho}^{(h)}},$$

where $R_{\rho Y}^{(h)}$ is the correlation between Y and ρ in stratum h ; $S_{\rho}^{(h)}$ and $S_Y^{(h)}$ are the standard errors of ρ and Y in stratum h , respectively; and $\bar{\rho}^{(h)}$ is the mean of the response probabilities in stratum h . The bias of the poststratification estimator is small if the biases within strata are small. A stratum bias is small in the following situations:

- There is little or no relationship between the target variable and the response behavior within the stratum. Then their correlation is small.
- All response probabilities within a stratum are more or less equal. Then their standard error is small.
- All values of the target variable within a stratum are more or less equal. Then their standard error is small.

These conclusions give some guidance with respect to the construction of strata. Preferably, strata should be used that are homogeneous with respect to the target variable, response probabilities, or both. The more elements resemble each other within strata, the smaller the bias will be.

EXAMPLE 10.4 Using poststratification for reducing nonresponse bias

Suppose one of the aims of the election survey in example 10.1 is to estimate the percentage of people voting for the National Elderly Party (NEP). To that end, a simple random sample of size 1,000 is drawn from the population of size 30,000. Three situations are compared. The first one is the ideal situation of a simple random sample from the target population with full response. The selection of the sample is repeated 1,000 times. For each sample the percentage of voters for the NEP is computed. The distribution of these estimates is represented in the upper box plot in Figure 10.1.

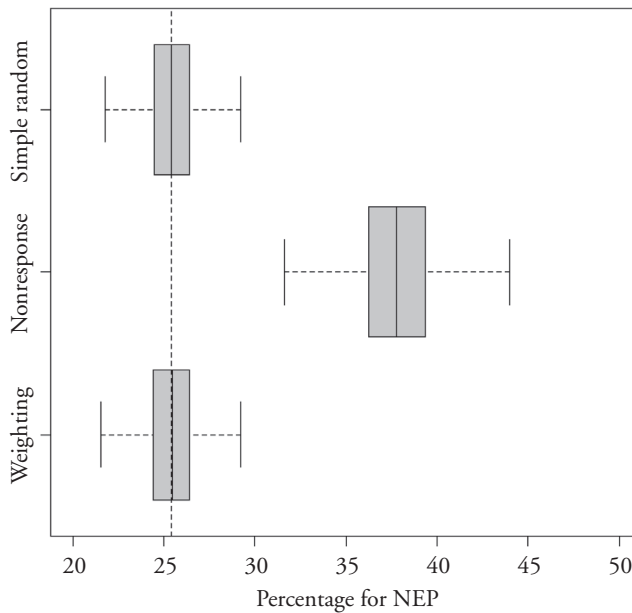


FIGURE 10.1 Estimating the percentage of voters for the NEP in the case of nonresponse

The vertical dotted line denotes the population percentage (25.4%). The box plot is symmetric around this line. It is clear that this estimator is unbiased.

The second situation describes what happens if nonresponse occurs. Each element is assigned a probability of response. This probability is equal to 0.2 for young people, 0.5 for middle-aged people, and 0.8 for the elderly. The box plot in the middle of Figure 10.1 shows the distribution of 1,000 estimates. This estimator is substantially biased. The expected value of this estimator is equal to 37.8%, which is much higher than 25.4%. This is not surprising. Young people have a low response probability and therefore are underrepresented in the sample. So, the elderly are overrepresented and they typically vote for the NEP.

The third situation examines the effect of poststratification on the nonresponse bias of the second situation. Age is used as the auxiliary variable. Adjustment weighting is successful here. The bias is completely removed. This could be expected as the response probabilities were equal within the age classes.

Poststratification is not always successful in reducing the bias. This is shown in a different example. Again, the same population of 30,000 is used, but now the percentage of voters for the New Internet Party (NIP) is estimated. There is a different nonresponse mechanism: persons with

Internet access have a response probability of 0.8, and those without it have a response probability of 0.2. Again, the effect of poststratification by age is explored. Figure 10.2 displays the results.

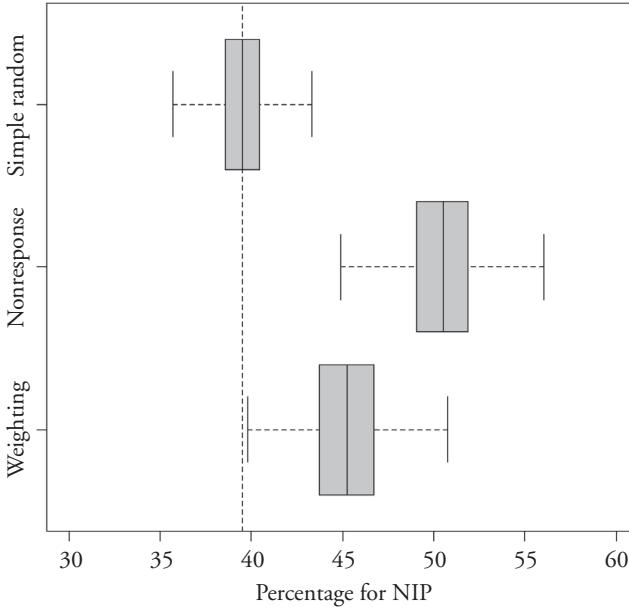


FIGURE 10.2 Estimating the percentage of voters for the NIP in the case of nonresponse

It is clear that a simple random sample results in an unbiased estimator for the population percentage (39.5%). Nonresponse leads to a substantial bias. The expected value of the estimator is now 50.7%. People with Internet access are overrepresented in the response, and they are the ones that typically vote for the NIP. Poststratification by age is not successful here. Only part of the bias is removed. The expected value goes down from 50.7% to 45.2%, but it is still too high. The reason is there is no direct relationship between age and response behavior. As there is a direct relation between access to the Internet and response behavior, and there is some relation between access to the Internet and age, part of the bias is removed.

Web surveys may suffer from *undercoverage*. This can happen when some people in the target population do not have access to the Internet. As a result, estimators may be biased. An expression for the bias was given in Section 10.1. The question is whether this bias can be removed or reduced by applying poststratification. It was shown in Chapter 8 that the bias after poststratification is equal to

$$(10.24) \quad B(\bar{y}_{I,PS}) = \sum_{b=1}^L W_b \frac{N_{NI,b}}{N_b} (\bar{Y}_I^{(b)} - \bar{Y}_{NI}^{(b)}),$$

where $N_{NI,b}$ is the number of people in stratum b without Internet access, $\bar{Y}_I^{(b)}$ is the mean of Y for those with Internet access in stratum b , and $\bar{Y}_{NI}^{(b)}$ is the mean for those without Internet access in stratum b . The bias will be small if there is (on average) no difference between elements with and without Internet access within the strata. This is the case if there is a strong relationship between the target variable Y and the stratification variable X . The variation in the values of Y will manifest itself in this case between strata but not within strata. In other words, the strata are homogeneous with respect to the target variable.

EXAMPLE 10.5 Using poststratification for reducing undercoverage bias

Suppose one of the aims of the election survey in example 10.1 is to estimate the percentage of people voting for the National Elderly Party (NEP). Three situations are considered:

- A simple random sample from the complete population
- A simple random sample from the Internet population
- A simple random sample from the Internet population, followed by poststratification

In all three situations, the distribution of the estimator is determined by repeating the selection of the sample 1,000 times. The sample size is always 1,000 cases. Figure 10.3 contains the results.

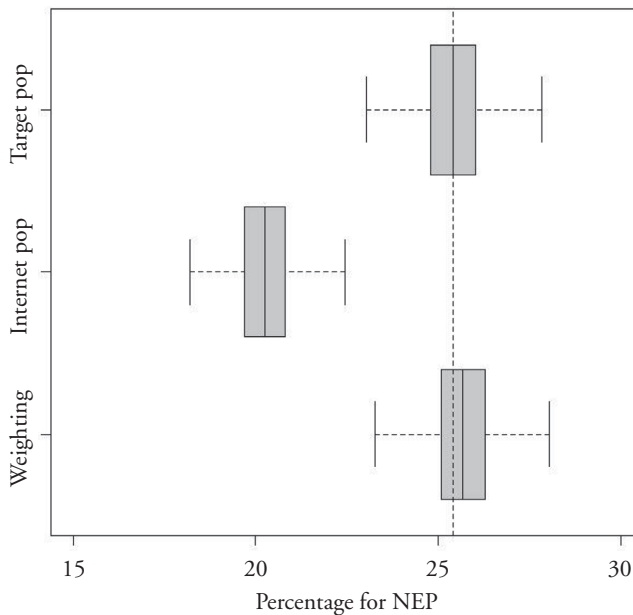


FIGURE 10.3 Estimating the percentage of voters for the NEP in the case of undercoverage

The population was constructed such that Internet access decreases with age. Moreover, Internet access for natives was much higher than for non-natives. Voting for the NEP depended on age only.

A simple random sample from the complete target population results in an unbiased estimator for the population percentage of 25.4%. Just sampling the Internet population leads to an estimator with a substantial bias. The expected value of this estimator is substantially too low: 20.3%. This can be explained by the fact that the elderly are underrepresented in the samples because they have access to Internet. The elderly typically vote for the NEP. Application of poststratification by age solves the problem. After weighting, the estimator is unbiased. There is a direct relation between voting behavior and age, and there is a direct relation between age and having Internet access. So, correcting the age distribution also corrects the estimator.

Figure 10.4 shows the analysis for voting for the New Internet Party (NIP). The same three situations are compared.

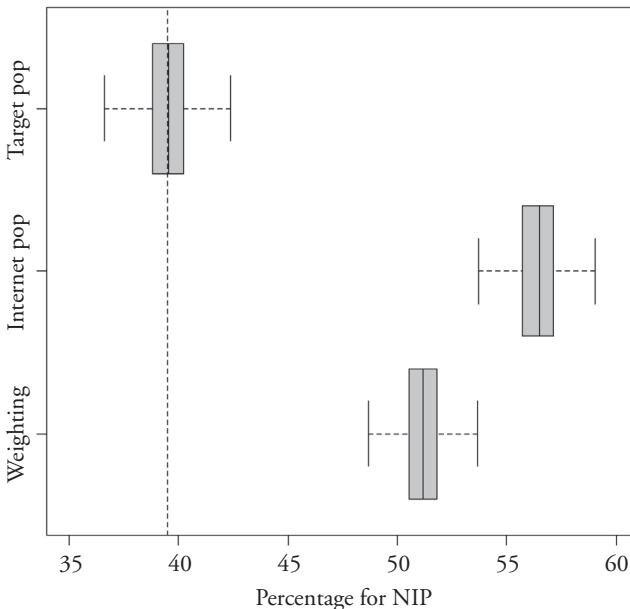


FIGURE 10.4 Estimating the percentage of voters for the NIP in the case of undercoverage

Just sampling the Internet leads to a biased estimator. The estimated values are substantially too high. The expected value of the estimator is 56.5%, whereas it should have been 39.5%. Poststratification is not successful. The expected value of the estimator decreases from 56.5% to 51.1%, but this is still too high. This is not surprising as there is a direct relation between voting for the NIP and having access to the Internet.

Web surveys may suffer from *self-selection*. This is the phenomenon that the sample is not selected by means of a probability sample. Instead, it is left to the Internet users themselves to participate in a web survey. The survey questionnaire is simply put on the web. Respondents are those people who happen to have Internet access, visit the website, and decide to participate in the survey. The survey researcher is not in control of the selection process. As a result, estimators may be biased. An expression for the bias was given in Section 10.1. The question is whether this bias can be removed or reduced by applying poststratification. It was shown in Chapter 9 that the bias after poststratification is equal to

$$(10.25) \quad B(\bar{y}_{S,PS}) = \sum_{b=1}^L W_b \frac{R_{\rho Y}^{(b)} S_{\rho}^{(b)} S_Y^{(b)}}{\bar{\rho}^{(b)}}$$

where the subscript b indicates that the respective quantities are computed just for stratum b . $R_{\rho Y}^{(b)}$ is the correlation coefficient between the target variable and the response behavior, $S_{\rho}^{(b)}$ is the standard deviation of the participation probabilities, $S_Y^{(b)}$ is the standard deviation of the target variable, and $\bar{\rho}^{(b)}$ is the average participation probability. The bias will be small if

- The participation propensities are similar within strata
- The values of the target variable are similar within strata
- There is no correlation between the participation behavior and the target variable within strata

These conditions can be realized if there is a strong relationship between the target variable Y and the stratification variable X . Then the variation in the values of Y manifests itself between strata and not within strata. In other words, the strata are homogeneous with respect to the target variable. Also if the strata are homogeneous with respect to the participation propensities, the bias will be reduced.

EXAMPLE 10.6 Using poststratification for reducing self-selection bias

The fictitious population of Section 9.3 is used to illustrate the possible effects of poststratification on a self-selection bias. This population consists of 100,000 persons. Most persons (99%) are passive Internet users. Active users make up only 1% of the population. Active users have a large participation probability of 0.99. Passive users have a small participation probability (0.01). The percentage of active Internet users decreases with age.

An election survey is conducted. The aim is to estimate the percentage of people voting for the National Elderly Party (NEP). Three situations are considered:

- A simple random sample from the complete population
- A self-election sample from the population
- A self-selection sample from the population, followed by poststratification

In all three situations, the distribution of the estimator determined by repeating the selection of the sample 1,000 times. The average participation probability in the population is 0.01971. Therefore, the expected sample size in a self-selection survey is equal to 1,971.

Figure 10.5 contains the results. The upper box plot shows that the estimator is unbiased in the case of simple random sampling (of size 1,971) from the target population. The expected value is equal to 25.6%. The middle box plot shows what happens if samples are selected by means of self-selection. The shape of the distribution remains more or less the same, but the distribution as a whole has shifted to the left. All values of the estimator are systematically too low. The expected value of the estimator is only 20.4%. The estimator is biased. The explanation of this bias is simple: Relative few elderly are active Internet users. Therefore, they are under-represented in the samples. They are typically people who will vote for the NEP. The lower box plot shows the distribution of the estimator in the case of poststratification by age. The bias is removed. This was possible because there is direct relation between participation and the weighting variable age.

Figure 10.6 shows the analysis for voting for the New Internet Party (NIP). The same three situations are compared.

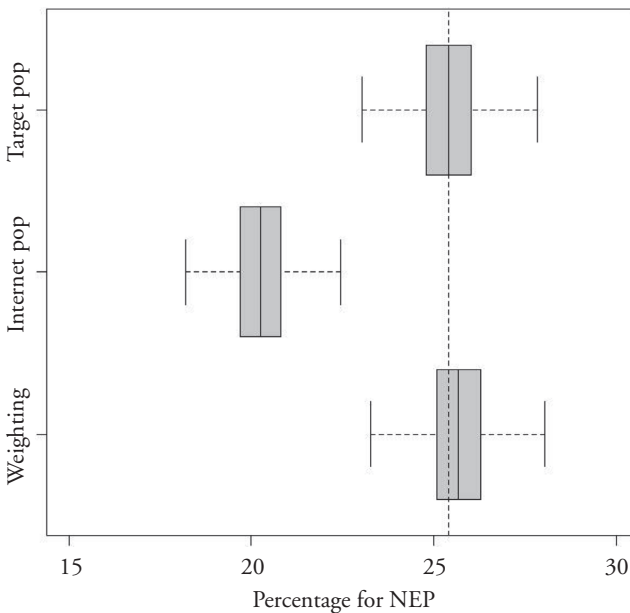


FIGURE 10.5 Estimating the percentage of voters for the NEP in the case of self-selection

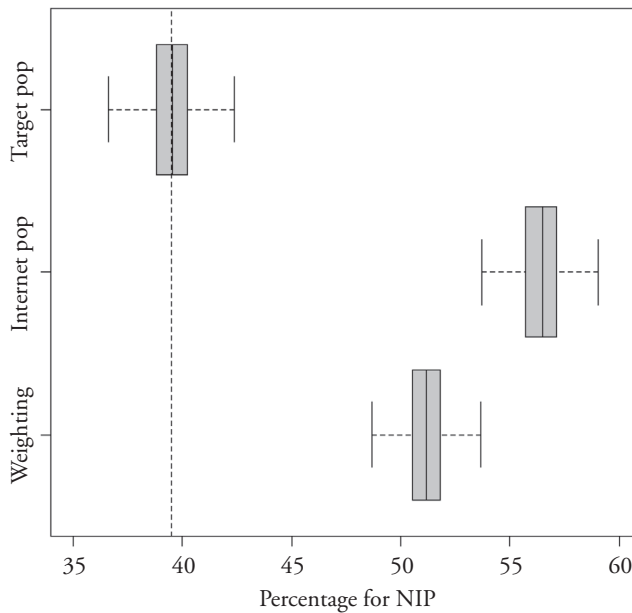


FIGURE 10.6 Estimating the percentage of voters for the NIP in the case of self-selection

In the case of simple random sampling, the estimator is unbiased. The population value of 39.5% is correctly estimated. Self-selection leads to a biased estimator. The estimated values are substantially too high. The expected value of the estimator is 56.5%, whereas it should have been 39.5%. Poststratification is not successful. The expected value of the estimator decreases from 56.5% to 51.1%, but this is still too high. This is not surprising as there is a direct relation between voting for the NIP and having access to the Internet.

Just one auxiliary variable (age) was used for weighting in the examples in this section. It is possible to use more than one variable for weighting. For example, if the two auxiliary variables Age and Education are available, they can be crossed. If age has three categories and Education has two categories, this leads to $3 \times 2 = 6$ strata. This weighting model is denoted by

Age \times Education.

The idea of crossing variables can be extended to more than two variables. As long as the table with population frequencies is available, and all response frequencies are greater than 0, weights can be computed. However, if there are no observations in a stratum, the corresponding weight cannot be computed. This leads to incorrect estimates. If the sample frequencies in the strata are very small, say less than 5, weights can be computed, but estimates will be unstable.

As more variables are included in a weighting model, there will be more strata. This increases the risk of empty strata or strata with too few observations. There are two solutions for this problem. One is not to have so many auxiliary variables in the model, but then a lot of auxiliary information is not used. Another is to *collapse strata*. This means merging a stratum with too few observations with another stratum. It is important to combine strata that resemble each other as much as possible. Collapsing strata is not a simple job, particularly if the number of auxiliary variables and strata is large.

Another problem of a large weighting model is that not all required population information is available. It may happen that the population distribution of the complete crossing of all auxiliary variables can simply not be obtained. A possible cause could be that the population distributions of these variables come from different sources. Without the complete population distribution, no weights can be computed.

One way to solve this problem is to use less auxiliary variables, but that would mean ignoring all available information with respect to the other variables. What is needed is a weighting technique capable of using partial population information. There are weighting techniques that can do this: generalized regression estimation and raking ratio estimation. These techniques are described in Sections 10.2.3 and 10.2.4.

EXAMPLE 10.7 Incomplete population information

The population distributions of the two variables Age (with categories Young, Middle, and Old) and Education (with two categories Low and High) are known separately, but the distribution in the cross-classification is not known. In this case, the poststratification Age \times Education cannot be carried out because weights cannot be computed for the strata in the cross-classification (See Table 10.2)

TABLE 10.2 Incomplete population information

	Population				Response		
	Low	High	Total		Low	High	Total
Young	?	?	19,925	Young	111	732	843
Middle	?	?	15,069	Middle	461	725	1,186
Old	?	?	15,006	Old	807	923	1,730
Total	29,963	20,037	50,000	Total	1,379	2,380	3,759

	Weights	
	Low	High
Young	?	?
Middle	?	?
Old	?	?

10.2.3 GENERALIZED REGRESSION ESTIMATION

The *generalized regression estimator* is based on a linear model that attempts to explain a target variable of the survey from one or more auxiliary variables. This estimator is not only capable of producing precise estimates, but it also can reduce a bias. It is shown that regression estimation in fact is a form of weighting.

The weights resulting from generalized regression estimation make the response representative with respect to the auxiliary variables in the model. It is shown that poststratification is a special case of linear weighting.

In principle, the auxiliary variables in the linear model have to be continuous variables (i.e., they measure a size or value). However, it is also possible to use categorical variables. The trick is to replace a categorical variable by a number of dummy variables, where each dummy variable indicates whether a person belongs to a specific category.

The theory is described assuming the data have been collected by means of simple random sampling without replacement. The theory can easily be generalized for other sampling designs. For more on this, see Bethlehem (1988).

First the ideal case (no bias) is considered. Suppose p (continuous) auxiliary variables are available. The p -vector of values of these variables for element k is denoted by

$$(10.26) \quad X_k = (X_{k1}, X_{k2}, \dots, X_{kp})'$$

The symbol $'$ denotes transposition of a matrix or vector. Let Y be the N -vector of all values of the target variable, and let X be the $N \times p$ -matrix of all values of the auxiliary variables. The vector of population means of the p auxiliary variables is defined by

$$(10.27) \quad \bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)'$$

This vector represents the population information assumed to be available. If the auxiliary variables are correlated with the target variable, then for a suitably chosen vector $B = (B_1, B_2, \dots, B_p)'$ of regression coefficients for a best fit of Y on X , the residuals $E = (E_1, E_2, \dots, E_N)'$, defined by

$$(10.28) \quad E = Y - XB$$

will vary less than the values of the target variable itself. In the ideal case of a perfect relation between Y and X , all residuals will be 0. Application of ordinary least squares results in

$$(10.29) \quad B = (X'X)^{-1}X'Y = \left(\sum_{k=1}^N X_k X_k' \right)^{-1} \left(\sum_{k=1}^N X_k Y_k \right).$$

For a simple random sample without replacement, the vector B can be estimated by

$$(10.30) \quad b = \left(\sum_{k=1}^N a_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^N a_k X_k Y_k \right) = \left(\sum_{i=1}^n x_i x_i' \right)^{-1} \left(\sum_{i=1}^n x_i y_i \right),$$

where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ denotes the p -vector of values of the p auxiliary variables for sample element i (for $i=1, 2, \dots, n$). The quantity a_k indicates whether element k is selected in the sample. The estimator b is an asymptotically design unbiased (ADU) estimator of B . It means the bias vanishes for large samples. The *generalized regression estimator* is now defined by

$$(10.31) \quad \bar{y}_{GR} = \bar{y} + (\bar{X} - \bar{x})' b,$$

where \bar{x} is the vector of sample means of the auxiliary variables.

The generalized regression estimator is an ADU estimator of the population mean of the target variable. If there exists a p -vector c of fixed numbers such that $Xc = J$, where J is a p -vector consisting of 1's, the generalized regression estimator can also be written as

$$(10.32) \quad \bar{y}_{GR} = \bar{X}' b.$$

This condition holds if there is a constant term in the regression model. The model also holds if the model contains a set of dummy variables corresponding to all the categories of a categorical variable. It can be shown that the variance of the generalized regression estimator is approximated by

$$(10.33) \quad V(\bar{y}_{GR}) = \frac{1-f}{n} S_E^2,$$

where S_E^2 is the population variance of the residuals E_1, E_2, \dots, E_N . Expression (10.31) is identical to the variance of the simple sample mean if the values Y_k are replaced by the residuals E_k . This variance will be small if the residual values E_k are small. Hence, use of auxiliary variables that can explain the behavior of the target variable will result in a precise estimator.

Bethlehem and Keller (1987) have shown that the generalized regression estimator (10.31) can be rewritten in the form of the weighted estimator (10.16). The adjustment weight w_i for observed element i is equal to $w_i = v' X_i$, and v is a vector of weight coefficients that is equal to

$$(10.34) \quad v = n \left(\sum_{i=1}^n x_i x_i' \right)^{-1} \bar{X}.$$

Poststratification is a special case of generalized regression estimation, where the auxiliary variables are categorical variables. To show this, categorical auxiliary variables are replaced by sets of dummy variables. Suppose there is one auxiliary variable with L categories. Then L dummy variables X_1, X_2, \dots, X_L are defined. For an observation in a certain stratum h , the corresponding dummy variable X_h is assigned the value 1, and all other dummy variables are set to 0. Consequently, the vector of population means of these dummy variables is equal to

$$(10.35) \quad \bar{X} = \left(\frac{N_1}{N}, \frac{N_2}{N}, \dots, \frac{N_L}{N} \right),$$

and v is equal to

$$(10.36) \quad v = \frac{n}{N} \left(\frac{N_1}{n_1}, \frac{N_2}{n_2}, \dots, \frac{N_L}{n_L} \right)'$$

If this form of v is used to compute $w_i = v'X_i$ and the result is substituted in expression (10.16) of the weighted estimator, the poststratification estimator is obtained.

EXAMPLE 10.8 Poststratification as a special case of generalized regression estimation

This example uses the same data as example 10.4. The objective of the survey is estimating voting behavior. A self-selection sample has been obtained from the population consisting of 30,000 people. The realized sample size is 3,696 people. There are two auxiliary variables: level of education (in two categories Low and High) and Ages (in three categories Young, Middle, and Old). Crossing these two variables produces a table with $2 \times 3 = 6$ cells. A dummy variable is introduced for each cell. So there are six dummy variables X_1, X_2, \dots, X_6 .

The possible values of these dummy variables are shown in Table 10.3. For example, X_4 is the dummy variable for the stratum consisting of young people with high education. Note that always one dummy variable has the value 1, whereas all other five dummy variables have the value 0.

The table also contains the vector of population means of the auxiliary variables. These values are equal to the population fractions in the cells of the population table. So the fraction of young people with high education in the population is equal to 0.161.

By comparing the population means with the sample means, it becomes clear that less educated young people are substantially under-represented in the sample. Their population fraction is 0.238, whereas their sample fraction is only 0.030.

TABLE 10.3 Poststratification by Education \times Age

Educ	Age	X_1	X_2	X_3	X_4	X_5	X_6
Low	Young	1	0	0	0	0	0
Low	Middle	0	1	0	0	0	0
Low	Old	0	0	1	0	0	0
High	Young	0	0	0	1	0	0
High	Middle	0	0	0	0	1	0
High	Old	0	0	0	0	0	1
Population means		0.238	0.181	0.181	0.161	0.121	0.119
Sample means		0.030	0.122	0.215	0.195	0.193	0.246
Weight coefficients		8.052	1.474	0.842	0.825	0.625	0.486

The weight coefficients in the vector v are given in the bottom row of the table. They have been computed using expression (10.36).

These weight coefficients are used to compute the adjustment weights for the observed elements. The weight of a person is obtained by summing the relevant weight coefficients. In the case of poststratification, there is always only one relevant weight coefficient. So the weight is equal to this weight coefficient. For example, the weight for a low educated young person is equal to 8.052. This implies that every sample person in this stratum counts for eight persons. Note that high educated old people are overrepresented. This is why they get a weight of only 0.486. Each person in this stratum counts for less than half a person.

In the case of undercoverage, the generalized regression estimator changes to

$$(10.37) \quad \bar{y}_{GR,I} = \bar{y}_I + (\bar{X} - \bar{x}_I)' b_I = \bar{X}' b_I.$$

The subscript I indicates that the corresponding quantities have been computed just using data from the Internet population. The vector of coefficients b_I is defined by

$$(10.38) \quad b_I = \left(\sum_{k=1}^N a_k I_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^N a_k I_k X_k Y_k \right),$$

where a_k is the sample indicator and I_k is the Internet indicator. So b_I is the analogue of b but just based on Internet population data. Bethlehem (1988) shows that the bias of estimator (10.37) is approximately equal to

$$(10.39) \quad B(\bar{y}_{GR,I}) = \bar{X}B_I - \bar{Y} = \bar{X}(B_I - B),$$

where B_I is defined by

$$(10.40) \quad B_I = \left(\sum_{k=1}^N I_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^N I_k X_k Y_k \right).$$

The bias of this estimator disappears if $B_I = B$. Thus, the regression estimator will be unbiased if undercoverage does not affect the regression coefficients. In Particular, if relationships are strong (the regression line fits the data well), the risk of finding a wrong relationship is small. By writing

$$(10.41) \quad B_I = B + \left(\sum_{k=1}^N I_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^N I_k X_k E_k \right),$$

the conclusion can indeed be drawn that the bias will be small if the residuals are small. This theory shows that use of the generalized regression estimator has the potential of reducing the bias caused by undercoverage.

In the case of self-selection, the generalized regression estimator changes to

$$(10.42) \quad \bar{y}_{GR,S} = \bar{y}_S + (\bar{X} - \bar{x}_S)' b_S = \bar{X}' b_S.$$

The subscript S indicates that the corresponding quantities have been computed using data from a self-selection sample. The vector of coefficients b_S is defined by

$$(10.43) \quad b_S = \left(\sum_{k=1}^N R_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^N R_k X_k Y_k \right),$$

in which R_k is the response indicator for element k . The bias of estimator (10.42) is approximately equal to

$$(10.44) \quad B(\bar{y}_{GR,S}) = \bar{X}B_S - \bar{Y} = \bar{X}(B_S - B),$$

where B_S is defined by

$$(10.45) \quad B_S = \left(\sum_{k=1}^N \rho_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^N \rho_k X_k Y_k \right).$$

The bias of this estimator disappears if $B_S = B$. Thus, the regression estimator will be unbiased if self-selection does not affect the regression coefficients. In Particular, if relationships are strong (the regression line fits the data well), the risk of finding a wrong relationship is small. By writing

$$(10.46) \quad B_S = B + \left(\sum_{k=1}^N \rho_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^N \rho_k X_k E_k \right),$$

the conclusion also can here be drawn that the bias will be small if the residuals are small. This theory shows that use of the generalized regression estimator has the potential of reducing the bias resulting from self-selection.

Generalized regression estimation can address the problem of the lack of sufficient population information. It is possible to include variables in the weighting model without having to know the population frequencies in the cells obtained by cross-tabulating all variables. The trick is to use a different set of dummy variables. Instead of defining one set of dummy variables for the complete crossing of all auxiliary variables, a set of dummy variables is defined for each variable separately or for each crossing of subsets of variables separately.

Suppose there are three auxiliary variables X_1 , X_2 , and X_3 . Poststratification would come down to crossing the three variables, using one set of dummy variables. If only the marginal population distributions of the three variables can be used, there are three sets of dummy variables, each corresponding to the categories of one auxiliary variable. And if, for example, the population distribution of the crossing of X_1 and X_2 is available and only the marginal distribution of X_3 , there are two sets of dummy variables: one for $X_1 \times X_2$ and one for X_3 . Of course, other combinations and subsets are possible, depending on the available auxiliary information and the number of observations in each cell of each cross-classification.

In the case of poststratification, the weight is equal to one of the weight coefficients. If the weighting model contains more than one set of dummy variables, there will also be more weight coefficients contributing to the weight. In fact, each set contributes a weight coefficient, and these weights are added to obtain the weight.

It should be noted that a weighting model containing more than one set of dummy variables will use less information than the model for the complete crossing of all auxiliary variables. Nevertheless, it uses more information than a poststratification corresponding to one of the subsets.

EXAMPLE 10.9 Generalized regression estimation using only marginal distributions

Continuing Example 10.8, it is now shown how to use just only the marginal distributions of education and age. Two sets of dummy variables are introduced: one set of two dummy variables for the categories of education, and another set of three dummy variables for the categories of age. Then there are $2 + 3 = 5$ dummy variables. In each set, always one dummy has the value 1, whereas all other dummies are 0. The possible values of the dummy variables are shown in Table 10.4.

TABLE 10.4 Weighting with the marginal distributions

Educ	Age	X_1	X_2	X_3	X_4	X_5	X_6
Low	Young	1	1	0	1	0	0
Low	Middle	1	1	0	0	1	0
Low	Old	1	1	0	0	0	1
High	Young	1	0	1	1	0	0
High	Middle	1	0	1	0	1	0
High	Old	1	0	1	0	0	1
Population means		1.000	0.599	0.401	0.399	0.301	0.300
Weight coefficients		1.359	0.673	-0.673	0.916	-0.255	-0.662

The first dummy variable X_1 represents the constant term in the regression model. It always has the value 1. The second and third dummy variable relate to the two categories of Education (Low and High), and the last three dummies represent the three Age categories. The vector of population means is equal to the fractions for all dummy variables separately. Note that in this weighting model always three dummies in a row have the value 1.

The weight for an observed element is now obtained by summing the appropriate elements of this vector. The first value corresponds to the dummy X_1 , which always has the value 1. So there is always a contribution 1.359 to the weight. The next two values correspond to the categories of education. Note that their sum equals zero. For a low education, an amount 0.673 is added, and for a high education, the same amount is subtracted. The final three values correspond to the categories of age. Depending on the age category, a contribution is added or subtracted. For example, the weight for a low-educated young person is equal to $1.359 + 0.673 + 0.916 = 2.948$.

No information is used about the crossing of Education by Age in Example 10.9. Only the marginal distributions are included in the computation of the weights. Therefore, a different notation is introduced. This weighting model is denoted by

$$\text{Education} + \text{Age}.$$

Because of the special structure of the auxiliary variables, the computation of the weight coefficients v cannot be carried out without imposing extra conditions. Here, for every categorical variable, the condition is imposed that the sum of the weight coefficients for the corresponding dummy variables must equal zero.

Example 10.9 uses only two auxiliary variables. More variables can be included in a weighting model. This makes it possible to define various weighting models with these variables. Suppose there are three auxiliary variables: Education, Age, and Gender. If the complete population distribution on the crossing of all three variables is available, then the weighting model

$$\text{Education} \times \text{Age} \times \text{Gender}$$

can be applied. If only the bivariate population distributions of every crossing of two variables are available, the following weighting scheme could be applied:

$$(\text{Education} \times \text{Age}) + (\text{Age} \times \text{Gender}) + (\text{Education} \times \text{Gender}).$$

Note that this scheme comes down to doing three poststratifications simultaneously. If only marginal frequency distributions are available, the model

$$\text{Education} + \text{Age} + \text{Gender}$$

could be considered. More details about the theory of generalized regression estimation can be found for example in Bethlehem and Keller (1987).

Until now only generalized regression estimation with categorical auxiliary variables was described. It is also possible to apply this estimation technique with continuous auxiliary variables or a combination of categorical and continuous variables. See Bethlehem (2009) for more details.

EXAMPLE 10.10 Using generalized regression estimation for reducing self-selection bias

A fictitious population is used to illustrate the possible effects of generalized regression estimation on a self-selection bias. This population consists of 50,000 eligible voters in a town. The aim of a web survey is to measure whether people intend to vote at the next local elections. Voting depends on age and level of education. Voting increases with age. Voting is higher among high educated people than among low educated people. The percentage of voters in the population is 46.1%.

Participation in the web survey depends on age and education: older people are more likely to participate than younger people. The average

participation probability in the population was 0.075. Hence, the expected response was $0.075 \times 50,000 = 3,750$.

A simulation was carried out in which 1,000 samples were selected from this population. The distribution of the five different estimators was compared:

- The mean of a simple random sample
- The mean of a self-selection sample
- Poststratification by education of the self-selection sample
- Poststratification by age the self-selection sample
- The generalized regression estimator that uses only the marginal distributions of education and age
- Poststratification by education and age of the self-selection sample

Figure 10.7 contains the results. The upper box plot shows the distribution of the sample mean in case of simple random sampling. It is clear that this estimator is unbiased.

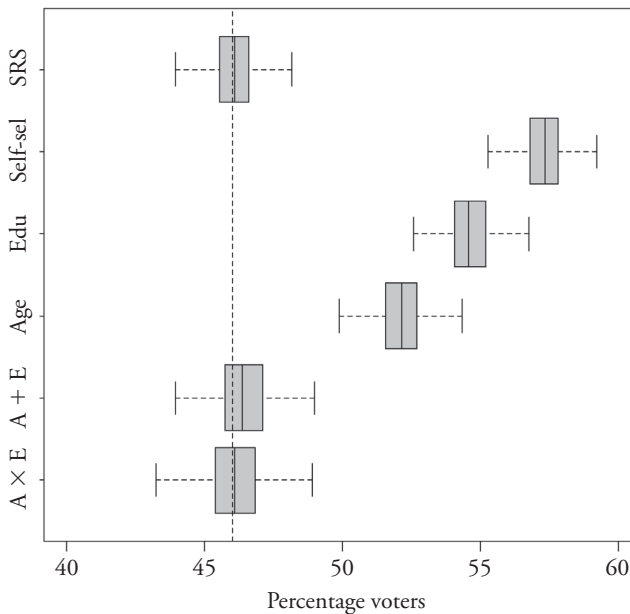


FIGURE 10.7 Estimating the percentage of voters for the NEP in the case of selection

The second box plot shows what happens in the case of self-selection. The estimator has a substantial upward bias. This is not surprising as people with a large participation probability also are more inclined to vote.

The third and fourth box plot shows to what extent the bias is reduced if poststratification is carried out with just one auxiliary variable. In both cases, the bias is reduced somewhat but not completely removed. Age seems to be more effective than Education.

The sixth box plot ($A \times E$, at the bottom) displays the distribution of the estimator in the case of poststratification by Age and Education. The bias is completely removed. This can be expected as the participation probabilities are equal within the strata formed by crossing Age and Education.

Finally, the fifth box plot shows the result of using the regression estimator with only the marginal distributions ($A + E$). Apparently, this estimator performs as well as poststratification with a complete crossing. This can be explained by the fact that there are no special interaction effects between the two auxiliary variables. This effect is often observed in practice. What matters is to have a specific set of auxiliary variables in the model. How they are used (crossed or marginally) is often less important.

10.2.4 RAKING RATIO ESTIMATION

If generalized regression estimation is applied, correction weights are obtained by taking the sum of a number of *weight coefficients*. It is also possible to compute correction weights in a different way, namely, as the product of a number of *weight factors*. This weighting technique is usually called *raking ratio estimation*, *raking*, or *multiplicative weighting*. Here it is denoted by *raking ratio estimation*. Weights are obtained as the product of several factors contributed by the various auxiliary variables in the model.

Raking ratio estimation can be applied in the same situations as generalized regression estimation as long as only categorical auxiliary variables are used. It computes correction weights by means of an iterative procedure. The resulting weights are the product of factors contributed by all cross-classifications in the model.

The technique of raking ratio estimation was already described by Deming and Stephan (1940). Skinner (1991) discussed application of this technique in multiple frame surveys. Little and Wu (1991) described the theoretical framework and showed that this technique comes down to fitting a log-linear model for the probabilities of getting observations in strata of the complete cross-classification given the probabilities for marginal distributions. To compute the weight factors, the following scheme must be carried out:

Step 1: Introduce a weight factor for each stratum in each cross-classification term. Set the initial values of all factors to 1.

Step 2: Adjust the weight factors for the first cross-classification term so that the weighted sample becomes representative with respect to the auxiliary variables included in this cross-classification.

Step 3: Adjust the weight factors for the next cross-classification term so that the weighted sample becomes representative for the variables involved.

Generally, this will disturb representativeness with respect to the other cross-classification terms in the model.

Step 4: Repeat this adjustment process until all cross-classification terms are dealt with.

Step 5: Repeat steps 2, 3, and 4 until the weight factors do not change any more.

EXAMPLE 10.11 Raking ratio estimation

Raking ratio estimation is illustrated using the same data as in Example 10.9. Two variables are used in the weighting model: Age (three categories) and Education (two categories).

Suppose only the marginal population distributions of Age (three categories) and Education (two categories) are available, not the cross-classification. Table 10.5 contains the starting situation. The upper-left part of the table contains the unweighted relative frequencies in the sample for each combination of Age and Education. The row and column denoted by “Weight factor” contain the initial values of the weight factors (1.000). The values in the row and column denoted by “Weighted sum” are obtained by first computing the weight for each sample cell (by multiplying the relevant row and column factor), and then summing the weighted cell fractions. Because the initial values of all factors are equal to 1, the weighted sums in the table are equal to the unweighted sample sums. The row and column denoted by “Population distribution” contain the fractions in the Age and Education categories in the population.

TABLE 10.5 The starting situation

	Low educ	High educ	Weight factor	Weighted sum	Population distribution
Young	0.030	0.195	1.000	0.225	0.399
Middle	0.122	0.193	1.000	0.315	0.301
Old	0.215	0.246	1.000	0.461	0.300
Weight factor	1.000	1.000			
Weighted sum	0.367	0.634		1.000	
Popul. Distr.	0.599	0.401			1.000

The iterative process must result in row and column factors with such values that the weighted sums match the population distribution. This is clearly not the case in the starting situation. First, the weight factors for the rows are adjusted. This leads to weight factors 1.773, 0.956, and 0.651 for the categories Young, Middle, and Old; see Table 10.6. The weighted sums for the rows are now correct, but the weighted sums for the columns are 0.310 and 0.690 and, thus, still show a discrepancy.

TABLE 10.6 Situation after adjusting for age

	Low educ	High educ	Weight factor	Weighted sum	Population distribution
Young	0.030	0.195	1.773	0.399	0.399
Middle	0.122	0.193	0.956	0.301	0.301
Old	0.215	0.246	0.651	0.300	0.300
Weight factor	1.000	1.000			
Weighted sum	0.310	0.690		1.000	
Popul. Distr.	0.599	0.401			1.000

The next step is to adjust the weight factors for the columns such that the weighted column sums match the corresponding population frequencies. Note that this adjustment for Education will disturb the adjustment for Age. The weighted sums for the age categories no longer match the relative population frequencies. However, the discrepancy is much smaller than in the initial situation. See Table 10.7.

TABLE 10.7 Situation after adjusting for education

	Low educ	High educ	Weight factor	Weighted sum	Population distribution
Young	0.030	0.195	1.773	0.304	0.399
Middle	0.122	0.193	0.956	0.333	0.301
Old	0.215	0.246	0.651	0.364	0.300
Weight factor	1.934	0.581			
Weighted sum	0.599	0.401		1.000	
Popul. distr.	0.599	0.401			1.000

The process of adjusting for Age and Education is repeated until the weight factors do not change any more. The final situation is reached after a few iterations. Table 10.8 contains the final results.

TABLE 10.8 Situation after convergence

	Male	Female	Weight factor	Weighted sum	Population distribution
Young	0.030	0.195	2.422	0.399	0.399
Middle	0.122	0.193	0.848	0.301	0.301
Old	0.215	0.246	0.521	0.300	0.300
Weight factor	2.080	0.525			
Weighted sum	0.599	0.401		1.000	
Popul. distr.	0.599	0.401			1.000

The adjustment weight for a specific sample element is now obtained by multiplying the relevant weight factors. For example, the weight for a young male is equal to $2.422 \times 2.080 = 5.038$. Note that for this example the adjustment weights differ from those obtained by the generalized regression estimator in Example 10.9.

There are many situations in which both the generalized regression estimator and the raking ratio estimator can be applied. This raises the question of whether the estimation method should be preferred. Several observations may help to take a decision.

In the first place, generalized regression estimation is based on a simple linear model that describes the relationship between a target variable and a number of auxiliary variables. If this model fits well, weighting adjustment will be effective. For raking ratio estimation, there is no straightforward model allowing for simple interpretation.

In the second place, computations for generalized regression estimation are straightforward. Weights are obtained by application of ordinary least squares. The weights of raking ratio estimation are obtained as the solution of an iterative process. There is no guarantee this process will always converge.

In the third place, for generalized regression estimation, it is possible to derive an analytical expression of the variance of weighted estimates. No simple expressions are available for estimates based on raking ratio estimation.

In the fourth place, weights produced by linear weighting may sometimes turn out to be negative. This seems counterintuitive, but it is simply a consequence of the linear model applied. Negative weights usually indicate that the linear model does not fit too well. A disadvantage of negative weights is that some statistical analysis packages do not accept negative weights. This may prevent weighted analysis of the survey outcomes.

In the fifth place, it has been shown (see Section 10.2.5) that in many situations estimators based on linear weights have asymptotically the same properties as those based on multiplicative weights.

10.2.5 CALIBRATION ESTIMATION

Deville and Särndal (1992) and Deville, Särndal, and Sautory (1993) have proposed a general framework for weighting of which generalized regression estimation and raking ratio estimation are special cases. Assuming simple random sampling, their starting point is that the correction weights c_i in $w_i = c_i \times d_i$ have to satisfy two conditions:

1. The correction weights c_i have to be as close as possible to 1.
2. The weighted sample distribution of the auxiliary variables has to match the population distribution; i.e.,

$$(10.47) \quad \bar{x}_W = \frac{1}{N} \sum_{i=1}^n w_i x_i = \bar{X}.$$

The first condition sees to it that the resulting estimators are unbiased, or almost unbiased, and the second condition guarantees that the weighted sample is representative with respect to the auxiliary variables used.

Deville and Särndal (1992) introduce general distance measure $D(c_i, 1)$ measuring the difference between c_i and 1. The problem is then to minimize

$$(10.48) \quad \sum_{i=1}^n D(c_i, 1)$$

under the condition (10.45). This problem can be solved by using the method of Lagrange. By choosing the proper distance function, generalized regression estimation and raking ratio estimation are obtained as special cases of this general approach. For generalized regression estimation, the distance function is defined by

$$(10.49) \quad D(c_i, 1) = (c_i - 1)^2,$$

which is the Euclidean distance, and for raking ratio estimation, the distance

$$(10.50) \quad D(c_i, 1) = c_i \log(c_i) - c_i + 1$$

must be used.

Deville and Särndal (1992) and Deville et al. (1993) only consider the full response situation. They show that estimators based on weights computed within their framework have asymptotically the same properties. This means that for large samples it does not matter which of the two estimation techniques is applied. Estimators based on both weighting techniques will behave approximately the same way. Note that, although the estimators behave similarly, the individual weights computed by means of generalized regression estimation or raking ratio estimation may differ substantially.

Under nonresponse, undercoverage, or self-selection, the situation is different. Then the asymptotic properties of both estimation techniques will generally not be equal. The extent to which the chosen weighting technique is able to reduce the bias depends on how well the corresponding underlying model can be estimated using the observed data. Generalized regression estimation assumes a linear model to hold with the target variable as dependent variable and the auxiliary variables as explanatory variables. Raking ratio estimation assumes a log-linear model for the cell frequencies. An attempt to use a correction technique for which the underlying model does not hold will not help to reduce the bias.

10.2.6 CONSTRAINING THE VALUES OF WEIGHTS

There are several reasons why survey researchers may want to have some control over the values of the adjustment weights. One reason is that extremely large weights are generally considered undesirable. Large weights usually correspond to

population elements with rare characteristics. Use of such weights may lead to unstable estimates of population parameters. To reduce the impact of large weights on estimators, a weighting method is required that is able to keep adjustment weights within prespecified boundaries, and that at the same time enables valid inference.

Another reason to have some control over the values of the adjustment weights is that application of generalized regression estimation may produce negative weights. Although the theory does not require weights to be positive, negative weights should be avoided; because they are counterintuitive, they cause problems in subsequent analyses, and they are an indication that the regression model does not fit the data well.

Negative weights can be avoided by using a better regression model. However, it is not always possible to find such models. Another solution is to use the current model and force weights within certain limits. Several techniques have been proposed for this. A technique developed by Deville et al. (1993) comes down to repeating the regression estimation process several times. First, a lower bound L and an upper bound U are specified. After the first run, weights smaller than L are set to L and weights larger than U are set to U . Then, the weighting process is repeated, but records from the strata with the fixed weights L and U are excluded. Again, weights may be produced not satisfying the conditions. These weights are also set to either the value L or the value U . The weighting process is repeated until all computed weights fall within the specified limits. Convergence of this iterative process is not guaranteed. In Particular, if the lower bound L and upper bound U are not far apart, the algorithm may not converge.

Huang and Fuller (1978) use a different approach. Their algorithm produces weights that are a smooth, continuous, monotone increasing function of the original weights computed from the linear model. The algorithm is iterative. At each step, the weights are checked against a user-supplied criterion value M . This value M is the maximum fraction of the mean weight by which any weight may deviate from the mean weight. For example, if M is set to 0.75, then all weights are forced into the interval with the lower bound equal to 0.25 times the mean weight and the upper bound equal to 1.75 times the mean weight. Setting the value to 1 implies that all weights are forced to be positive. Huang and Fuller (1978) prove that the asymptotic properties of the regression estimator constructed with their algorithm are asymptotically the same as those of the generalized regression estimator. So, restricting the weights has (at least asymptotically) no effect on the properties of population estimates computed with these weights.

10.2.7 CORRECTION USING A REFERENCE SURVEY

Poststratification, generalized regression estimation, and raking ratio estimation can be effective bias reduction techniques provided auxiliary variables are available that have a strong correlation with the target variables of the survey. If such variables cannot be used because their population distribution is not

available, one might consider estimating these population distributions in a different survey, a so-called *reference survey*. This reference survey must be based on a probability sample, where data collection takes place with a mode different from the web, e.g., CAPI (computer-assisted personal interviewing, with laptops) or CATI (computer-assisted telephone interviewing). Preferably, the sample size of this reference survey must be small to keep costs within limits. The reference survey approach has been applied by several market research organizations. See, for example, Börsch-Supan et al. (2004) and Duffy et al. (2005).

Under the assumption of no nonresponse, or ignorable nonresponse, this reference survey can produce unbiased estimates of quantities that also have been measured in the web survey. Unbiased estimates for the target variable can be computed, but because of the small sample size, these estimates will have a substantial variance. The question is now whether estimates of population characteristics can be improved by combining the large sample size of the web survey with the unbiasedness of the reference survey.

First, it will be explored whether a reference survey can reduce an undercoverage bias. Then the effect on a self-selection bias is analyzed. Only poststratification with one auxiliary variable is considered as an adjustment method.

It is assumed that one categorical auxiliary variable is observed in both the web survey and the reference survey, and that this variable has a strong correlation with the target variable of the survey. Then a form of poststratification can be applied where the stratum means are estimated using the web survey data and the stratum weights are estimated using the reference survey data. Suppose that m is the sample size of the reference survey and that m_b is the number of observed elements in stratum b . This leads to the poststratification estimator:

$$(10.51) \quad \bar{y}_{I,RS} = \sum_{b=1}^L \frac{m_b}{m} \bar{y}_I^{(b)},$$

where $\bar{y}_I^{(b)}$ is the web-survey-based estimate for the mean of stratum b of the Internet population (for $b = 1, 2, \dots, L$) and m_b / m is the relative sample size in stratum b as estimated in the reference survey sample (for $b = 1, 2, \dots, L$). Under the conditions described above, the quantity m_b / m is an unbiased estimate of $W_b = N_b / N$.

Let I denote the probability distribution for the web survey, and let P be the probability distribution for the reference survey. Then the expected value of the poststratification estimator is equal to

$$(10.52) \quad E(\bar{y}_{I,RS}) = E_I E_P(\bar{y}_{I,RS} | I) = E_I \left(\sum_{b=1}^L \frac{N_b}{N} \bar{y}_I^{(b)} \right) = \sum_{b=1}^L W_b \bar{Y}_I^{(b)} = \tilde{Y}_I.$$

where $W_b = N_b / N$ is the relative size of stratum b in the target population and $\bar{Y}_I^{(b)}$ is the mean of the target variable of stratum b of the Internet population. The expected

value of this estimator is identical to that of the poststratification estimator (8.16). The bias of this estimator is equal to

$$\begin{aligned}
 (10.53) \quad B(\bar{y}_{I,RS}) &= E(\bar{y}_{I,RS}) - \bar{Y} = \tilde{Y}_I - \bar{Y} = \sum_{b=1}^L W_b (\bar{Y}_I^{(b)} - \bar{Y}^{(b)}) \\
 &= \sum_{b=1}^L W_b \frac{N_{NI,b}}{N_b} (\bar{Y}_I^{(b)} - \bar{Y}_{NI}^{(b)}).
 \end{aligned}$$

If a strong relationship exists between the target variable and the auxiliary variable, there is little or no variation of the target variable within the strata. This implies that if the stratum means for the Internet population and for the target population do not differ much, this results in a small bias. So, using a reference survey with the proper auxiliary variables can substantially reduce the bias of web survey estimates.

Note that the expression for the bias of the reference survey estimator is equal to that of the poststratification estimator. An interesting aspect of the reference survey approach is that any variable can be used for adjustment weighting as long as it is measured in both surveys. For example, some market research organizations use “webographics” or “psychographic” variables that divide the population into “mentality groups”. People in the same groups have more or less the same level of motivation and interest to participate in such surveys. Deployment of effective weighting variables resembles the Missing At Random (MAR) situation. This implies that within weighting strata, there is no relationship between participating in a web survey and the target variables of the survey.

Bethlehem (2007) shows that if a reference survey is used, the variance of the poststratification estimator is equal to

$$\begin{aligned}
 (10.54) \quad V(\bar{y}_{I,RS}) &= \frac{1}{m} \sum_{b=1}^L W_b (\bar{Y}_I^{(b)} - \tilde{Y}_I)^2 + \frac{1}{m} \sum_{b=1}^L W_b (1 - W_b) V(\bar{y}_I^{(b)}) \\
 &\quad + \sum_{b=1}^L W_b^2 V(\bar{y}_I^{(b)}).
 \end{aligned}$$

The quantity $\bar{y}_I^{(b)}$ is measured in the web survey. Therefore its variance $V(\bar{y}_I^{(b)})$ will be of the order $1/n$. This means that the first term in the variance of the poststratification estimator will be of the order $1/m$, the second term of order $1/mn$, and the third term of order $1/n$. As n will generally be much larger than m in practical situations, the first term in the variance will dominate; i.e., the (small) size of the reference survey will determine the accuracy of the estimates. So, the large number of observations in the web survey does not help to produce accurate estimates. One could say that the reference survey approach reduces the bias of estimates at the cost of a higher variance.

EXAMPLE 10.12 Using a reference survey for reducing undercoverage bias

Suppose one aim of the election survey in Example 10.1 is to estimate the percentage of people voting for the National Elderly Party (NEP). Four situations are considered:

- A simple random sample from the complete population
- A simple random sample from the Internet population
- A simple random sample from the Internet population, followed by poststratification
- A simple random sample from the Internet population, followed by poststratification based on a reference survey

In all four situations, the distribution of the estimator was determined by repeating the selection of the sample 1,000 times. The sample size is always 1,000 cases. Figure 10.8 contains the results.

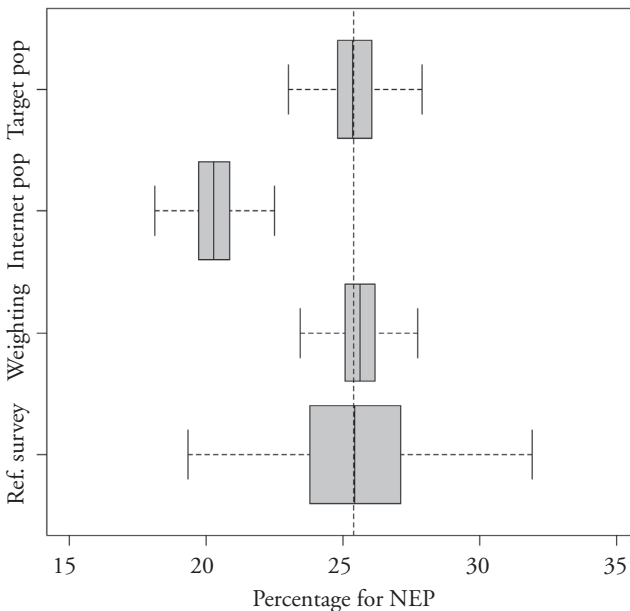


FIGURE 10.8 Estimating the percentage of voters for the NEP in the case of undercoverage

The population was constructed such that Internet access decreases with age. Moreover, Internet access for natives was much higher than for non-natives. Voting for the NEP depended on age only.

A simple random sample from the complete target population results in an unbiased estimator for the population percentage of 25.4%. Just sampling the Internet population leads to an estimator with a substantial bias. The expected value of this estimator is substantially too low: 20.3%. This can be explained by the fact that the elderly are underrepresented in the samples because they have less access to Internet. The elderly typically vote for the NEP.

Application of poststratification by age solves the problem. After weighting, the estimator is unbiased. There is a direct relation between voting behavior and age, and there is a direct relation between age and having Internet access. So, correcting the age distribution also corrects the estimator.

The lower box plot shows the distribution of the estimator if the population distribution of the weighting variable age is estimated in the reference survey with a sample size of $m=100$. The bias is removed but at the cost of a substantial increase of the variance. This is from the small sample size of the reference survey. Of course, one could consider increasing this sample size, but this also increases the costs. One may even wonder why to conduct a web survey at all, if also a reference survey is carried out.

Figure 10.9 shows the analysis for voting for the New Internet Party (NIP). The same four situations are compared.

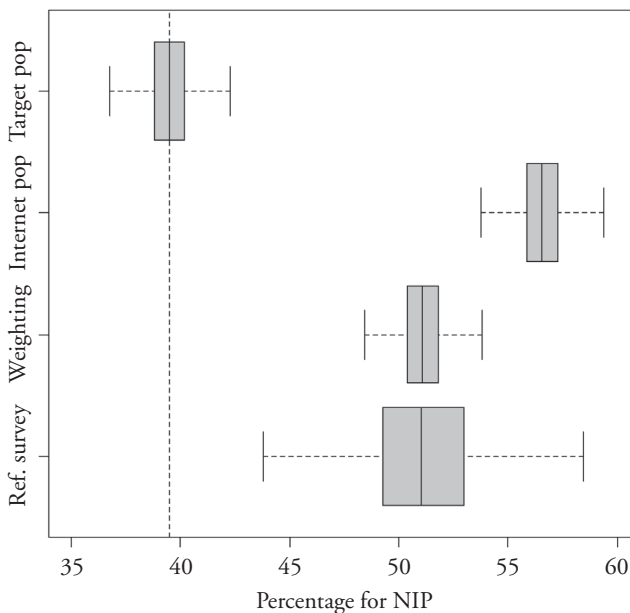


FIGURE 10.9 Estimating the percentage of voters for the NIP in the case of undercoverage

Just sampling the Internet leads to a biased estimator. The estimated values are substantially too high. The expected value of the estimator is 56.5%, whereas it should have been 39.5%. Poststratification is not successful. The expected value of the estimator decreases from 56.5% to 51.1%, but this is still too high. This is not surprising as there is a direct relation between voting for the NIP and having access to the Internet. Poststratification based on a reference survey will not solve the problem here. The bias remains, and at the same time the variance increases.

Now the case of self-selection is considered. Again, it is assumed that one categorical auxiliary variable is observed in the web survey and the reference survey, and that this variable has a strong correlation with the target variable of the survey. Poststratification is applied where the stratum means are estimated using web survey data and the stratum weights are estimated using the reference survey data. This leads to the poststratification estimator

$$(10.55) \quad \bar{y}_{S,RS} = \sum_{b=1}^L \frac{m_b}{m} \bar{y}_S^{(b)},$$

where $\bar{y}_S^{(b)}$ is the web-survey-based estimate for the mean of stratum b of the target population (for $b=1, 2, \dots, L$) and m_b / m is the estimated relative sample size in stratum b using the reference survey (for $b=1, 2, \dots, L$). Under the conditions described, the quantity m_b / m is an unbiased estimate of $W_b = N_b / N$.

Let I denote the probability distribution for the web survey, and let P be the probability distribution for the reference survey. Then the expected value of the poststratification estimator is equal to

$$(10.56) \quad E(\bar{y}_{S,RS}) = E_I E_P(\bar{y}_{S,RS} | I) = E_I \left(\sum_{b=1}^L \frac{N_b}{N} \bar{y}_b \right) = \sum_{b=1}^L W_b \bar{Y}_b^* = \tilde{Y}^*.$$

So, the expected value of this estimator is identical to that of the poststratification estimator (9.34). The bias of this estimator is equal to

$$(10.57) \quad \begin{aligned} B(\bar{y}_{S,RS}) &= E(\bar{y}_{S,RS}) - \bar{Y} = \tilde{Y}^* - \bar{Y} \\ &= \sum_{b=1}^L W_b (\bar{Y}_b^* - \bar{Y}_b) = \sum_{b=1}^L W_b \frac{R_{\rho Y}^{(b)} S_{\rho}^{(b)} S_Y^{(b)}}{\bar{\rho}^{(b)}}. \end{aligned}$$

If there is a strong relationship between the target variable and the auxiliary variable used for computing the weights, there is little or no variation of the

target variable within the strata. Consequently, the correlation between target variable and response behavior will be small, and the same applies to the standard deviation of the target variable. So, using a reference survey with the proper auxiliary variables can substantially reduce the bias of web survey estimates.

Bethlehem (2008) shows that the variance of estimator (10.55) is equal to

$$(10.58) \quad V(\bar{y}_{S,RS}) = \frac{1}{m} \sum_{b=1}^L W_b (\bar{Y}_b^* - \tilde{Y}^*)^2 + \frac{1}{m} \sum_{b=1}^L W_b (1 - W_b) V(\bar{y}_S^{(b)}) \\ + \sum_{b=1}^L W_b^2 V(\bar{y}_S^{(b)}).$$

The quantity $\bar{y}_S^{(b)}$ is measured in the on-line survey. Therefore its variance $V(\bar{y}_S^{(b)})$ will be at most of the order $1/E(n_S) = 1/(N\bar{p})$, where n_S is the size of the self-selection sample. This means that the first term in the variance of the poststratification estimator will be of the order $1/m$, the second term of order $1/(mE(n_S))$, and the third term of order $1/E(n_S)$. As $E(n_S)$ will generally be much larger than m in practical situations, the first term in the variance will dominate; i.e., the (small) size of the reference survey will determine the accuracy of the estimates.

Moreover, because strata preferably are based on groups of people with the same psychographic characteristics, and target variables may very well be related to the psychographic variables, the stratum means \bar{Y}_b^* may vary substantially. This also contributes to a large value of the first variance component.

The conclusion is that a large number of observations in the web survey does not help to produce accurate estimates. The reference survey approach may reduce the bias of estimates, but it does so at the cost of a higher variance.

The effectiveness of a survey design is sometimes also indicated by means of the *effective sample size*. This is the sample size of a simple random sample of elements that would produce an estimator with the same precision. Use of a reference survey implies that the effective sample size is much lower than the size of the web survey.

EXAMPLE 10.13 Using a reference survey for reducing self-selection bias

The fictitious population of Section 10.5 is used to illustrate the possible effects of poststratification with a reference survey on a self-selection bias. This population consists of 100,000 persons. Most persons (99%) are passive Internet users. Active users make up only 1% of the population. Active users have a large participation probability of 0.99. Passive users

have a small participation probability (0.01). The percentage of active Internet users decreases with age.

An election survey is conducted. The aim is to estimate the percentage of people voting for the National Elderly Party (NEP). Three situations are considered:

- A simple random sample from the population
- A self-selection sample from the population
- A self-selection sample from the population, followed by poststratification
- A self-selection sample from the Internet population, followed by poststratification with a reference survey

In all four situations, the distribution of the estimator is determined by repeating the selection of the sample 1,000 times. The average participation probability in the population is 0.01971. Therefore, the expected sample size in a self-selection survey is equal to 1,971.

Figure 10.10 contains the results. The upper box plot shows that the estimator is unbiased in the case of simple random sampling (of size 1,971) from the target population. The expected value is equal to 25.6%.

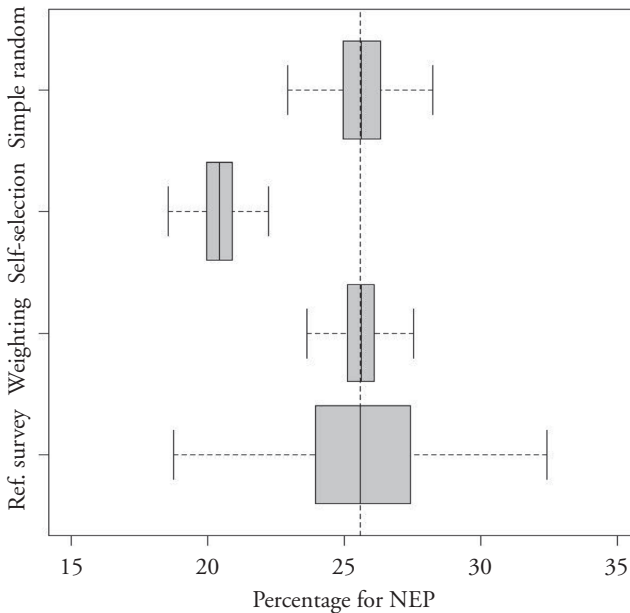


FIGURE 10.10 Estimating the percentage of voters for the NEP in the case of self-selection

The middle box plot shows what happens if samples are selected by means of self-selection. The shape of the distribution remains more or less the same, but the distribution as a whole has shifted to the left. All values of the estimator are systematically too low. The expected value of the estimator is only 20.4%. The estimator is biased. The explanation of this bias is simple: Relative few elderly are active Internet users. Therefore, they are underrepresented in the samples. They are typically people who will vote for the NEP.

The third box plot shows the distribution of the estimator in the case of poststratification by age. The bias is removed. This was possible because there is direct relation between participation and the weighting variable age.

The lower box plot shows the distribution of the estimator if the population distribution of the weighting variable age is estimated in the reference survey with a sample size of $m = 100$. The bias is removed but at the cost of a substantial increase of the variance. This is from the small sample size of the reference survey.

Figure 10.11 shows the analysis for voting for the New Internet Party (NIP). The same four situations are compared. In the case of simple random sampling, the estimator is unbiased. The population value of 39.5% is correctly estimated.

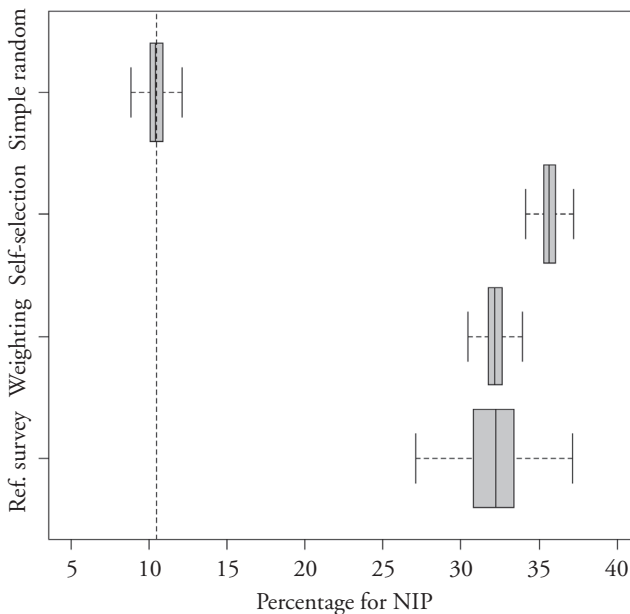


FIGURE 10.11 Estimating the percentage of voters for the NIP in the case of self-selection

Self-selection leads to a biased estimator. The estimated values are substantially too high. The expected value of the estimator is 56.5%, whereas it should have been 39.5%.

Poststratification is not successful. The expected value of the estimator decreases from 56.5% to 51.1%, but this is still too high. This is not surprising as there is a direct relation between voting for the NIP and having access to the Internet.

Poststratification based on a reference survey does not work here. The bias remains, and at the same time the variance increases.

10.3 Application

There are three nationwide public TV channels in the Netherlands. One of these channels (“Nederland 1”) has a current affairs program called “EenVandaag.” This program maintains a web panel. It is used to measure public opinion with respect to topics that are discussed in the program. The “EenVandaag Opinion Panel” started in 2004. In 2008, it contained approximately 45,000 members.

The panel is a self-selection panel. Participants were recruited among the viewers of the program. For these reasons, the panel lacks representativity. It is explored how unbalanced the composition of the panel is and whether estimates can be improved by applying some form of weighting adjustment.

In the period before the start of the Olympic Games in Beijing in August 2008, there was a lot of discussion in the Netherlands about a possible boycott of the games. Suggestions ranged from not showing up at the opening ceremony to athletes not participating in the games at all. This boycott was considered because of the lack of respect of the Chinese for the human rights of the Tibetan people. One wave of the opinion panel was conducted in April 2008 in order to determine the public opinion of the Dutch people with respect to this issue. The members of the panel were invited to complete a questionnaire. This questionnaire also contained topics about other issues, like preference for political parties. The questionnaire was completed by 19,392 members of the panel.

The representativity of the response is affected by two phenomena. In the first place, the panel was constructed by means of self-selection. In the second place, not all members of the panel responded to the request to fill in the questionnaire. The response rate was $100 \times 19,392 / 45,632 = 42.5\%$. Possible deviations from representativity are analyzed in this section. It is also explored to what extent weighting adjustment can improve the situation.

If persons apply for membership of the panel, they have to fill in a basic questionnaire with a number of demographic questions. These demographic variables can be used as auxiliary variables. The following variables were used in the analysis:

- Gender in two categories: male and female
- Age in five categories: 18–24, 25–39, 40–54, 55–64, and 65+

- Marital status in four categories: never married, married, divorced, and widowhood
- Province of residence in 12 categories: Groningen, Friesland, Drenthe, Overijssel, Flevoland, Gelderland, Noord-Holland, Zuid-Holland, Zeeland, Noord-Brabant, and Limburg
- Ethnic background in three categories: native, first-generation non-native, and second-generation non-native
- Voting at the 2006 general elections in 12 categories: CDA (Christian-democrats), PvdA (social-democrats), SP (socialists), VVD (liberals), PVV (right-wing populists), GroenLinks (green party), ChristenUnie (right-wing Christians), D66 (liberal-democrats), PvdD (party for the animals), SGP (right-wing Christians), other party, and did not vote.

The population distributions were available for all these variables. Most distributions could be found in Statline, the on-line statistical database of Statistics Netherlands (www.cbs.nl). The distribution of the voting variable came from a different source. For this reason, it was not possible to cross this variable with other auxiliary variables.

The first step in the analysis was to compare the response distribution of each variable with its population distribution. Table 10.9 contains the result for the variable Gender. It is clear that males are substantially overrepresented.

Table 10.10 compares the response distribution of the variable Age with its population distribution. Persons in the age group from 18 to 39 years are underrepresented. Also the elderly (65+) are somewhat underrepresented. People in the age group from 55 to 64 years are clearly overrepresented.

TABLE 10.9 The distribution of gender in the response and the population

Gender	Panel response	Population	Difference
Male	72.4%	49.0%	23.4%
Female	27.6%	51.0%	-23.4%
Total	100.0%	100.0%	

TABLE 10.10 The distribution of age in the response and the population

Age	Panel response	Population	Difference
18-24	5.5%	10.7%	-5.2%
25-39	15.5%	25.6%	-10.1%
40-54	29.0%	28.8%	0.3%
55-64	33.4%	16.2%	17.2%
65 +	16.5%	18.8%	-2.3%
Total	99.9%	100.1%	

Table 10.11 contains the comparison for the variable Marital status. People that never married and widowed persons are underrepresented. Married people are overrepresented. The patterns in Tables 10.10 and 10.11 may partly coincide. Young people are typically found in the category Never married. Widowed persons will often belong to the elderly.

Table 10.12 shows the regional distribution of panel response. Response percentages and population percentages are compared for the 12 Dutch provinces. The differences are small. One could conclude that the response is more or less representative with respect to the variable Province.

It should be noted that the pattern is usually different for the response to surveys based on probability sampling. Because of high nonresponse rates in highly urbanized areas, the response rates are usually low in the provinces of Utrecht, Noord-Holland, and Zuid-Holland. This is not the case for the EenVandaag Opinion Panel.

TABLE 10.11 The distribution of marital status in the response and the population

Marital status	Panel response	Population	Difference
Never married	26.9%	31.5%	-4.6%
Married	61.4%	53.6%	7.7%
Divorced	8.6%	6.8%	1.8%
Widowed	3.2%	8.1%	-4.9%
Total	100.1%	100.0%	

TABLE 10.12 The distribution of province of residence in the response and the population

Province	Panel response	Population	Difference
Groningen	3.6%	3.6%	0.0%
Friesland	3.5%	3.9%	-0.4%
Drenthe	3.0%	3.0%	0.0%
Overijssel	6.2%	6.7%	-0.5%
Flevoland	2.3%	3.9%	-1.6%
Gelderland	12.4%	12.0%	0.4%
Utrecht	8.1%	7.2%	0.8%
Noord-Holland	16.9%	16.1%	0.8%
Zuid-Holland	21.3%	21.1%	0.2%
Zeeland	2.2%	2.3%	-0.1%
Noord-Brabant	13.8%	14.8%	-1.0%
Limburg	6.4%	7.1%	-0.7%
Total	99.7%	101.7%	

Table 10.13 compares the response distribution and the population distribution of the variable Ethnic background. There is a problem with the first-generation non-natives. These persons and at least one of their parents are born outside the Netherlands. A language problem is one of the main causes for not participating in surveys. Note that the problem is less severe for second-generation non-natives. They are better integrated into the population.

Finally, Table 10.14 shows the difference between the response distribution and the population distribution for voting behavior in the 2006 general elections in the Netherlands.

The more traditional CDA (Christian-democrats) voters are underrepresented in the response, and the more activist SP (socialist party) voters are overrepresented. Also notable is the fact that apparently voters are substantially overrepresented among the respondents. This phenomenon also can be observed in other surveys. There is a relationship between voting and participating in surveys. Voters tend to participate in surveys, and nonvoters tend to refuse.

TABLE 10.13 The distribution of ethnic background in the response and the population

Marital status	Panel response	Population	Difference
Native	89.4%	81.2%	8.2%
Non-native (first generation)	3.4%	11.7%	-8.3%
Non-native (second generation)	7.2%	7.1%	0.1%
Total	100.0%	100.0%	

TABLE 10.14 The distribution of voting in 2006 in the response and the population

Party	Panel response	Population	Difference
CDA (Christian-democrats)	15.3%	21.3%	-6.0%
PvdA (social-democrats)	20.0%	17.0%	3.0%
SP (socialists)	20.2%	13.3%	6.9%
VVD (liberals)	14.6%	11.8%	2.8%
PVV (right-wing populists)	7.1%	4.7%	2.4%
GroenLinks (green party)	5.8%	3.7%	2.1%
ChristenUnie (right-wing Christians)	4.4%	3.2%	1.2%
D66 (liberal-democrats)	3.3%	1.6%	1.7%
PvdD (party for the animals)	2.1%	1.5%	0.6%
SGP (right-wing Christians)	0.3%	1.3%	-1.0%
Other party	1.6%	1.0%	0.6%
Did not vote	5.2%	19.8%	-14.6%
Total	99.9%	100.2%	

It is now explored how these auxiliary variables can be used for weighting adjustment. One target variable is selected here. It asks respondents whether they have a paid job for at least 12 hours per week. The response percentage is 48.6%. The population distribution is also available. The population percentage of people with a paid job is 57.6%. The response estimate is significantly too low. The question is whether this estimate can be improved by weighting.

As a first step, each auxiliary variable separately is used in poststratification weighting. The results are presented in Table 10.15.

The effects are small for most variables. There is almost no change in the estimate, and there is no reduction of the standard error. The effect is different for the variable Age. The estimate is adjusted in correct direction, from 48.6 to 50.9. The standard error is also smaller. This is an indication that Age should be included in the weighting model. Note also that the estimate is still far from the population value.

The next step is to explore whether the weighting model can be improved by adding another variable to Age. Table 10.16 contains the results for five

TABLE 10.15 Poststratification weighting with a single variable: Objective is estimation of the percentage with a paid job

Weighting model	Estimate	Standard error
No weighting	48.55	0.36
Age	50.92	0.29
Gender	47.44	0.36
Marital status	47.58	0.36
Province	48.45	0.36
Ethnic background	48.49	0.36
Vote in 2006	46.70	0.36
Population	57.62	

TABLE 10.16 Generalized regression estimation with two variables. Objective is estimation of the percentage with a paid job.

Weighting model	Estimate	Standard error
No weighting	48.55	0.36
Age + Gender	48.50	0.29
Age + Marital status	50.77	0.29
Age + Province	50.85	0.29
Age + Ethnic background	50.81	0.29
Age + Vote in 2006	50.66	0.29
Age \times Province	51.03	0.29
Population	57.62	

generalized regression models with only main effects. The interaction effects of variables are not included.

It is clear that the estimate cannot be improved by adding another variable to Age. The best weighting model is the one containing Age and Province, but it is not better than the model just containing the variable Age.

Table 10.16 contains also the estimates for the model obtained by crossing the variables Age and Province (Age \times Province). There is some improvement with respect to the model Age + Province. This is the best estimate that can be obtained with this set of auxiliary variables. The conclusion can be that weighting adjustment can reduce the bias of the estimate for the percentage of people with a paid job. Unfortunately, the bias cannot be removed. Apparently, the proper auxiliary variables are lacking.

Analysis of another target variable shows that not always the same auxiliary variables are effective in adjustment weighting. This time the target variable measures whether one intends to vote for the Socialist Party in the next elections. The percentage of voters in the response is 15.43%. There is no population value available, but other opinion polls suggest a percentage between 11% and 12%. This would indicate the response-based estimate is too high. Table 10.17 contains the results of various weighting adjustment attempts.

If poststratification is applied with just one variable, then there is no effect for five of the six weighting variables. The estimate changes only for the variable Vote 2006. The estimate goes down by 3% from 15.43% to 12.43%. Also the standard error of the estimate is smaller. This result is not surprising as one can expect there may be a relationship between current and past voting behavior.

Table 10.17 also shows that adding another auxiliary variable to the weighting model does not lead to substantial changes. By adding the variable Gender, the estimate is reduced from 12.43% to 12.30%. One can conclude that

TABLE 10.17 Poststratification weighting with a single variable

Weighting model	Estimate	Standard error
No weighting	15.43	0.28
Age	15.18	0.28
Gender	15.92	0.28
Marital status	15.59	0.28
Province	15.53	0.28
Ethnic background	15.82	0.28
Vote in 2006	12.43	0.21
Vote 2006 + Age	12.43	0.21
Vote 2006 + Gender	12.30	0.21
Vote 2006 + Marital status	12.43	0.21
Vote 2006 + Province	12.44	0.21
Vote 2006 + Ethnic background	12.67	0.21

this estimate is in line with the results of the other polls, although there is no guarantee that they reflect the true population value.

Note that it is not possible to cross the variable *Vote 2006* with other auxiliary variables. This is because *Vote 2006* is obtained from a different source: the Electoral Council of the Netherlands. This council does not record other variables like gender and age.

This application shows that weighting adjustment may help to reduce the bias of web-survey-based estimates. However, there is no guarantee that the bias is completely removed. The examples in this section also make clear that different target variables may need different weighting adjustment models.

10.4 Summary

There can be several reasons to carry some kind of weighting adjustment on the response to a web survey:

- The sample is selected with unequal probability sampling. To obtain unbiased estimates, an estimator such as the Horvitz–Thompson must be used. This comes down to weighting adjustment where the weights are equal to the design weights (which are equal to 1 over the first-order inclusion probabilities).
- Nonresponse may cause estimators of population characteristics to be biased. This happens when specific groups are over- or underrepresented in the survey response and these groups behave differently with respect to the survey variables.
- If the target population is wider than the Internet population, people without Internet can never be selected for the survey. This is called under-coverage, and it may lead to biased estimates.
- If the sample is selected by means of self-selection, the true selection probabilities are unknown, assuming equal selection probabilities leads to biased estimates.

Weighting adjustment techniques may help to reduce a bias. These techniques assign weights to observed elements, where underrepresented elements get a weight larger than 1 and overrepresented elements get a weight smaller than 1.

To be able to compute adjustment weights, auxiliary variables are required. Such variables have to be measured in the survey and the population distribution (or complete sample distribution) must be available.

Weighting adjustment will only help to reduce a bias if there is a strong relationship between the survey variables and the auxiliary variables and/or response behavior and auxiliary variables.

Poststratification is the most frequently used weighting adjustment technique. Using auxiliary variables, the population is divided into several strata (subpopulations). All observed elements in a stratum are assigned the same

weight. Poststratification reduces a bias if the strata are homogeneous (i.e., all elements within a stratum resemble each other).

Practical limitations may hinder application of poststratification. If many auxiliary variables are used to form strata, there may be strata without observations. Consequently it is impossible to compute weights for such strata. It can also happen that insufficient population is available with respect to the distribution of the auxiliary variables.

If it is not possible to carry out poststratification, there are two alternative weighting methods. One is generalized regression estimation. It is based on a linear regression model that predicts the values of the target variable of the survey from a set of auxiliary variables. Such regression models offer more flexibility with respect to the way auxiliary information is used to compute adjustment weights. Another possibility is to use raking ratio estimation. This weighting model is based on iterative proportional fitting. It can be shown that estimates based on generalized regression estimation and raking ratio estimation behave approximately the same in many situations.

Calibration is a theoretical framework for adjustment weighting. Poststratification, linear weighting, and multiplicative weighting are special cases for this framework. It has more possibilities, like imposing constraints on the values of the weights.

If auxiliary variables cannot be used because their population distribution is not available, one might consider estimating them with a reference survey. This is a survey not affected by the problems of web surveys. It might be a CAPI or CATI survey. This approach can be effective in reducing or removing a bias, but the price to be paid is a substantial increase of the variances of estimates.

KEY TERMS

Auxiliary variable: A variable that has been measured in the survey and for which the distribution in the population (or the complete sample) is available.

First-order inclusion probability: The probability that a population element is selected in the sample. The first-order inclusion probability is determined by the sampling design.

Generalized regression estimation: A weighting technique that computes weights using a generalized linear regression model that predicts the target variable of the survey from a set of auxiliary variables. This is sometimes also called linear weighting.

Homogeneous: A stratum (subpopulation) is called homogeneous if all its elements resemble each other with respect to the target variables of the survey.

Internet population: The subpopulation of the target population consisting of only elements that have access to the Internet.

Nonresponse: The phenomenon that elements in the selected sample, which are also eligible for the survey, do not provide the requested information or that the provided information is not usable.

Poststratification: A weighting method that divides the population in strata and subsequently assigns the same weight to all observed elements within a stratum.

Raking ratio estimation: A weighting technique that computes weights using an iterative fitting procedure that adjusts the weight so that the weighted sample distributions of the auxiliary variables fit their population distributions. This technique is also known as iterative proportional fitting or multiplicative weighting.

Reference survey: A survey conducted with the objective to obtain unbiased estimates of the population distributions of auxiliary variables.

Representative: The (weighted) survey response is representative with respect to a variable if the (weighted) response distribution is equal to its population distribution.

Self-selection survey: A survey for which the sample has been recruited by means of self-selection. It is left to the persons themselves to decide to participate in a survey. No probability sample is selected.

Stratification: A division of the population into several subpopulations (strata) by cross-classifying many auxiliary variables.

Undercoverage: The sampling frame does not cover completely the target population of the survey. There are persons in the population who do not appear in the sampling frame. They will never be selected in the sample.

EXERCISES

Exercise 10.1. Which property of an auxiliary variable makes it useful for including in a weighting adjustment model?

- a. The response distribution of the variable is approximately equal to its population distribution.
- b. The sample distribution of the variable is approximately equal to its population distribution.
- c. The response distribution of the variable differs considerably from its sample distribution.
- d. The response distribution of the variable is approximately equal to its sample distribution.

Exercise 10.2. A large company has 2,500 employees. The management has installed coffee machines everywhere in the building. After a while, the management wants to know whether the employees are satisfied with the coffee machines. It is decided to conduct a web survey. A simple random sample without replacement of 500 employees is drawn. It turns out that 380 employees complete the web questionnaire form. Of those, 310 are satisfied with the coffee machines.

- a. Compute the 95% confidence interval of the percentage of employees in the company who are satisfied with the coffee machines.

Only 380 of the 500 selected employees responded. So there is a nonresponse problem.

- b. Compute a lower bound and an upper bound for the percentage of employees in the sample who are satisfied with the coffee machines.

Previous research has showed that employees with a higher level of education are less satisfied with the coffee facilities. The management knows the level of education of each employee in the company: 21% has a high education and 79% has a low education. The table below shows the relationship between coffee machine satisfaction and level of education for the 380 respondents:

	Low education	High education	Total
Satisfied	306	4	310
Not satisfied	40	30	70
Total	346	34	380

A weighting adjustment procedure is carried out to reduce the nonresponse bias.

- c. Compute weights for low and high educated employees.
- d. Compute the weighted estimate of the percentage of employees in the company satisfied with the coffee facilities.

Exercise 10.3. There are plans in the Netherlands to introduce a system of road pricing. It means car drivers are charged for the roads they use. Such a system could lead to better use of the available road capacity and, therefore, could reduce traffic congestion. An Automobile Association wants to know what the attitude of the Dutch people is toward road pricing. It conducts a web survey. People are asked two questions:

- Are you in favor of road pricing?
- Do you have a car?

The results are summarized below:

	In favor of road pricing?	
	Yes	No
Has a car?		
Yes	128	512
No	60	40

- a. Using the available data, and assuming simple random sampling, estimate the percentage in favor of road pricing.
- b. From another source, it is known that 80% of the target population owns a car and that 20% does not have one. Use this additional information to apply weighting adjustment. Compute a weight for car owners and a weight for those without a car.
- c. Make a table like the one above but with weighted frequencies.
- d. Compute a weighted estimate for the percentage in favor of road pricing.
- e. Explain the difference between the weighted and the unweighted estimate.

Exercise 10.4. A transport company carries out a web survey to determine how healthy its truck drivers are. Only 21 drivers complete the web questionnaire form. Each respondent was asked whether he or she has visited a doctor because of medical problems. Also the experience of the driver (little, much) and age (young, middle, old) were recorded. The results are in the table below:

No.	Age	Experience	Doctor visits
1	Young	Much	2
2	Young	Much	3
3	Young	Much	4
4	Young	Little	3
5	Young	Little	4
6	Young	Little	4
7	Young	Little	5
8	Middle	Much	5
9	Middle	Much	6
10	Middle	Much	7
11	Middle	Little	5
12	Middle	Little	6
13	Middle	Little	6
14	Middle	Little	7
15	Old	Much	8
16	Old	Much	10
17	Old	Much	10
18	Old	Much	8
19	Old	Little	8
20	Old	Little	9
21	Old	Little	10

- a. Estimate the average number of doctor visits, assuming the response can be seen as a simple random sample.

- b. Assume the population distributions of experience and age are available for the population of all drivers of the company:

Experience	Percentage	Age	Percentage
Much	48%	Young	22%
Little	52%	Middle	30%
		Old	48%

Establish whether the response is selective. Explain which of these two auxiliary variables should be preferred for computing adjustment weights.

- c. For each auxiliary variable separately, carry out weighting adjustment. Compute weights for each of the categories of the auxiliary variable.
- d. Compute for both weighting adjustments a weighted estimate of the average number of doctor visits.
- e. Compare the outcomes under 10.4.a and 10.4.d. Explain the differences and/or similarities.

REFERENCES

- Bethlehem, J. G. (1988), Reduction of Nonresponse Bias through Regression Estimation. *Journal of Official Statistics*, 4, pp. 251–260.
- Bethlehem, J. G. (2002), Weighting Nonresponse Adjustments Based on Auxiliary Information. In: Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (eds.), *Survey Nonresponse*. John Wiley & Sons, New York.
- Bethlehem, J. G. (2007), *Reducing the Bias of Web Survey Based Estimates*. Discussion Paper 07001. Statistics Netherlands, Voorburg/Heerlen, the Netherlands.
- Bethlehem, J. G. (2009), *Applied Survey Methods, A Statistical Perspective*. John Wiley & Sons, Hoboken, NJ.
- Bethlehem, J. G. & Keller, W. J. (1987), Linear Weighting of Sample Survey Data. *Journal of Official Statistics*, 3, pp. 141–154.
- Börsch-Supan, A., Elsner, D., Faßbender, H., Kiefer, R., McFadden, D., & Winter, J. (2004), *Correcting the Participation Bias in an Online Survey*. Report, University of Munich, Germany.
- Deming, W. E. & Stephan, F. F. (1940), On a Least Squares of Adjustment of a Sampled Frequency Table when the Expected Totals are Known. *Annals of Mathematical Statistics*, 11, pp. 427–444.
- Deville, J. C. & Särndal, C. E. (1992), Calibration Estimation in Survey Sampling. *Journal of the American Statistical Association*, 87, pp. 376–382.
- Deville, J. C., Särndal, C. E., & Sautory, O. (1993), Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88, pp. 1013–1020.
- Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005), Comparing Data from Online and Face-to-Face Surveys. *International Journal of Market Research*, 47, pp. 615–639.

- Holt, D. & Smith, T. M. F. (1979), Post Stratification. *Journal of the Royal Statistical Society*, Series A, 142, pp. 33–46.
- Horvitz, D. G. & Thompson, D. J. (1952), A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, pp. 663–685.
- Huang, E. T. & Fuller, W. A. (1978), Nonnegative Regression Estimation for Survey Data. *Proceedings of the Social Statistics Section of the American Statistical Association*, Washington, DC, pp. 300–303.
- Kruskal, W. & Mosteller, F. (1979a), Representative Sampling, I: Non-scientific Literature. *International Statistical Review*, 47, pp. 13–24.
- Kruskal, W. & Mosteller, F. (1979b), Representative Sampling, II: Scientific Literature. Excluding Statistics, *International Statistical Review*, 47, pp. 111–127.
- Kruskal, W. & Mosteller, F. (1979c), Representative Sampling, III: The Current Statistical Literature. *International Statistical Review*, 47, pp. 245–265.
- Little, R. J. A. & Wu, M. M. (1991), Models for Contingency Tables with Known Margins when Target and Sampled Populations Differ. *Journal of the American Statistical Association*, 86, pp. 87–95.
- Skinner, C. J. (1991), On the Efficiency of Raking Ratio Estimates for Multiple Frame Surveys. *Journal of the American Statistical Association*, 86, pp. 779–784.

Use of Response Propensities

11.1 Introduction

Some of the main problems of web surveys are caused by undercoverage and nonparticipation. Both phenomena may lead to biased estimators of population characteristics, and therefore, wrong conclusions are drawn from the web survey data. To avoid this, some kind of correction technique is required. Weighting adjustment is one such technique. This topic is treated in Chapter 10. It is also possible to use response propensities to correct biased estimates. This chapter is about response propensities, and it describes what they are, how they can be computed, and what can be done with them.

The problems described above may occur if a general-population survey is conducted using Internet data collection. The target population is usually wider than just those with access to the Internet. This implies that individuals without Internet access cannot be selected for the survey. The sample for a general-population web survey is selected from a sampling frame. Such a frame does not contain information about which people have Internet access and which do not. It will also not contain the e-mail addresses of those having Internet access. The sample is usually selected using a different mode. For example, selected individuals are sent a letter with an invitation to complete the web questionnaire. Those without Internet access will not respond. For those with Internet access, it is up to them whether they will respond. In conclusion, the ultimate group of respondents is the result of a selection process with unknown probabilities.

Even if the target population coincides with all individuals having Internet access, the problems are not solved. Selecting a proper probability sample requires a sampling frame containing the e-mail address of all individuals in the

population. Such sampling frames do not exist, and probably they will never exist. The way out is, again, to use a different mode to recruit the sample. Selected individuals can be sent an invitation letter, or they can be called by telephone. It is up to them whether they will respond. Again, the response may be considered the result of a probability mechanism with unknown probabilities.

In any case, if the target population is the general population, and only Internet users can complete the survey questionnaire, the question originates of whether estimates of population characteristics are biased and, if so, of whether these estimates can be corrected so that the bias is reduced or eliminated.

Summing up, web-survey-based estimates may be biased because of undercoverage or nonresponse. Although such a selection bias can have various causes, the methodological consequences are similar. Therefore, the treatment of these problems is also more or less the same.

Various methods have been proposed in the literature to deal with selection bias. Adjustment weighting is one of them (see Chapter 10). Use of response propensity is another. Although this approach was developed already many years ago, it witnessed a revival with the emergence of web surveys.

An important problem of causal inference is how to estimate treatment effects in observational studies. This is a situation (like in an experiment) in which a group of objects is exposed to a well-defined treatment, and another group (the control group) does not receive this treatment. However, it is not a controlled experiment. Therefore, observed effects can be biased. They may partially be artifacts caused by the way the treatment group and the control are composed.

So-called *propensity score matching* methods can be used to correct for a sample selection bias because of observable differences between the treatment and the control group. Matching involves pairing objects in the treatment group in and the control group that are similar in terms of their observable characteristics. When the relevant differences between any two objects are captured in observable covariates, which occurs when outcomes are independent of assignment to treatment conditional on the covariates, the matching methods can yield an unbiased estimate of the treatment effect.

The propensity score matching techniques was developed in the 1980s in the context of biomedical studies (Rosenbaum and Rubin, 1983), and its original conceptual framework is described in the study by Rubin (1974).

The method of propensity score matching can only be applied if certain conditions are satisfied. Nevertheless, the method is widely applied, not only in biomedical analyses, but in many other research fields like labor market research and policy evaluation. See, for example, the study by Dehejia and Wahba (1999). The basic assumptions underlying the method are as follows:

- Selection in the treatment group (or control group) can be explained purely in terms of observable characteristics. If one can control for observable differences in characteristics between the treatment and the control group, the observed effect can be attributed to the treatment. It is a true treatment effect. This assumption is called the *conditional independence assumption* or

the *strong ignorability assumption*. If the conditional independence assumption holds, the matching process is analogous to creating an experimental data set where, conditional on observed characteristics, the selection process is random.

- The decision to assign the treatment to an object does not depend on the decision to assign the treatment to other objects.
- The observed outcome of a variable for an object depends only on the object itself and not on the mechanism assigning treatment to objects.

Application of the method of propensity scores requires the individual values of all variables explaining the group differences to be available. In particular, if the two groups are the participants and the nonparticipants, it may be difficult to collect all these values for the nonparticipants. This restriction may prevent application. Moreover, all values must not have been affected by measurement errors.

In this chapter, the method of propensity score matching is applied to web surveys. This is only possible if the conditions stated above are satisfied. In addition, the web survey situation should be put in the proper theoretical framework. This means

- Identifying what the treatment is.
- Identifying the treatment group and the control group.
- Selecting the observable variables to be included in the matching process.
- Assessing of the conditional independence assumption.

As mentioned, application of propensity scores to the web survey situation may be hampered by the availability of proper data.

The idea of applying the propensity score method in survey methodology was introduced by Harris Interactive. See, for example, the study by Taylor, Overmeyer, Siegel, and Terhanian (2001). Harris Interactive used the propensity score method to solve the problems of undercoverage and self-selection in Internet panels. Terhanian, Smith, Bremer, and Thomas (2001) proposed the use of this technique as a tool for weighting self-selection samples of web respondents. The idea is to define the propensity score as the probability that an object selects itself for the web survey. Consequently, the term *response propensity* will be used in this context. The treatment group consists of all objects in the sample, and the control group consists of all objects not in the sample.

The main idea is now that within a group of objects with the same response propensities, there are no selection effects. So within-group estimates based on sample elements from this group will be unbiased. By combining the group estimates into a population estimate (taking into account group sizes), an unbiased population estimate is obtained.

There is a growing literature on experiences with the use of the response propensities method in web surveys. The approach seems promising. Nevertheless, further theoretical and empirical work is required to prepare it for regular application in web survey and web panel environments. There are still some

unanswered questions. There are, for example, studies that criticize the application of propensity matching and show that its application critically depends on assumptions about the nature of the process by which participants select themselves, and on the data available to the survey researcher. See the studies by Smith and Todd (2000) and by Heckman, Ichimura, Smith, and Todd (1998).

■ EXAMPLE 11.1 Response propensities using a reference survey

Schonlau et al. (2004) and Lee (2006) applied the propensity score method in web surveys. To obtain groups that are similar with respect to the response propensities, they divided the sample into strata using variables that explain the response behavior. Such variables can be measured in the survey, but their values are not available for individuals not participating in the survey.

To solve this problem, they conducted another survey that does not suffer from undercoverage and self-selection. Moreover, nonresponse in this survey is considered Missing Completely At Random (MCAR). The idea is to measure just the variables that are required to estimate response propensities. Such a survey is called here a *reference survey*.

This reference survey can be seen as a benchmark for the web survey participants by balancing the distribution of these variables for the web respondents so that it becomes similar to its distribution for the reference survey respondents.

Duffy, Smith, Terhanian, and Bremer (2005) used behavioral, attitudinal, and sociodemographic variables for this purpose. These variables are sometimes called *webographic* or *psychographic variables*. Examples of questions measuring this kind of variables are “How often do you watch tv programs alone?” and “Do you often feel unhappy?”

The response propensity is defined as the conditional probability that a sample element responds in the web survey, given the values of the explanatory variables. The estimated values of the response propensities are used to construct groups with (approximately) the same scores. Stratification based on strata corresponding to these groups will remove the selection bias provided all conditions underlying the group are satisfied.

Bethlehem (2007) showed that use of reference surveys for propensity score matching is not without problems. The bias may be reduced but at the cost of a large increase in variance. In addition, it is not realistic to assume that the reference survey does not experience problems with undercoverage or nonresponse. Furthermore, the reference survey will probably be a computer-assisted personal interviewing (CAPI) or computer-assisted telephone interviewing (CATI) survey, whereas the main survey is a web survey. So there may be also mode effects. Use of a reference survey is treated in some more detail in Chapter 10.

This chapter discusses various problems and ways to use response propensities to improve estimators in web surveys. A definition of response propensity is given, and models for response propensities are described. Methods for correcting bias using response propensities are illustrated. In particular, response propensity weighting and stratification are proposed as correction techniques. At the end, an application comparing methods for correcting bias is described.

11.2 Theory

Let the target population U of the survey consist of N identifiable elements, which are labeled $1, 2, \dots, N$. Therefore, the target population can be denoted by

$$(11.1) \quad U = \{1, 2, \dots, N\}.$$

Associated with each element k is a value Y_k of the target variable Y . The aim of the web survey is assumed to be estimation of the population mean

$$(11.2) \quad \bar{Y} = \frac{1}{N} \sum_{k=1}^N Y_k$$

of the target variable Y .

Two cases will be discussed here, in which response propensities can be used in an attempt to reduce a selection bias. The first case is that of a simple random sample in which nonresponse occurs. The second case is that of a self-selection survey.

It will be assumed that every individual in the population has Internet access. This is an ideal situation. In practice, this will usually not be the case, which introduces an extra bias resulting from undercoverage.

11.2.1 A SIMPLE RANDOM SAMPLE WITH NONRESPONSE

Suppose a simple random sample is selected without replacement from the population U . It means that each element can appear at most once in the sample. Therefore, the sample can be represented by a series of indicators

$$(11.3) \quad a = a_1, a_2, \dots, a_N.$$

The k th indicator a_k assumes the value 1 if element k is selected in the sample, and otherwise it assumes the value 0. The expected value (i.e., the mean value over all possible samples) of a_k is denoted by

$$(11.4) \quad \pi_k = E(a_k).$$

The quantity π_k is the *first-order inclusion probability* of element k (for $k = 1, 2, \dots, N$). In a simple random sample, all first-order inclusion probabilities are equal to $\pi_k = n/N$, where n is the sample size, which can be written as

$$(11.5) \quad n = \sum_{k=1}^N a_k.$$

The Horvitz–Thompson estimator (Horvitz and Thompson, 1952) is defined by

$$(11.6) \quad \bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^N a_k \frac{Y_k}{\pi_k}.$$

This is an unbiased estimator of the population mean. Note that for a simple random sample, the Horvitz–Thompson estimator reduces to the simple sample mean

$$(11.7) \quad \bar{y} = \frac{1}{n} \sum_{k=1}^N a_k Y_k.$$

The Horvitz–Thompson estimator can only be applied if every element in the sample responds. Unfortunately, this is not the case in practical situations. There is always nonresponse. To investigate the consequences of nonresponse, it is assumed that the *random response model* applies. This model assumes every element k in the population to have an (unknown) response probability ρ_k . If element k is selected in the sample, a random mechanism is activated that results in probability ρ_k in response and in probability $1 - \rho_k$ in nonresponse. Under this model, a set of response indicators

$$(11.8) \quad R_1, R_2, \dots, R_N$$

is introduced, where $R_k = 1$ if the corresponding element k responds, and where $R_k = 0$ otherwise. So, $P(R_k = 1) = \rho_k$, and $P(R_k = 0) = 1 - \rho_k$.

Now suppose a simple random sample without replacement of size n is selected from this population. The response only consists of those elements k for which $a_k = 1$ and $R_k = 1$. Hence, the number of available cases is equal to

$$(11.9) \quad n_R = \sum_{k=1}^N a_k R_k.$$

Note that this realized sample size is a random variable. The number of non-respondents is equal to

$$(11.10) \quad n_{NR} = \sum_{k=1}^N a_k (1 - R_k),$$

where $n = n_R + n_{NR}$.

The values of the target variable become only available for the n_R responding elements. The mean of these values is denoted by

$$(11.11) \quad \bar{y}_R = \frac{1}{n_R} \sum_{k=1}^N a_k R_k Y_k.$$

It can be shown (see the study by Bethlehem (2009)) that the expected value of the response mean is approximately equal to

$$(11.12) \quad E(\bar{y}_R) \approx \tilde{Y},$$

where

$$(11.13) \quad \tilde{Y} = \frac{1}{N} \sum_{k=1}^N \frac{\rho_k}{\bar{\rho}} Y_k$$

and

$$(11.14) \quad \bar{\rho} = \frac{1}{N} \sum_{k=1}^N \rho_k$$

is the mean of all response probabilities in the population. From expression (11.12), it is clear that, generally, the expected value of the response mean is unequal to the population mean to be estimated. Therefore, this estimator is biased. This bias is approximately equal to

$$(11.15) \quad B(\bar{y}_R) = \tilde{Y} - \bar{Y} = \frac{S_{\rho Y}}{\bar{\rho}} = \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}},$$

where $S_{\rho Y}$ is the covariance between the values of the target variable and the response probabilities, $R_{\rho Y}$ is the corresponding correlation coefficient, S_Y is the standard deviation of the variable Y , and S_{ρ} is the standard deviation of the response probabilities.

The bias of the estimator can be repaired by introducing the modified Horvitz–Thompson estimator

$$(11.16) \quad \bar{y}_{HT}^* = \frac{1}{N} \sum_{k=1}^N a_k R_k \frac{Y_k}{\pi_k \rho_k}.$$

Because $E(a_k R_k) = \pi_k \rho_k$, this is an unbiased estimator. The problem is, however, that this estimator cannot be computed because the response probabilities are unknown. This can be solved by first estimating the response probabilities and then substituting these estimates in expression (11.6). The method of

response propensities can be used for this purpose. For selection bias in web surveys the study by see also Biffignandi and Bethlehem (2011).

11.2.2 A SELF-SELECTION SAMPLE

Participation in a self-selection web surveys requires that respondents are aware of the existence of the web survey (they have to visit the website accidentally, or they have to follow up a banner or an e-mail message). They also have to decide to fill in the questionnaire on the Internet. This means that each element k in the Internet population has an unknown probability ρ_k of participating in the survey, for $k = 1, 2, \dots, N$.

The responding elements are denoted by a set

$$(11.17) \quad R_1, R_2, \dots, R_N$$

of N indicators, where the k th indicator R_k assumes the value 1 if element k participates, and otherwise it assumes the value 0, for $k = 1, 2, \dots, N$. The expected value $\rho_k = E(R_k)$ can be called the *response probability* of element k . The realized sample size is denoted by

$$(11.18) \quad n_S = \sum_{k=1}^N R_k.$$

Lacking any knowledge about the values of the response probabilities, survey researchers usually implicitly assume all these probabilities to be equal. In other words, simple random sampling is assumed. Consequently, the sample mean

$$(11.19) \quad \bar{y}_S = \frac{1}{n_S} \sum_{k=1}^N R_k Y_k$$

is used as an estimator for the population mean. The expected value of this estimator is approximately equal to

$$(11.20) \quad E(\bar{y}_S) \approx \tilde{Y} = \frac{1}{N\bar{\rho}} \sum_{k=1}^N \rho_k Y_k,$$

where $\bar{\rho}$ is the mean of all response probabilities in the population. This expression was derived by Bethlehem (1988).

It is clear from expression (11.20) that, generally, the expected value of the sample mean is not equal to the population mean of the population. One situation in which the bias vanishes is that in which all response probabilities in the population are equal. In terms of nonresponse correction theory, this comes down to MCAR. This is the situation in which the cause of missing data is completely independent of all variables measured in the survey. For more

information on MCAR and other missing data mechanisms, see the study by Little and Rubin (2002). Indeed, in the case of MCAR, self-selection does not lead to an unrepresentative sample because all elements have the same selection probability. Bethlehem (2002) shows that the bias of the sample mean (11.20) can be written as

$$(11.21) \quad B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y} \approx \tilde{Y} - \bar{Y} = \frac{C_{\rho Y}}{\bar{\rho}} = \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}},$$

in which

$$(11.22) \quad C_{\rho Y} = \frac{1}{N} \sum_{k=1}^N (\rho_k - \bar{\rho})(Y_k - \bar{Y}).$$

The bias can be repaired by introducing a modified estimator defined by

$$(11.23) \quad \bar{y}_S^* = \frac{\bar{\rho}}{n_S} \sum_{k=1}^N \frac{R_k Y_k}{\rho_k}.$$

This estimator is approximately unbiased. Note that the mean response probability can simply be estimated by n_S/N , changing estimator (11.23) into

$$(11.24) \quad \bar{y}_S^* = \frac{1}{N} \sum_{k=1}^N \frac{R_k Y_k}{\rho_k}.$$

Again, the problem is that the response probabilities are unknown. However, they can be estimated using the method of response propensities.

11.2.3 THE RESPONSE PROPENSITY DEFINITION

Let $U = \{1, 2, \dots, N\}$ be the target population of the survey. Associated with each element k is a value Y_k of the target variable Y . Furthermore, there is a set of auxiliary variables. For every element k , the vector of values of these auxiliary variables is denoted by $X_k = (X_{k1}, X_{k2}, \dots, X_{kp})'$. It is assumed that these values are known for all elements in the sample (in the case of a random sample) or for all elements in the population (in the case of a self-selection survey). Such information can, for instance, be found in a population register.

It is assumed that every element k in the population has a nonzero, unknown response probability, denoted by ρ_k . If element k is selected in a sample, a random mechanism is activated that results in probability ρ_k in response and in probability $(1 - \rho_k)$ in nonresponse. In the case of a self-selection survey, there is no sample selection. The whole population could be considered the sample. Also, here ρ_k is the probability to respond.

The response probability ρ is a latent variable; it is not observed. Instead the corresponding response indicator R is observed. The response indicator R_k equals

1 if element k responds, and otherwise it is 0. A vector R of response indicators can be introduced (i.e., $R = (R_1, R_2, \dots, R_N)'$, where $P(R_k = 1) = \rho_k$ and $P(R_k = 0) = 1 - \rho_k$). These probabilities can be estimated using an appropriate model based on auxiliary information.

The first step is to introduce the *response propensity* $\rho(X_k)$. It is defined by

$$(11.25) \quad \rho_k(X) = P(R_k = 1 | X = X_k).$$

It can be interpreted as the probability of response given the values of the set of auxiliary variables X . It is assumed that all auxiliary variables required to explain the response behavior are included in the set X . To say it otherwise, given the values of the auxiliary variables, response behavior is independent of the target variables of the surveys. This comes down to the assumption of *Missing At Random* (MAR). This assumption is also known as the *conditional independence* assumption (Lechner, 1999), *selection on observables* (Barnow, Cain, and Goldberger, 1980), *unconfoundedness* assumption or *ignorable treatment* assumption (Rosenbaum and Rubin, 1983), and *exogeneity* (Imbens, 2004).

The definition of the concept of response probability is not straightforward. It involves at some stage a decision on how to deal with the dependence of the response probabilities on the circumstances under which the survey is being conducted. They may, for example, depend on the number and timing of contact attempts and on the interviewer characteristics. If these circumstances change, it is very likely that the individual response probabilities also change.

In addition, response probabilities may vary over time. However, the more conditions are imposed on the response probabilities, the more fixed they become. The fixed-response model develops as a special case of the random response model when the response probabilities are viewed conditionally on very detailed circumstances. No variation is left, and response becomes deterministic.

11.2.4 MODELS FOR RESPONSE PROPENSITIES

To be able to estimate the response propensities, a model must be chosen. The most frequently used one is the *logistic regression model*. It assumes the relationship between response propensity and auxiliary variables can be written as

$$(11.26) \quad \log \text{it}(\rho_k(X)) = \log \left(\frac{\rho_k(X)}{1 - \rho_k(X)} \right) = \sum_{j=1}^p X_{kj} \beta_j,$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ is a vector of p regression coefficients. The *logit* transformation ensures that estimated response propensities are always in the interval $[0, 1]$.

Another model sometimes used is the *probit model*. It assumes the relationship between response propensity and auxiliary variables can be written as

$$(11.27) \quad \text{probit}(\rho_k(X)) = \Phi^{-1}(\rho_k(X)) = \sum_{j=1}^p X_{kj}\beta_j,$$

in which Φ^{-1} is the inverse of the standard normal distribution function. Both models are special cases of the generalized linear model (GLM)

$$(11.28) \quad g(\rho_k(X)) = \sum_{j=1}^p X_{kj}\beta_j,$$

where g is called the *link function* that has to be specified. It is also possible to use the *identity* link function. This means the relationship between response propensity and auxiliary variables can be written as

$$(11.29) \quad \rho_k(X) = \sum_{j=1}^p X_{kj}\beta_j.$$

This is a simple linear model. It has advantages and disadvantages. A first advantage of the linear model is that coefficients are much easier to interpret. They simply represent the effects of the auxiliary variables on the response propensity. These effects are “pure” effects. The coefficient of a variable is corrected for the interdependencies of the other auxiliary variables in the model. Interpretation of a logit or probit model is not so straightforward. The logit or probit transform hampers the interpretation of the linear parameters.

A second advantage is that the computations are simpler. Estimates of the coefficients can be obtained by ordinary least squares. Estimation of the logit and probit model requires maximum likelihood estimation.

An advantage of the probit and logit model is that estimated response propensities are always in the interval $[0, 1]$. The linear model does not prevent estimated probabilities to be negative or larger than 1. However, according to Keller, Verbeek, and Bethlehem (1984), the probability of estimates outside the interval $[0, 1]$ vanishes asymptotically if the model is correct and all response probabilities are strictly positive.

It should be noted that the linear model is not necessarily a worse approximation of reality than the probit or the logit model. The logit and probit transformations were introduced for convenience only, and not because their models were “more likely.” Dehija and Wabha (1999) conclude that the choice of model does not influence the results very strongly.

Figure 11.1 contains the graphs of the logit and the probit function. It can be observed that both functions are more or less linear for values of p between, say, 0.2 and 0.8. So, the linear link functions can be considered an approximation of the other two link functions.

The logit model is the most commonly used model for estimating response propensities. It will also be applied in this chapter. Rewriting model (11.26) leads to the expression

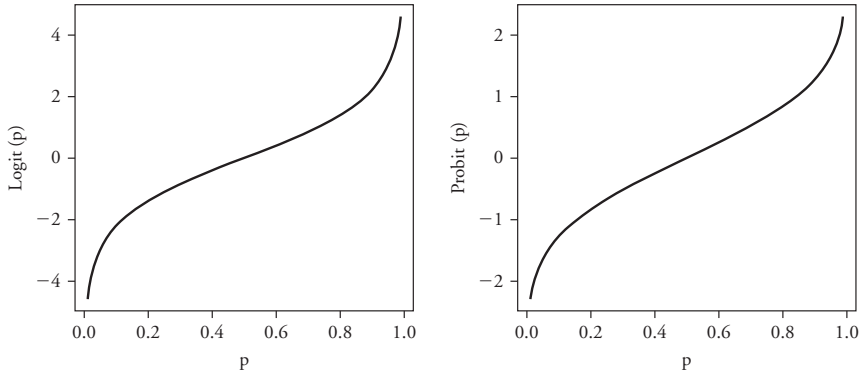


FIGURE 11.1 The logit and probit link functions

$$(11.30) \quad \rho(X_k) = \frac{\exp(X'_k \beta)}{1 + \exp(X'_k \beta)}$$

for the response propensities. The response propensities have to be estimated using the available data. Estimation is only possible if there are both respondent and nonrespondents for each set of values of X . This comes down to what is called the *matching assumption*. It states that

$$(11.31) \quad 0 < \rho(X_k) < 1.$$

This assumption ensures that for each value of X there are elements that participate in the web surveys and elements that do not participate in the web survey. Note that elements with a response probability equal to 0 or 1 cannot be compared because for these elements there are no counterparts. This is not important for elements k with $\rho(X_k) = 1$ because they are all observed in the sample. It is a problem for elements k with $\rho(X_k) = 0$ because they are never observed, and thus, they may be the cause of a bias.

Estimation also requires the individual values of the auxiliary variables to be known for the nonrespondents (in the case of a probability sample with non-response) or for the nonparticipants (in the case of a self-selection survey). This is often not the case. Such values may be available if the sample is selected from a sampling frame that contains all relevant variables.

EXAMPLE 11.2 Estimating response propensities

Response propensities are estimated for a real survey data set. This data set is based on a Dutch survey that has been carried out by Statistics Netherlands. It is called the General Population Survey (GPS). The sampling frame was the population register in the Netherlands. The

sampling design was such that each person had the same probability of being selected (a so-called self-weighting sample). The sample of the GPS consisted of 32,019 persons. The number of respondents was 18,792.

To find the model for the response propensities, auxiliary variables are required. Statistics Netherlands has an integrated system of social statistics. This system is called the *Social Statistics Database* (SSD). The SSD contains a wide range of characteristics on each individual in the Netherlands. There are data on demography, geography, income, labor, education, health, and social protection. These data are obtained by combining data from registers and other administrative data sources. By linking the sample file of the GPS to the SSD, the values of a large set of auxiliary variable become available for both respondents and non-respondents. Table 11.1 contains the subset of variables that turned out to have a significant contribution in the response propensity model.

TABLE 11.1 Auxiliary variables in the response propensity model

Variable	Description	Categories
Gender	Gender	2
Married	Is married	2
Age13	Age (in 13 age groups)	13
Ethnic	Type of non-native	5
HHSize	Size of the household	5
HHType	Type of household	5
Phone	Has listed phone number	2
Hasjob	Has a job	2
Region	Region of the country	5
Urban	Degree of urbanization	5

Note that all variables in this table are categorical variables. To include them in model (11.26), they have to be replaced by sets of dummy variables, where there is a dummy variable for each category of each variable. Furthermore, to be able to estimate the model parameters, extra restrictions must be imposed. This is usually accomplished by setting one of the parameters for each set of dummies to 0.

A logit model has been fitted with the variables in Table 11.1 as explanatory variables. Just the main effects were included in the model, no interaction effects. The estimated model was used to estimate the response propensity for each sample person. This distribution of these probabilities is displayed in Figure 11.2

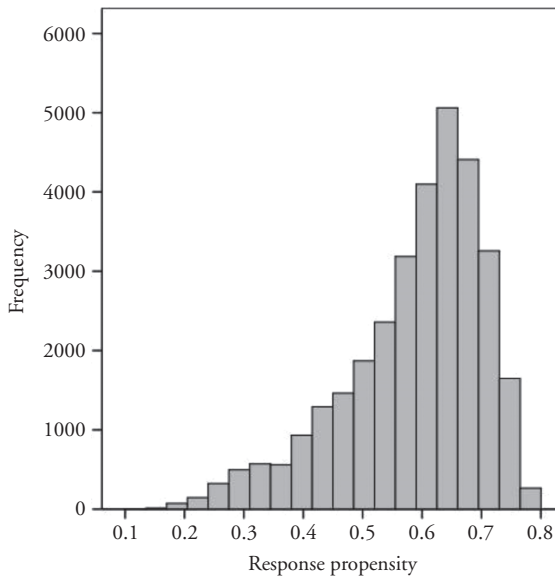


FIGURE 11.2 The distribution of the estimated response propensities

There is a wide variation in these probabilities. The values of the response propensities range roughly between 0.1 and 0.8. The average value is 0.587. This is equal to the response rate. Table 11.2 contains the characteristics of the persons with the highest and the lowest response propensity.

TABLE 11.2 The lowest and highest response propensity

Variable	Value for lowest propensity	Value for highest propensity
Gender	Male	Female
Is married	No	No
Age in 13 age groups	45–49	18–19
Type of non-native	First generation non-western	Native
Size of the household	2	5 or more
Type of household	Other	Couple with children
Has listed phone number	No	Yes
Has a job	No	Yes
Region of the country	Metropolis	Woodlands
Degree of urbanization	Very strong	Not
Response propensity	0.109	0.798

The person with the lowest response propensity is an unmarried, middle-aged, non-native male. He lives in a big city and has no job. There are two people in the household, but the type of household is unclear. He does not have a listed phone number.

The person with the highest response probability is a native young girl. She is one child in a larger household (five persons or more) living in a rural area. She has a job and a listed phone number.

Another issue is that it is not always clear which variables are required in the model for explaining response behavior. Often one has to do it with the available variables, whereas proper estimation of response propensities may require more variables. Imbens and Rubin (2010) devote a whole chapter to the estimation of the response propensities. They describe how the model can be built with first- and second-term interactions of explanatory variables. They also describe how the adequacy of the model specification can be assessed by inspection of the estimated response propensities.

Schouten (2004) shows that just including main terms in the model is often sufficient for estimating response propensities. Adding interaction terms does not improve the explanatory power of the models.

EXAMPLE 11.3 Constructing a model for response propensities

Bethlehem, Cobben, and Shouten (2011) describe how a model for the response propensities can be constructed. First the bivariate relationship between the auxiliary variables and the response indicator is evaluated using Cramér's V . The value of this quantity is always between 0 and 1, where 0 means no relationship and 1 is a very strong relationship. The results are displayed in Table 11.3.

TABLE 11.3 Cramér's V statistic: The strength of the relationship between the auxiliary variables and the response behavior

Auxiliary variable	V
Region of the country	0.163
Degree of urbanization	0.153
Has listed phone number	0.150
Percentage non-natives in neighborhood	0.138
Percentage non-western non-natives in neighborhood	0.133
Average house value in neighborhood	0.115
Type of non-native	0.112
Type of household	0.106

Size of the household	0.099
Marital status	0.097
Is married	0.097
Is non-native	0.087
Has social allowance	0.077
Age in 13 classes	0.061
Has an allowance	0.061
Children in household	0.056
Has a job	0.037
Age in 3 classes	0.030
Has disability allowance	0.021
Gender	0.011
Has unemployment allowance	0.000

Apparently, variables related to the degree of urbanization have the strongest relationship. Also the variable indicating whether someone has a listed telephone number seems to be important. A variable like gender is less important.

Next, a multivariate response model is constructed by starting with the most significant variable and stepwise including less significant variables until no more significant relationships with the response indicator remain. Table 11.4 contains the resulting model.

TABLE 11.4 Multivariate model for the response propensities

Variable	Wald χ^2
Having a listed phone	242.0
Region	164.8
Ethnic background	93.7
Size of household	52.6
Age (3 categories)	11.0
Has a job	23.7
Marital status	74.1
Gender	14.7
Type of household	23.6
Has a social allowance	8.6
Average housevalue	25.3
Degree of urbanization	16.3
Pseudo R^2	0.042
χ^2	1805.62
d.f.	40

The significance is tested with a Wald test. The χ^2 -value for each variable and for the full model is given. Furthermore, the Nagelkerke pseudo R^2 (Nagelkerke, 1991) is reported as a measure of the model fit.

11.2.5 CORRECTION METHODS BASED ON RESPONSE PROPENSITIES

Once response propensities have been estimated, they can be used to reduce a possible selection bias. There are two general approaches: *response propensity weighting* and *response propensity stratification*. They will be described in the subsequent sections.

The theory will be restricted to the situation in which first a probability sample has been selected, and then problems are encountered in obtaining the required information from every sample element.

11.2.5.1 Response Propensity Weighting. *Response propensity weighting* is an approach that recognizes the principle of survey sampling that unbiased estimators can only be constructed and computed if the selection probabilities of the observed elements are known. In the case of selection problems (under-coverage, nonresponse, and self-selection), the true selection probability of an element is the product of selection probability as defined in the sampling design and the response probability. To obtain (estimates of) these true selection probabilities, the (unknown) response probabilities are replaced by estimated response propensities.

In the ideal situation, in which every sample element can be observed, the Horvitz–Thompson estimator defined by

$$(11.32) \quad \bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^N a_k \frac{Y_k}{\pi_k}.$$

This is an unbiased estimator of the population mean. The indicator variable a_k denotes whether element k is selected in the sample ($a_k = 1$) or not ($a_k = 0$), and π_k is the first-order inclusion probability of element k .

In the case of, for example, nonresponse, each sample element k has a certain, unknown probability ρ_k of responding. To avoid a possible bias, the Horvitz–Thompson estimator could be modified to include these response probabilities:

$$(11.33) \quad \bar{y}_{HT,R} = \frac{1}{N} \sum_{k=1}^N a_k R_k \frac{Y_k}{\pi_k \rho_k},$$

where R_k indicates whether element k responds. This is an unbiased estimator, but it cannot be computed because the values of the ρ_k are unknown. The solution is to replace each ρ_k by its estimated response propensity $\rho(X_k)$, resulting in

$$(11.34) \quad \hat{y}_{HT,R} = \frac{1}{N} \sum_{k=1}^N a_k R_k \frac{Y_k}{\pi_k \hat{\rho}(X_k)}.$$

Note that it is not so easy to establish the statistical properties of this estimator. Its distribution is not only determined by the sampling design but also by the response behavior mechanism and response propensity model. However, if the proper model is used, this estimator should be approximately unbiased.

Kalton and Flores-Cervantes (2003) note that if only categorical auxiliary variables are used in the logistic regression model, and there are no interactions included in the model, weighting with estimated response propensities is similar to raking ratio estimation. This type of adjustment weighting is described in Chapter 10. There is also a difference: If raking ratio estimation is applied, the weighted marginal distributions of the auxiliary variables are equal to their corresponding population distributions. So representativity with respect to these variables is guaranteed. This is not the case for weighting with response propensities.

The Horvitz–Thompson estimator is only a simple estimator that does not make use of any additional information. If a relationship exists between the target variable of the survey and a number of auxiliary variables, and the distribution of these auxiliary variables in the population is known, better estimators can be constructed producing more precise estimates. An example of such an estimator is the *generalized regression estimator*. It was already described in Chapter 3.

Suppose that p auxiliary variables are available. The p -vector of values of these variables for element k is denoted by

$$(11.35) \quad X_k = (X_{k1}, X_{k2}, \dots, X_{kp})'.$$

The vector of population means of the p auxiliary variables is denoted by

$$(11.36) \quad \bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)'$$

This vector is supposed to be known. If the auxiliary variables are correlated with the target variable, then for a suitably chosen vector $B = (B_1, B_2, \dots, B_p)'$ of regression coefficients for a best fit of Y on X , the residuals E_k , defined by

$$(11.37) \quad E_k = Y_k - X_k B$$

vary less than the values Y_k of the target variable itself. Application of ordinary least squares results in

$$(11.38) \quad B = \left(\sum_{k=1}^N X_k X_k' \right)^{-1} \left(\sum_{k=1}^N X_k Y_k \right).$$

If all sample elements respond, then for any sampling design, the vector B can be estimated by

$$(11.39) \quad b = \left(\sum_{k=1}^N a_k \frac{X_k X_k'}{\pi_k} \right)^{-1} \left(\sum_{k=1}^N a_k \frac{X_k Y_k}{\pi_k} \right).$$

The estimator b is an asymptotically design unbiased (ADU) estimator of B . It means the bias vanishes for large samples. Using expression (11.39), the generalized regression estimator is defined by

$$(11.40) \quad \bar{y}_{GR} = \bar{y}_{HT} + (\bar{X} - \bar{x}_{HT})' b,$$

where \bar{x}_{HT} and \bar{y}_{HT} are the Horvitz–Thompson estimators for the population means of X and Y , respectively. The generalized regression estimator is an ADU estimator of the population mean of the target variable.

If there are selection problems, each sample element k has a certain, unknown probability ρ_k of responding. To avoid a possible bias, the generalized regression estimator can be modified to include these response probabilities, resulting in

$$(11.41) \quad b_R = \left(\sum_{k=1}^N a_k R_k \frac{X_k X_k'}{\pi_k \rho_k} \right)^{-1} \left(\sum_{k=1}^N a_k R_k \frac{X_k Y_k}{\pi_k \rho_k} \right),$$

where R_k indicates whether element k responds. This is an approximately unbiased estimator, but it cannot be computed because the values of the ρ_k are unknown. The solution is to replace each ρ_k by its estimated response propensity $\hat{\rho}(X_k)$, resulting in

$$(11.42) \quad \hat{b}_R = \left(\sum_{k=1}^N a_k R_k \frac{X_k X_k'}{\pi_k \hat{\rho}(X_k)} \right)^{-1} \left(\sum_{k=1}^N a_k R_k \frac{X_k Y_k}{\pi_k \hat{\rho}(X_k)} \right).$$

Again, computation of the statistical properties of this estimator is not straightforward. Its distribution is determined by the sampling design, the response behavior mechanism, and the response propensity model. However, if the proper models for the response model and for the target variable are used, this estimator should be approximately unbiased and be more precise than the Horvitz–Thompson estimator.

It was already mentioned that response propensity weighting does not force the weighted distribution of the auxiliary variables in the logit model to be equal to their population distributions. To perform this kind of calibration, the auxiliary variables can be included in the generalized regression model for the target variable.

More information about response propensity weighting can be found, for example, in the studies by Kalton and Flores-Cervantes (2003), Särndal (1981), Särndal and Lundström (2005, 2008), and Little (1986).

11.2.5.2 Response Propensity Stratification. *Response propensity stratification* takes advantage of the fact that estimates will not be biased if all response probabilities are equal. In this case, selection problems will only lead to fewer observations, but the composition of the sample is not affected. The idea is to divide the sample in strata in such a way that all elements within a stratum have (approximately) the same response probabilities. Consequently, unbiased estimates can be computed within strata. Next, stratum estimates are combined into a population estimate.

In the case of response propensity stratification, the final estimates rely less heavily on the correctness of the model that is used to calculate the response propensities. The reason is that the exact values are not used in the computation. They are just used to construct strata. Hence, the propensity score $\rho(X)$ is smoothed.

Suppose the sample is stratified into L strata based on the response propensities. Cochran (1968) suggests that five strata are sufficient (i.e., $L = 5$). The strata are denoted by U_1, U_2, \dots, U_L . The sample size in stratum h is denoted by n_h . These sample sizes are random variables and not fixed numbers. Assuming simple random sampling, the *response propensity estimator* for the population mean of the target variable Y is now defined by

$$(11.43) \quad \bar{y}_{RPS} = \frac{1}{n} \sum_{h=1}^L n_h \bar{y}_R^{(h)},$$

where $\bar{y}_R^{(h)}$ is the mean of the responding elements in stratum h , for $h = 1, 2, \dots, L$.

Note that this poststratification estimator calibrates the response to the sample level instead of the population level (just like the poststratification estimator in Chapter 3).

There are several ways to construct strata. Preferably, the strata should be constructed in such a way that the response propensities vary as little as possible within strata. The starting point is the distribution of the estimated response propensities. Then decisions have to be made about the number of strata and about the width of the strata (in terms of values of the response propensities).

Imbens and Rubin (2010) propose an iterative procedure to determine the number of strata in which the auxiliary variables are balanced between participants and nonparticipants. According to Cochran (1968), five strata are enough for stratification purposes. This is a rule of thumb. However, one should notice that the more strata there are, the less variation there will be within the strata and the more the distribution of the strata will resemble a continuous distribution.

Once the strata are constructed, the conditional independence can be checked by testing for a bivariate relationship between the response indicator and the auxiliary variables within each of the strata. This can be done, for instance, by Cramér's V statistic, which is based on the χ^2 test statistics.

EXAMPLE 11.4 Constructing response propensity strata

Example 11.2 described how the response propensity model was constructed for the General Population Survey (GPS). The estimated response propensities varied between 0.10 and 0.80. The response propensities were divided into five strata by dividing the interval $[0.20, 0.80]$ in five intervals of equal width. Table 11.5 summarizes this stratification. Note that the number of responding elements increases as the response propensity increases, whereas the number of sample elements in the strata does not necessarily increase.

TABLE 11.5 The response propensity strata

Stratum	Range	Sample size	Respondents
1	0.10 – 0.24	303	63
2	0.24 – 0.38	1,913	609
3	0.38 – 0.52	5,385	2,504
4	0.52 – 0.66	14,690	8,777
5	0.66 – 0.80	9,728	6,839
Total		32,019	18,792

Figure 11.3 uses a kernel density technique to show the distribution of the estimated response propensities. The five bars represent the strata.

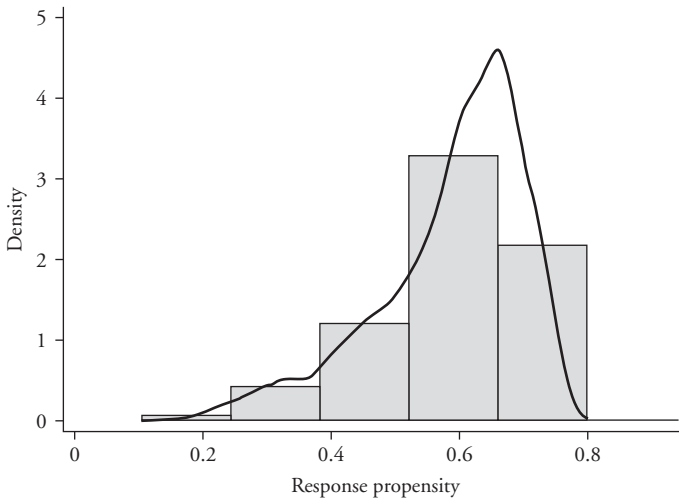


FIGURE 11.3 Response propensity stratification with 5 strata

In addition to the five strata, a kernel density line has been estimated. This represents the continuous distribution of the response propensity in the sample. It is clear that the histogram is not a very accurate approximation of the distribution of the response propensities.

Figure 11.4 shows what happens if 25 strata are constructed instead of 5. The histogram is now much closer to the density function.

One should be careful in constructing too many strata. This will reduce the number of observations per strata and, therefore, may lead to less stable estimates.

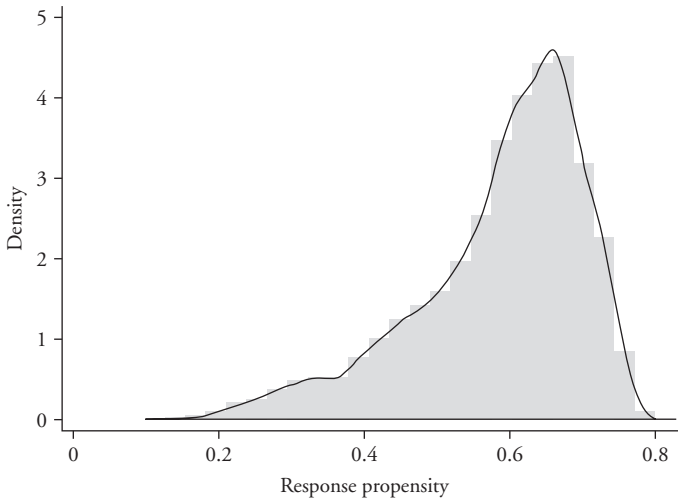


FIGURE 11.4 Response propensity stratification with 25 strata

11.3 Application

The theory described in the previous sections is now applied to the case of a self-selection in web surveys. Putting the web survey in the context of an experimental design, there are two possible “treatments”: participating in the web survey and not participating in the web survey. In line with Biffignandi and Pratesi (2004), the treatment is participation in the web survey. Note that in a web survey, the assignment to the treatment is the result of being a web user in the general population. The untreated units are those who belong to the target population but do not participate into the web survey. For the target variable of the survey, only the values of participants in the web survey become available.

The population is generated for this application. This has the advantage that the properties of the population and relationships between variables are exactly known. Web survey samples are generated to study the properties of estimators

and correction techniques that attempt to reduce the bias using response propensities. It is noted once more that if the values of the response propensities are known, the inference problem could be solved by traditional estimation methods like the Horvitz–Thompson estimator.

It is assumed in this simulation study that the values of auxiliary variables for participating and nonparticipating individuals can be retrieved from a population register. So the reference survey approach is not used. Three different estimation techniques are compared: (1) naïve estimation without any kind of correction, (2) response propensity weighting, and (3) response propensity stratification.

11.3.1 GENERATION OF THE POPULATION

A small population has been generated consisting of 260 web users and 1,500 nonweb users. There are three auxiliary variables X_1 , X_2 , and X_3 . The values of these variables have been generated differently for the web users and the nonweb users.

For the web users, the values of X_1 were obtained as drawn from a normal distribution, the values of X_2 from a Gamma distribution, and the values of X_3 from a distribution that describes the fact that the probability of response decreases when t (time from first contact) increases:

- $X_1 \sim Normal(6, 4)$
- $X_2 \sim Gamma(5)$
- $X_3 \sim e^{-t} + 3$

For the nonweb users, covariates were obtained using the same distributions, but with different parameters:

- $X_1 \sim Normal(10, 5)$
- $X_2 \sim Gamma(3)$
- $X_3 \sim e^{-t}$

Table 11.6 shows the mean and standard deviations of the three variables in the two subpopulations. The table also contains the p -values of the test of equality of the means. It is clear that the means differ significantly.

Given the values of the auxiliary variables X_1 , X_2 , and X_3 , the response propensities can be estimated using a model that explains being a web user from the auxiliary variables. The target variable Y is constructed as a linear function of the auxiliary variables and a noise term:

$$Y = 0.2X_1 + 0.3X_2 + 0.5X_3 + 10U,$$

where the random variable U has a uniform distribution on the interval $[0, 1]$.

TABLE 11.6 Characteristics of the generated population (size, mean, and standard deviation)

	Web users	Nonweb users	<i>p</i> -value
<i>N</i>	260	1500	
X_1	5.87 ± 3.95	9.97 ± 5.00	< 0.005
X_2	5.13 ± 2.05	3.01 ± 1.73	< 0.005
X_3	4.07 ± 1.12	1.01 ± 0.99	< 0.005

11.3.2 GENERATION OF RESPONSE PROBABILITIES

The probabilities of the elements of being web users are generated using a logistic regression model. The auxiliary variables X_1 , X_2 , and X_3 are used as explanatory variables. The dependent variable is the indicator of being a web user. These probabilities were used as response probabilities of the elements in the self-selection web survey.

It is noted again that the values of the three auxiliary variables are assumed to be available in a population register. Thus, the problem is avoided of what to do if the auxiliary variables are only available for all elements in the sample. This would raise the question of whether to use adjustment weights. The literature is not clear about this.

11.3.3 GENERATION OF THE SAMPLE

Samples were selected from the population of web users. The response propensity were taken equal to the probability of being a web user. So it was assumed that a large probability web user will also have a large probability to participate in web surveys when confronted with an invitation to do so.

The sample selection mechanism can be considered a form of Poisson sampling with individual participation probabilities being equal to the generated response probabilities.

11.3.4 COMPUTATION OF RESPONSE PROPENSITIES

The information about the responding elements was used to estimate the response propensities. Again, a logistic regression model (11.30) was used with the three auxiliary variables X_1 , X_2 , and X_3 . It was assumed that the values of these variables for the nonparticipating elements can be retrieved from a population register. The participation indicator R (with values 0 and 1) is the dependent variable in the model. The expected value of R_i for element i is the response propensity of element i . Note that the self-selection mechanism causes the sample size n_S to be a random variable.

11.3.5 MATCHING RESPONSE PROPENSITIES

The estimated response propensities were used to find similar elements in the target population. So, participating elements are matched to other (nonparticipating) elements using their response propensities.

The first step was to put elements in the same stratum if their response propensities are equal for the first five decimal digits. The next step was to match the remaining unmatched web survey participants on the basis of four decimal digits. This was continued down to a one-digit match.

Figures 11.5, and 11.6 show the response probability distribution for the original population (separately for web users and nonweb users). Figure 11.7 shows the response propensity distribution for the web survey participants.

The set of web survey respondents after matching on the response propensities was reduced to only those participants having matches among the nonparticipants. This ensured that, conditionally on the auxiliary variables in the model for the response propensities, the assumption of conditional independence was satisfied.

There is a substantial difference between the range of response probabilities for the web users and the nonweb users. For the web users, the minimum response probability was 0.088422 and the maximum response probability was

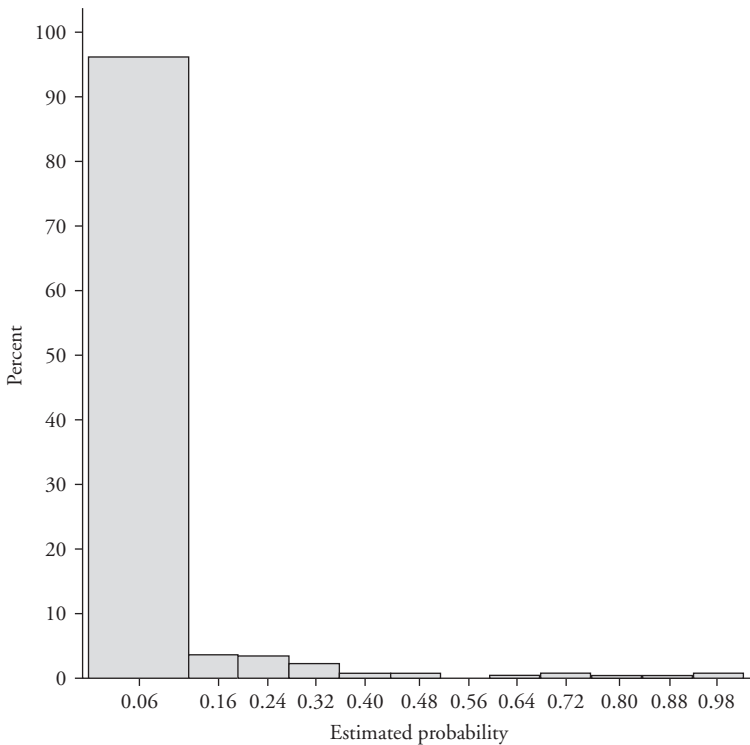


FIGURE 11.5 Response probabilities for nonweb users

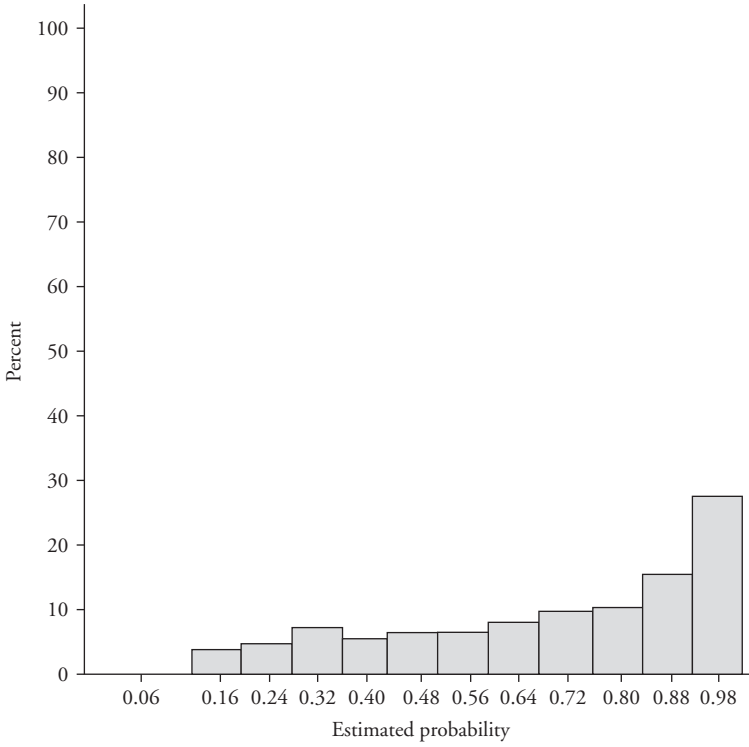


FIGURE 11.6 Response probabilities for web users

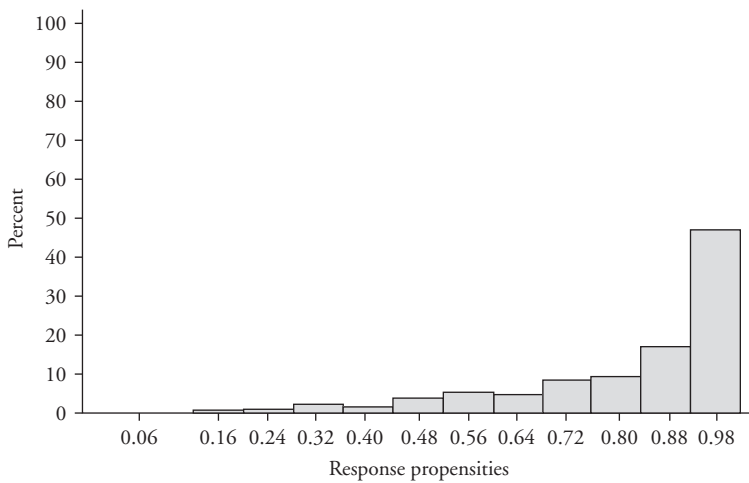


FIGURE 11.7 Estimated propensity scores for web survey participants

TABLE 11.7 Characteristics of response probabilities and response propensities

Quantile	Response probabilities for web users	Response probabilities for nonweb users	Response propensities for participants
100% (maximum)	0.999997	0.997206	0.999997
75% (quantile 3)	0.960129	0.011523	0.977122
50% (median)	0.825558	0.002353	0.904643
25% (quantile 1)	0.591330	0.000704	0.733165
0% (minimum)	0.088422	0.000021	0.135205

0.999997. For the nonweb users, all response probabilities should theoretically be equal to 0. However, the estimation process introduces small deviations from 0. In any case, 90% of the probabilities are equal to 0.

For the generated population of potential web survey participants, the minimum response propensity was 0.135205 and the maximum propensity was 0.999997. Table 11.7 shows some characteristics for each distribution. The web survey participants with the highest propensity and the nonweb users with the smallest probability have been excluded.

11.3.6 ESTIMATION OF POPULATION CHARACTERISTICS

Weights were assigned to the respondents of the web survey. These weights were the reciprocals of the propensity scores. The weighted data were used to estimate the mean, the total, and the standard deviation of the target variable Y . Two approaches were explored. The first approach was *response propensity weighting*. This is an application of the Horvitz–Thompson estimator where the inclusion probabilities are replaced by the estimated response propensities. This approach is described in Section 11.2.5.1. In the case of self-selection, estimator (11.34) is replaced by

$$(11.44) \quad \hat{y}_{HT,SS} = \frac{1}{N} \sum_{k=1}^N R_k \frac{Y_k}{\hat{\rho}(X_k)}.$$

A similar expression can be derived for estimating population totals and population standard deviations.

The second approach was to apply *response propensity stratification*. This is a form of poststratification estimation where strata are constructed on the basis of the values of the estimated response propensities. This approach is described in Section 11.2.5.2. The basic formula is

$$(11.45) \quad \bar{y}_{PS,SS} = \frac{1}{n_S} \sum_{b=1}^L n_{S,b} \bar{y}_R^{(b)},$$

where $n_{S,b}$ is the number of responding elements in stratum b and $\bar{y}_R^{(b)}$ is the mean of the target variable of the responding elements in stratum b . Similar expressions can be derived for estimating other population characteristics, like population totals and population standard deviations.

11.3.7 EVALUATING THE RESULTS

To illustrate the performance of the proposed estimators based on propensity scores relative to other traditional estimation methods (no adjustment and poststratification using auxiliary variables), estimates of the population mean, total, and standard deviation were compared.

Estimators were compared using the empirical relative bias (RB)

$$(11.46) \quad \text{RB} = \frac{T - \theta}{\sqrt{\text{MSE}(T)}},$$

in $\text{MSE}(T)$ is the empirical mean square error of the estimator T and θ is the parameter of interest. The performance of the response propensity estimation approaches described above were compared for two different situations:

1. Comparing estimators for parameters of the whole target population with estimators for parameters of the population of web users.
2. Comparing response propensity approaches with traditional estimators like no correction (i.e., no weights) and poststratification with auxiliary variables.

With respect to 1), the relative bias was lower if the target was the web population. The performance of the proposed estimators was not poor for the whole population, especially for estimating the total (RB = 0.02 versus RB = -0.03 for propensity weighting, and RB = 0.01 versus RB = -0.02 for the propensity stratification). Estimation of the standard deviation produced the same level of relative bias both for the whole and for the web population.

With respect to 2), the results were good for propensity stratification. The relative bias was smaller compared with traditional estimation methods. The results for propensity weighting were not as good. This was caused by the relatively high values and small variation of the response propensities among the participants, which makes this approach more or less similar to estimation without weighting. Note that in the case of poststratifying with auxiliary variables, only two strata were constructed for each variable: one stratum with individuals having a value below the mean and a stratum with individuals having a value above the mean.

Also note that the response propensity strata have to be constructed so that within the strata the response behavior is homogeneous. Cochran (1968) suggests that it is enough to use five strata.

11.3.8 MODEL SENSITIVITY

A final step in this experiment was investigating how sensitive the results were for changes in the model that was used for generating the population. To get an

answer to this question, the linear relationship between the study variable and the auxiliary variables was replaced by a nonlinear relationship. Moreover, the shapes of the distributions of the auxiliary variables were made differently for web users than for nonweb users.

Response propensity stratification performed best when the relationship between the target variable and the auxiliary variables was not linear any more. There was no difference in performance between response propensity weighting and response propensity stratification if the distributions of the auxiliary variables differed for web users than for nonweb users.

The results of this experiment show that the response propensity matching combined with response propensity stratification is a promising approach for reducing the self-selection bias of web surveys. More research is needed to implement further adjustments to the propensity weighting. In fact, several studies are going on. For instance, Schonlau, Van Soest, Kapteyn, and Couper (2009) investigate whether adjustment using weights or matching on a small set of auxiliary variables makes the distribution of the target variables representative of the population. They extract an Internet sample from the Health and Retirement Study (HRS), which is representative of the U.S. population of 50 year old or older. Several studies are in progress with respect to selection bias in volunteer web panels, too. For example, Lee and Valliant (2009) and Bethlehem, Cobben, and Schouten (2011) study the effects of a combination of response propensity weighting or stratification with traditional correction techniques.

11.4 Summary

In web surveys, selecting a proper probability sample requires a sampling frame containing the e-mail addresses of all individuals in the population. Such sampling frames rarely exist. Actually, general-population sampling frames do not contain information about which people have Internet access and which do not. Thus, one should bear in mind that people not having Internet access will not respond to a web questionnaire (if no special data contact and collection strategies are performed). Moreover, people having Internet access will also not always participate. Taking these facts into account, it is evident that the ultimate group of respondents is the result of a selection process (mostly self-selected) with unknown probabilities. Even if the target population coincides with all individuals having Internet access, some problems remain as a result of self-selection.

One possible solution for correcting the bias from selection problems is using response propensities. The response propensity is the conditional probability that a person responds to the survey request, given the available background characteristics. To compute response propensities, auxiliary information for all sample elements is needed.

The response propensities can be used in a direct way for estimation of the target variables directly by using the response propensities as weights. This is called response propensity weighting. The direct approach attempts to estimate the true selection probabilities by multiplying the first-order inclusion probabilities with the estimated response propensities. Bias reductions will only be

successful if the available auxiliary variables are capable of explaining the response behavior.

The response propensities also can be used indirectly, by forming strata of elements having (approximately) the same response propensities. This is called response propensity stratification. The final estimates rely less heavily on the accuracy of the model for the response propensities.

Some studies show that response propensity matching combined with response propensity stratification is a promising strategy for the adjustment of the self-selection bias in web surveys. Research is ongoing to implement further improvements for response propensity weighting.

Propensity score adjustment is a frequently adopted solution to improve the representativity of web panels. It should be noted that there is no guarantee that correction techniques are successful. See also Chapter 12.

KEY TERMS

Propensity score: The conditional probability of assignment to a particular treatment given a vector of observed auxiliary variables X .

Propensity score method: The propensity score method originates from evaluation studies that estimate average treatment effects. In treatment effect studies, there usually are two groups involved: one group that receives the treatment, and one group that serves as a control group and does not receive the treatment. The statistic of interest is the effect of the treatment. However, to measure this effect without bias, it is necessary to remove all possible differences in outcome that originate as a result of a different composition of the treatment and the control group. For this purpose, the propensity score is used to balance the composition of the two groups.

Response probability: The probability that a person responds to a survey. This is a theoretical quantity.

Response propensity: The response propensity $\rho_k(X)$ is the probability of response of element k given the values of a set of auxiliary variables X .

Response propensity stratification: A poststratification method to adjust for nonresponse bias, where the strata are based on the response propensities. This method uses response propensities to construct strata.

Response propensity weighting: A weighting method to adjust for nonresponse bias, where the weights are based on the response propensities and directly used in the estimation of the target variable.

EXERCISES

Exercise 11.1. The aim of this exercise is to perform a propensity weighting adjustment as described in Section 11.2.5.1 using the data from the GPS. The data can be found on the website: www.web-survey-handbook.com.

First, estimate response probabilities using a binary logistic regression model and all the available auxiliary information. In SPSS (SPSS Corporation Chicago, IL)

one can use the forward selection strategy based on the Wald criterion to do this. Save the estimated response propensity as *rprob*, and label the variable *Response propensity*. Compute weights by taking the inverse of the estimated response probabilities *rprob*.

Use **Transform | Compute** to introduce a new variable *propwght*. Its value is obtained for a respondent by $propwght = 1 / rprob$.

To scale the weights properly, they have to be divided by the average response probability of the respondents. Consequently, the weights will average to 1 over the response. First, select the respondents with **Data | Select cases**. Then compute the average weight of the respondents with **Analyze | Descriptive statistics | Descriptives**. Finally, use **Transform | Compute** again to adjust the values of *propwght*.

Weight the response data using **Data | Weight cases**. Be sure that only respondents have been selected. Compute weighted estimates of the two target variables “Owns a house” and “PC in household” by **Analyze | Descriptive Statistics | Frequencies**. What are the estimates? Compare them with the response means.

Exercise 11.2. The aim of this exercise is to perform a response propensity stratification adjustment as described in Section 11.2.5.2 using the data from the GPS. This comes down to poststratification using strata constructed on the basis of response probabilities.

First, a categorical variable *propclas* is needed. This is obtained by aggregating individuals into five classes, based on the estimated response probabilities *rprob*. To determine these classes, look at the distribution of the estimated response probabilities by plotting a histogram of the estimated response probabilities for the complete, unweighted sample. **Graphs | Histogram** can be used for this.

Classes can be formed to divide the individuals equally over the response probabilities (i.e., every class contains the same number of individuals). Another option is to form classes according to the response probabilities (i.e., every class has the same width of the response probabilities). In this exercise, classes with the same width in response propensities are constructed. The first class comprises sample persons with a response propensity between 0 and 0.2; the second class between 0.2 and 0.4; the third between 0.4 and 0.6; the fourth between 0.6 and 0.8; and the last between 0.8 and 1. The last class, however, is empty because the maximum response propensity is < 0.8 . So there will be only four classes.

Use **Transform | Recode | Into different variables** to recode the variable *rprob* into the four desired classes. Recode to an output variable *propclas*, and assign the label “propensity class.” The resulting variable *propclas* will be the stratification variable for the nonresponse adjustment.

First compute the four weights. As we weight to the sample, use

$$w_i = \frac{n_b/n}{r_b/r}$$

for the weight of an element in class h , and use the estimator

$$\bar{y}_{PS} = \frac{1}{n} \sum_{h=1}^L n_h \bar{y}^{(h)},$$

to estimate the target variables “Owns a house” and “Has a PC.” Here, n_h is the sample size of class h , n is the sample size, r_h is the response size in class h , r is the response size, and $\bar{y}^{(h)}$ is the response mean of the target variable in class h . To be able to compute the poststratification estimator, these quantities need to be computed first. To this end, complete the table below:

h	n_h	r_h	w_h
1			
2			
3			
4			
Total	32,019	18,792	

Derive the class sizes n_h and r_h by **Analyze | Descriptive Statistics | Crosstabs** by crossing the response indicator *response* with the variable *propclas*. Compute the weights by hand and import them in SPSS by constructing a new weight variable *propwgt2* using **Transform | Compute** (you have to do this four times, for every category of *propclas*).

Finally, estimate the target variables by activating the weights with **Data | Weight cases**, respondent selection in **Data | Select cases**, and weighted estimation in **Analyze | Descriptive Statistics | Frequencies**. What are the estimates? Are they different from the first approach? Also compare them with the response means.

REFERENCES

- Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980), Issues in the Analysis of Selectivity Bias. In: Stromsdorfer, E. & Farkas, G. (eds.), *Evaluation Studies, Volume 5*, Sage, San Francisco, CA, pp. 42–59.
- Bethlehem, J. G. (1988), Reduction of Nonresponse Bias through Regression Estimation. *Journal of Official Statistics*, 4, pp. 251–260.
- Bethlehem, J. G. (2002), Weighting Nonresponse Adjustments Based on Auxiliary Information. In: Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (eds.). *Survey Nonresponse*. John Wiley & Sons, New York.
- Bethlehem, J. G. (2007), *Reducing the Bias of Web Survey Based Estimates*. Discussion paper 07001, Statistics Netherlands, Voorburg/Heelen, the Netherlands.
- Bethlehem, J. G. (2009), *Applied Survey Methods: A Statistical Perspective*, John Wiley & Sons, NJ.

- Bethlehem, J. G., Cobben, F., & Schouten, B. (2011), *Handbook on Nonresponse in Household Surveys*. John Wiley & Sons, Hoboken, NJ.
- Biffignandi, S. & Bethlehem, J. G. (2011), Web Surveys: Methodological Problems and Research Perspectives. In: *Theoretical and Applied Statistics*, Springer.
- Biffignandi, S. & Pratesi, M. (2004), Indagini WEB: Propensity Score Matching e Inferenza. Un'analisi Empirica ed uno Studio di Simulazione (Web Surveys: Propensity Scores Matching and Inference. An Empirical Analysis and a Simulation Study). In: Falorsi, P., Pallara, A., & Russo, A. (eds.), *L'Integrazione di Dti di Fonti Diverse. Tecniche e Applicazioni del Record Linkage e Metodi di Stima Basati Sull'uso Congiunto di Fonti Statistiche e Amministrative*. Franco Angeli, Milano. Italy.
- Cochran, W. (1968), The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24, pp. 205–213.
- Dehejia, R. & Wahba, S. (1999), Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94, pp. 1053–1062.
- Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005), Comparing Data from Online and Face-to-Face Surveys. *International Journal of Market Research*, 47, pp. 615–639.
- Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998), Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66, pp. 1017–1098.
- Horvitz, D. G. & Thompson, D. J. (1952), A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, pp. 663–685.
- Imbens, G. & Rubin, D. B. (2010), *Causal Inference in Statistics*. Forthcoming.
- Imbens, G. (2004), Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economic and Statistics*, 86, pp. 4–29.
- Kalton, G. & Flores-Cervantes, I. (2003), Weighting Methods. *Journal of Official Statistics*, 19, pp. 81–97.
- Keller, W. J., Verbeek, A., & Bethlehem, J. G. (1984), *ANOTA: Analysis of Tables*. Report 5766–84-M1–3. Statistics Netherlands, Department for Statistical Methods, Voorburg, the Netherlands.
- Lee, S. (2006), Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *Journal of Official Statistics*, 22, pp. 329–349.
- Lee, S. & Valliant, R. (2009), Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods Research*, 37, pp. 319–343.
- Lechner, M. (1999), *Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption*. IZA Discussion Papers 91, Institute for the Study of Labor (IZA), Bonn, Germany
- Little, R. (1986), Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review*, 54, pp. 139–157.
- Little, R. & Rubin, D. (2002), *Statistical Analysis with Missing Data*, 2nd ed. John Wiley & Sons, New York.
- Nagelkerke, N. J. D. (1991), Miscellanea, A Note on a General Definition of the Coefficient of Determination, *Biometrika*, 78, pp. 691–692.
- Rosenbaum, P. R. & Rubin, D. B. (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, pp. 41–55.

- Rubin, D. B. (1974), Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66, pp. 688–701.
- Särndal, C. E. (1981), Frameworks for Inference in Survey Sampling with Application to Small Area Estimation and Adjustment for Non-Response. *Bulletin of the International Statistical Institute*, 49, pp. 494–513.
- Särndal, C. E. & Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester, U.K.
- Särndal, C. E. & Lundström, S. (2008), Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. *Journal of Official Statistics*, 24, pp. 167–191.
- Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., Spranca, M., Kan, H., Turner, R., & Berry, S. H. (2004), A Comparison between Responses from a Propensity-weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review*, 22, pp. 128–138.
- Schonlau, M., Van Soest, A., Kapteyn, A., & Couper, M. P. (2009), Selection Bias in Web Surveys and the Use of Propensity Scores, *Sociological Methods & Research*, 37, pp. 291–318.
- Schouten, B. (2004), *Adjustment for Bias in the Integrated Survey on Living Conditions (POLS) 1998*. Discussion paper 04001, Statistics Netherlands, Voorburg, the Netherlands.
- Smith, J. & Todd, P. (2000), *Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?* Paper presented at the University of North Carolina, Chapel Hill, NC.
- Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W., & Terhanian, G. (2001), The Record of Internet-Based Opinion Polls in Predicting the Results of 72 Races in the November 2000 U.S. Elections. *International Journal of Market Research*, 43, pp. 127–136.
- Terhanian, G., Smith, R., Bremer, J., & Thomas, R. K. (2001), *Exploiting Analytical Advances: Minimizing the Biases Associated with Internet-Based Surveys of Non-Random Samples*, ARF/ESOMAR: Worldwide Online Measurement, ESOMAR Publication Services, Vol. 248, pp. 247–272.

Web Panels

12.1 Introduction

A *web panel* (also called an *on-line panel*, *Internet panel*, or *access panel*) is a survey in which the same individuals are interviewed via the web at different points in time. Information is therefore collected in a longitudinal way using the same group of individuals. Panels differ from cross-sectional surveys, which if repeated over time select a new sample with each survey release. Because of their specific characteristics, web panels have several advantages:

- Web panels offer the possibility of continuously updating information about the individuals. As a consequence, they allow for comparisons of survey results over time. This means that a panel is particularly useful for investigating changes over time.
- A single web panel can be used to collect data in many different areas, such as market research, medical research, psychological research, and sociological research. Also, many different issues can be addressed. For example, data can be collected over time on employment status, education, health, buying and consuming behavior, and product preferences. Information on rare events, such as crimes or accidents, also can be adequately surveyed using web panels.
- Different types of objects can be investigated in a web survey, such as individuals, households, and companies. It is also possible to select samples from web panels for specific purposes.
- Data can be collected fast and cheap.

- The reliability of the information provided by the respondents can be partially checked automatically.
- Survey results can be linked to demographic information.

Market research (Postoaca, 2006a; Comley, 2007) makes extensive use of web panels. For many years, web panels have been used in the United States. Recent trends indicate a large-scale diffusion of this data collection mode in most European countries. This mode is now being recognized as one of the most important market research survey tools. In 2006, ESOMAR (the European Society for Opinion and Marketing Research) organized a Conference on Panel Research, with a special focus on web panels. Note that ESOMAR's mission is to promote the use of opinion and market research for improving decision making in business and society worldwide.

■ EXAMPLE 12.1 The Dutch Online Panel Study

Vonk, Van Ossenbruggen, and Willems (2006) describe how web panels became the primary means of data collection for market research in the Netherlands. One reason was the rapidly improving Internet technology and the acceptance of on-line data collection by consumers and users of market research. Another reason was the substantial increase of the costs of traditional data collection by interviewers. And a third reason was the dropping response rates for telephone surveys.

Around 2006 there were approximately 30 commercial web panels in the Netherlands. Together they had 1,650,000 members. Note that the population of people aged 18 years and older in the Netherlands was around 12,752,000. This would suggest that approximately 13% of the population was a member of a panel.

Vonk, Van Ossenbruggen, and Willems, (2006) conducted a comparison study across 19 of these panels that together contained 90% of all panel members. It turned out that the 1,650,000 members were not unique. In fact, there were only 900,000 unique panel members. Of those, 700,000 participated in only one panel. The other 200,000 individuals were member of more panels. On average, they were members of 4.7 panels.

Vonk et al. (2006) also investigated the representativity of these self-selection panels by taking a sample of 1,000 members from each of the 19 panels. It turned out that these panels were representative with respect to basic demographic variables, like age, gender, level of education, and region of the country. However, there was a lack of representativity with respect to other variables. The panel contained too few non-Western non-natives and too many heavy Internet users. There were too few church-goers. Also, voters for the Christian-Democrats were underrepresented and voters for the Socialist Party were overrepresented.

Currently, medical research (Couper, 2007) also relies for a large part on web panels. According to Tortora (2009) and Göritz (2007), many sociological and psychological phenomena are being investigated as well by web panels.

Web panels are mainly applied in commercial market research and, to a lesser extent, in academic research, whereas in official statistics, this mode of data collection has not yet been implemented or even considered.

One of the main reasons why web panels are becoming more widespread and attractive in the world of survey research is that they are cheap to manage and to maintain. Furthermore, they can provide a lot of information that is useful for operational tasks and for augmenting the knowledge base, both in business and in sociopolitical planning.

Most web panel providers stress the representativity of their products, although the concept of representativity is not well defined and often lacks a scientific basis. Probably the lack or limitation of a methodological theory (especially for inference purposes) is one of the principal reasons for not considering web panels for making official statistics so far.

However, although the representativity of web panels remains an issue from a methodological point of view, professional survey researchers are becoming more and more conscious of the need to provide standard quality rules for web panels. They are focusing attention on the need to avoid uncontrolled web data collection processes in which there are no rules on how to achieve a required minimal level of precision, on rigor in the survey procedures, and on criteria for evaluating the minimum standards for measuring and reporting response rates. As a consequence of the above-mentioned need, many operational guides are now available. Furthermore, a certain amount of research on statistical problems has been undertaken and studies are continuing at present. See, for example, the use of the propensity score method as described in Chapter 11. This chapter also contains some examples.

EXAMPLE 12.2 Sources of quality guidelines

Many organizations and associations have provided guidelines on the use of web panels. Some sources for these guidelines are as follows:

- A consortium of German market and social research institutes released their “standards for quality assurance for online surveys.” See the report by the Arbeitskreis Deutscher Markt (2001).
- The European Federation of Associations of Market Research Organizations (EFAMRO) drafted a document on “Quality Standards for Access Panels.” See the report by the European Federation of Associations of Market Research Organizations (2004). This document is for a large part based on the European Society for Opinion and Marketing Research (ESOMAR) guidelines for on-line surveys.

- ESOMAR issued a document title, “Conducting Market and Opinion Research Using the Internet” (2005). A more recent publication by ESOMAR is “26 Questions to Help Research Buyers of Online Samples” (2008).
- The American Association for Public Opinion Research (AAPOR) established a task force on on-line panels in 2010.
- The Interactive Market Research Organization (IMRO) released a document called “Guidelines for Best Practices in Online Sample and Panel Management” (2006).

This chapter presents an overview of web panel types, together with basic theory and operational rules. The concept of representativity is discussed, and probability-based web panels are compared with self-selection panels. An overview is given of the advantages and disadvantages of web panels. The examples of existing web panels are also briefly described.

12.2 Theory

12.2.1 WEB PANEL DEFINITION AND RECRUITMENT

A web panel is a survey based on a list of objects (e.g., companies, households, or individuals) that are interviewed at different time points (called panel waves). A web panel is expected to include a large number of objects, of which certain “demographic” characteristics (in a wide sense) are known. Samples are drawn from the panel for conducting surveys on specific issues.

The way people are recruited for a web panel has an impact on what can be done with it in terms of statistical inference. In particular, the recruitment approach may lead to a selection bias. Therefore, web panels can be classified according to the recruitment method. See the studies by Lozar Manfreda and Vehovar (2008) and by Sikkel and Hoogendoorn (2008). With respect to recruitment, there are two approaches:

- Panels based on *probability sample recruitment*. These panels are composed of respondents who have been selected with some kind of probability sampling method. One example is telephone recruitment by means of random digit dialing (RDD). Other examples are recruitment by e-mail (if a list of e-mail addresses is available), mail (if addresses are available), or face-to-face (if addresses are available). The essential aspect of this approach is that a sampling frame is available. If such a frame contains auxiliary variables, it is possible to apply nonresponse correction techniques like adjustment weighting.
- Panels based on *self-selection recruitment*. Such panels are also called *volunteer panels* or *opt-in panels*. These panels are composed of respondents who

voluntarily sign up (opt-in) for the panel. Particularly in market research, there are many self-selection panels (Comley, 2007). In the field of psychology, an increasing number of studies involving random allocation experiments are now conducted using opt-in panels (Görizt, 2007; Reips, 2007). In the European literature (European Federation of Associations of Market Research Organizations, 2004), the term *access panel* is frequently used for proprietary volunteer panels containing individuals who have agreed to participate regularly in surveys run by a specific organization, generally a market research organization. Access panels can be classified into two categories according to the enrollment procedures adopted (Postoaca, 2006a):

- *Single opt-in enrollment.* No e-mail confirmation is required, and potential respondents are sent directly to a *recruitment questionnaire* page (Postoaca, 2006a). In this case, the demographic data are collected at the end of an introductory survey or their collection is spread among different surveys. Web panel members immediately become “active panel members.”
- *Double opt-in enrollment.* Potential respondents receive an e-mail confirmation. The recruitment is completed when an e-mail with panel participation confirmation is received. The e-mail confirmation contains a link that must be clicked to access the enrollment page. At this stage, a survey on the member profile is conducted. When profiling or demographic data for each respondent member have become available, an “active panel” has been obtained that can be used for surveys (Lee, 2006). The Ipsos Interactive Service (IIS) is an example of an opt-in panel. It uses a double opt-in process for all panelists. Individuals wishing to join the IIS panel first complete the on-line recruitment survey and accept the terms and conditions of membership. Ipsos informs the panelists that it agrees to keep all personal information regarding its panelists confidential (their habits, preferences, personal addresses, etc.) and that this information is used only for Ipsos research.

Respondents for self-selection panels are recruited in various, fairly spontaneous ways (Miller, 2006; Comley, 2007). Here are some examples of how one can become a member of a panel:

- Participants go to the specific panel recruitment portal themselves. They could, for example, know about the web panel via an advertisement in the media.
- Participants are redirected through banners. Pop-ups also are often used for recruiting panel members.
- Some websites are designed to “sign up” participants to several opt-in panels immediately on entering the website. Examples are www.surveymonster.net and www.yellowsurveys.com.
- Participants are asked to register in a panel at the end of another, possibly off-line, survey. In that case, the panel is populated with a subset of the respondents of the initial survey.

EXAMPLE 12.3 Some web panels in the United States and in Europe

In the United States many on-line panels exist and still more probability-based panels are in the process of being built.

Three panels that have existed for quite a long time are briefly described: Knowledge Panel, Gallup Panel, and American Life Panel.

- *The Knowledge Panel*

Knowledge Networks recruited the first on-line research panel. Its aim was to create a panel that was representative of the entire U.S. population. This panel is representative because it is recruited using high-quality probability sampling techniques, and it is not limited to current web users or computer owners. The basic characteristics of the panel are described in Table 12.1.

TABLE 12.1 Main characteristics of the Knowledge Panel

Feature	Knowledge Panel
Coverage %	97% of U.S. households
Sampling	Probability-based random
Non-Internet Population	Included
Latino Households without Internet access	Included
Sample Representativeness	Comparable with high-quality RDD with cell phone sample supplementation
Survey Frequency	2 to 4 surveys per month

The Knowledge Panel is a probability-based panel. Profiling data are surveyed every two of four months. This web panel has been collecting data for more than 10 years. It contains interesting information for researchers in government and academia, media, retail and consumer products, as well as pharmaceutical/health care firms.

- *The Gallup Panel*

Gallup began using web-based surveys in 1996. By 1998, Gallup had constructed a comprehensive web-based system with full-time programming staff devoted exclusively to supporting its web operations. Today, more than 450 simultaneous web surveys — fielded in multiple languages — are supported by the panel. Web panel members reside in all 50 states. Those who choose to join the panel are committed to the completion of two to three surveys per month. The typical survey lasts 10 to 15 minutes. The Gallup Panel uses RDD methodology to recruit its members. Households willing to participate are sent a “membership

packet” by mail with a demographic questionnaire to be completed and mailed back. The company tries to recruit all members of the household aged 13 years and older.

- *The American Life Panel*

The American Life Panel is maintained by the RAND Corporation. It consists of approximately 1,500 respondents recruited by telephone by the University of Michigan Survey of Consumer Attitudes. Non-Internet households are provided with a WebTV3. For details, the article see by Couper and Dominitz (2007).

In Europe, on-line panels are widespread in the area of market research. More recently, medical research has been collecting information using panels. These panels are mainly commercially oriented. They are not discussed here. More information can be found in ESOMAR publications. Here, some panels are mentioned that focus on collecting data related to socioeconomic phenomena, such as employment, education, and consumer behavior. These panels are not part of systems of official statistics, but they have been built up within the framework of specific projects financed by institutions committed to sponsoring research and of experiments with the use of web panels for collecting socioeconomic information.

- *The CentERpanel*

The *CentERpanel* is an Internet panel that reflects the composition of the Dutch-speaking population. It is maintained by CentERdata. CentERdata is a research institute located on the campus of the Tilburg University in the Netherlands. Participants who do not have Internet access are provided with a Net.Box, allowing them to access the Internet through their television. Households that have neither a PC nor an Internet connection are given a set-top box, which is plugged into the TV, thus enabling Internet access. See Hoogendoorn and Daalmans (2008). CentERdata recruits households via the telephone. The household members are asked to participate in survey research projects. If so, they are included in a database of potential panel members. Some of them are selected to become part of the panel. All members of the selected households are invited to join. The web panel members complete a questionnaire once a week. This self-administered approach is considered to have the major advantage of allowing respondents to answer questions at their own convenience.

- *The LISS Panel*

The LISS (Longitudinal Internet Studies for the Social Sciences) panel in the Netherlands is a relatively new representative panel of 5,000 Dutch households built in a fashion similar to the CentERpanel. It is also maintained by CentERdata. Recruitment is based on a probability

sample. Respondents of 16 years and older are asked to participate. For a short description of this panel, see Chapter 3, and for more details, see Scherpenzeel (2008).

- *The PAADEL panels*

The PAADEL (Agri-food and Demographic Panels for Lombardy) project is managed by the Center for Statistical Analyses and Interviewing (CASI) of the University of Bergamo and is supported by a grant from the Lombardy region in Italy. This project manages two representative panels with the aim of continuously collecting data related to company innovation in the agricultural sector and food consumer behavior. One panel is for agricultural sector companies, and the other is for households, viewed as consumers.

12.2.2 USE OF WEB PANELS

Web panels can be used to collect data for two different types of research: longitudinal research and cross-sectional research.

In the case of *longitudinal research*, the same group of people is surveyed at different points in time. The group may consist of all panel members or just of a sample of panel members. The aim is to study changes over time. Two different types of surveys are distinguished:

1. A *demographic or profiling survey*. Each active member of the panel is interviewed regularly, for example, every three months. Participants are asked to enter or update their basic demographic characteristics, including an e-mail address. Thus, their demographic profile is kept up-to-date. Among other things, this information is used to correct panel outcomes for the lack of representativity. Almost all web panels have such a demographic component.
2. A *core survey* collects data on certain key market research or socioeconomic issues. These data are required to answer the main research questions of the survey. Core surveys can be conducted longitudinally over many data collection waves. They can be based on the entire web panel or on a sample extracted from it.

In the case of *cross-sectional research*, the state of a population is examined at one point in time. Interest is not in studying changes over time. Usually, the panel is used only once to collect data on a specific topic. For this reason, they are also called *specific surveys*.

To sum up, web panels can satisfy several information needs. A major advantage is that data collected in many surveys can be brought together in one database. All these data may be combined with the demographic information into a very rich database.

EXAMPLE 12.4 Core surveys and specific studies in the LISS panel

Web panels produce a series of data sets corresponding to different research stages or different research topics. Three types of data sets are distinguished:

- Profiling or demographic data sets. This is a collection of background variables on socioeconomic characteristics of household and individuals. In the case of a business panel, the profiling database includes stratification variables such as employment, size of the company, type of economic activity, and location.
- Core studies data sets. These are data sets including variables that are collected in several survey waves.
- Specific studies data sets. These are data sets containing variables on specific topics. They may have been collected once or several times.

The LISS on-line panel (see Example 12.3 in this chapter and Example 9.3 in Chapter 9) is a probability-based web panel. At the recruitment stage, a profiling survey is administered among recruited members. In this way, a profiling database with background variables on education and social stratification is generated. Additionally, 10 core surveys are carried out. They cover the following topics: health, religion and ethnicity, social integration and leisure, family and household, work and schooling, personality, politics and values, as well as economic situation regarding assets, income, and housing. Each of these surveys is longitudinal.

Some special study surveys are single-wave surveys. Examples of topics investigated are expectations for the future, monthly spending during the working-life and retirement, and European values. Finally, an example of a longitudinal survey is one the topics of crime victimization. It is based on a vignette experiment.

12.2.3 WEB PANEL MANAGEMENT

12.2.3.1 Recruitment Steps. Self-selection web panels do not require special recruitment activities using samples from sampling frames. Possible ways of recruiting people for a volunteer web panel were briefly described in Section 12.2.1. The situation is different for probability-based recruitment. To obtain a proper random sample requires a careful and accurate recruitment strategy. The following steps are distinguished:

1. Contacting and inviting potential panel members via a traditional data collection mode, such as telephone, mail, or face-to-face.
2. Administering a profile or welcome survey in order to collect basic demographic information about potential panel members. Different survey

organizations have various ways of collecting such information, but there is at least one survey or data collection step to be completed in order for the recruited respondents to become part of what can be termed the *active panel*. Administering a detailed “enrollment registration survey” allows sufficient information to be acquired to assess the eligibility for specific surveys.

3. Once the active panel is formed from the recruited members, for which a background database is available, a specific survey can select
 - All active panel members.
 - A sample, via random selection.
 - A sample selected on specific qualifying criteria available from the profile interview, such as only people aged 65 years and older.
 - A selection based on specific criteria obtained from screening questions asked of sampled panel members.

In the *initial recruitment* step, eligible and ineligible cases can be determined. The panel manager should try to recruit as many people as possible from among the eligible cases. These respondents will be asked to become members of the web panel.

When a profile survey questionnaire is offered to the eligible cases, the contact phase can have three different outcomes, namely:

- Refusal and break-off.
- Noncontacts.
- Other noninterviews.

EXAMPLE 12.5 Recruitment procedures and sampling strategies

The Knowledge Panel

The basic characteristics of the Knowledge Panel were described in Example 12.3. The survey is administered using a web browser and enables the inclusion of multimedia content. The telephone is used for the first recruitment stage.

The panel is recruited applying list-assisted RDD sampling techniques using a sampling frame containing the entire U.S. telephone population. This sampling frame is updated quarterly. In the recruitment stage, households are selected and every member of the household who is 13 years of age and older becomes part of the web panel. Knowledge Networks’s panel recruitment methodology uses the quality standards established by the best RDD surveys conducted for the federal government. Knowledge Networks excludes only those banks of telephone numbers (consisting of 100 telephone numbers) that have zero directory-

listed phone numbers. Knowledge Networks's telephone numbers are selected from the 1+ banks with an equal probability of selection for each number. Sampling is done without replacement to avoid duplication of households. Having generated an initial list of telephone numbers, the sample preparation system excludes confirmed disconnected and non-residential telephone numbers.

Telephone numbers for which Knowledge Networks is able to recover a valid postal address (approximately 50%) are sent an advance mailing informing the households that they have been selected to participate in the Knowledge Networks Panel. A monetary incentive is included in the advance mailing to encourage cooperation when the interviewer calls.

After the mailing, the telephone recruitment process begins. The numbers called by interviewers consist of all those for which an advance mailing was sent, as well as 50% of the numbers without advance mailing. The advance mailing approach is more cost-effective than the approach without advance mailing. Households are called by telephone over a period of up to 90 days, with at least 15 call attempts in cases where no one answers the phone and 25 call attempts for telephone numbers that are known to be associated with households. Refusals are coded and extensively described.

Once a person has been recruited to the web panel, contact continues via e-mail. After recruitment, the potential households are classified into two groups—those who have Internet access and those who do not. The following actions are carried out:

- a. A Web TV or, more recently, a laptop computer is provided to the non-Internet households. The WebTV, via a telephone modem, makes it possible to browse the web on a TV. Thus, selected households that are not computer users are provided with free hardware and Internet access.
- b. All members receive a welcome survey that teaches them how to navigate and complete web surveys.
- c. Members complete a profile questionnaire to collect basic demographic information about the respondents and households they live in. See the studies by Huggins and Eyerman (2001) and by Pineau, Nukulki, and Tang (2006).

Respondents are prepared for the panel by asking them to complete a household roster. Each household member of age 13 years and older has to provide initial demographic information and background information about prior computer and Internet use. Individuals having completed the profile questionnaire become members of the active panel.

Once an active panel has been recruited and profiled, members become eligible for selection for specific surveys. In most cases, a simple random sample will be selected from the panel. For some specific studies, a stratified random sampling may be drawn using strata that are constructed using the profile data.

Knowledge Networks applies a specific sampling rule that a panel member may be assigned to not more than one survey per week.

For certain studies, the sample is selected after prescreening. For example, suppose that a survey on the clothes-related buying behavior of women is set up. Then sample members will be drawn from a subpanel of the web panel consisting of just females.

The Gallup Panel

The Gallup Organization recruits respondents for the Gallup Panel via RDD. Households willing to participate are sent a “membership packet” by mail with a demographic questionnaire to be completed and mailed back. Household members returning a completed membership packet become part of the active panel. See the studies by Arens and Miller-Steiger (2006), Sayles and Arens (2007), and Tortora (2009). The Gallup Panel collects survey data using different modes (web, mail, telephone, and IVR).

In 2008, Stanford University recruited a national area probability sample of adults and equipped them all with a free laptop computer and high-speed Internet access. This project tested the feasibility of recruiting such a panel to complete monthly surveys for a year. Also in the same year, the American National Election Study (run by the University of Michigan and Stanford University) recruited a web-enabled panel via RDD providing non-Internet households with a WebTV.2.

The LISS panel

Members for the LISS panel of CentERdata are selected from the population register in the Netherlands. A probability sample is used. The next step is to link telephone numbers to the sample addresses. This is not always possible because many people have unlisted telephone numbers.

Recruitment involves three stages:

1. The initial contact, with an invitation to become a member of the panel. This contact is by telephone where possible, and, otherwise, by a face-to-face interview.
2. Obtaining the consent of the respondent to be re-contacted.
3. Obtaining the final approval to become a member of the panel.

By matching data from the recruitment process with variables in the population register, selectivity with respect to certain variables can be identified at all stages. Age and income selectivity are particularly monitored, as well as PC ownership (note that this is possible only during the web questionnaire response phase). Interestingly, there is almost no selectivity with respect to key variables on living conditions.

12.2.3.2 Panel Attrition and Maintenance. Web *panel attrition* is the phenomenon that panel members drop out of the panel in the course of time. Because of attrition, web panel maintenance is necessary. Maintenance has to deal with efficiency, replacement, and eligibility concerns so as to eliminate attrition effects. A distinction has to be made between self-selection panels and probability-based panels. Attrition is less of concern for self-selection panels. Recruitment does not require special action. For probability-based web panels, representativity is an important issue. It is important the panel keeps its proper composition. To correct the effects of attrition, additional sampling of specific groups in the population may be required. This increases the costs of panel maintenance.

A first challenge is to maintain the size of the active panel at the desired level. Without any maintenance, the size of the panel would gradually decrease. There are always members that leave the panel, for either voluntary or involuntary reasons. Various rules for panel size maintenance can be adopted, and each organization has its own rules. See, for example, the studies by Postoaica (2006a, 2006b). Possible reasons for active members to leave the panel are as follows:

- Members are forced to leave the panel after a specific period of time. Panel organizations may have rules about the maximum period of membership. This is to avoid panel condition (i.e., members change their attitude or behavior as result of being in the panel for a long time). There can be other attrition rules. For example, members are removed from the Gallup Panel if they do not respond to six consecutive surveys (Sayles and Arens, 2007).
- Some active members may become temporarily inactive for a variety of reasons (holiday, illness, and so on). This may also happen in the case of business panels (e.g., temporary inactivity).
- The panel organization may have a system to spread the response burden. It is not uncommon for such organizations to have rules that members should participate in not more than a given number of surveys in a specific period. If a members reaches this burden threshold, they will be assigned the status “not available for sampling.” Many on-line panels have such rules to avoid overburdening members. For example, the Knowledge Networks panel has a sampling rule of not assigning more than one survey per week to its members.
- If a panel survey is run for a specific client, certain restrictions on the eligibility criteria may be defined by the client. For example, the client may not want sampled members to have completed a recent survey on a similar topic or on a topic that might influence the survey outcome. For such quarantine-like criteria, see the discussion by Postoaica (2006a).

To guarantee a constant inflow of new members in the web panel, special recruitment strategies may be implemented. One is the use of *incentives*. Several web panels offer incentives to people to become or to stay a member. Different types of incentives can be used: monetary incentives (even if the amount of

money offered is very small, such as just a few dollars), charity donations, lottery tickets, product gifts, and so on.

Several experiments on the effect of incentives on panel participation and data quality, and on testing the impact of alternative forms of incentives, can be found in the literature. No general conclusions can be drawn. In general, no relevant bias effect is found. There are some indications of effects on estimates of background variables.

As an example, Ipsos (one of the largest market research companies in the world) uses incentives. This company currently conducts more than six million surveys, in more than 100 countries, each year). Participation is rewarded with points that can be redeemed for cash or prizes, and with other opportunities to win great prizes for completing on-line surveys.

12.2.4 RESPONSE RATES

The response rate is an important indicator for the quality of single, cross-sectional surveys because low response rates present a serious risk of unreliable estimates. See, for example, the study by Bethlehem, Cobben, and Schouten, (2011). Use of response rates in the context of panels can be less meaningful. If the response is high in one wave of a panel, but it was low during the recruitment phase, the high-wave response rate does not guarantee the bias of the estimates to be small. A good sample from a bad panel does not produce reliable estimates.

There is a lot of literature focusing on the question of the usefulness and meaning of response rates. See, for example, the reports by the American Association of Public Opinion Research (AAPOR, 2006) and Eysenbach (2004). Vonk, Ossenbruggen, and Willems (2006) states that the “response percentage does not indicate sample or panel quality. It reflects a panel business strategy. The response rate is an indication of the level of efficiency of the panel provider” (page 20). In some cases, a different terminology has been suggested, like “view rate,” “participation rate,” or “completion rate.” Taking different literature and ideas into account, it can be acknowledged that response rates alone are not a very good indicator of the nonresponse bias and thus of web survey and web panel quality. See the studies by Bethlehem, Cobbin, and Schouten (2011), Groves (2006), and Groves and Peytcheva (2008).

A basic characteristic of self-selection recruitment for a panel is that everyone can volunteer to participate. There is no sampling frame. Although a target population may have been defined, it is not always clear whether each respondent belongs to this target population. Therefore, it is not possible to compute the exact response rates for these panels. This problem is discussed by Fricker and Schonlau (2002), Schonlau, Fricker, and Elliott (2002), and the AAPOR (2006). Only completion rates can be calculated for self-selection panels (see Section 12.2.4.4).

For panels based on probability sample recruitment, the cumulative or multiplicative response rate can be used. This computation takes into account the response rate of the recruitment phase and the response rates of the subsequent surveys. See the studies by Huggins and Eyerman (2001),

Schlengen et al. (2002), Tourangeau (2003), the Office of Management and Budget (2006), Schonlau, Van Soest, Kaypten and Couper (2009), and Couper (2007). Nevertheless, even in the case of probability-based web panels, the computation of response rates can be a difficult task if a possibly complex procedure has to be taken into account.

No rigorous and well-accepted terminology and definition of response rate in web panels exists. An example of the definition of response rate and of completion rate is given by the Interactive Market Research Organization (2006):

- The *response rate* is “based on the people who have accepted the invitation to the survey and have started to complete the survey. Even if they are disqualified during screening, the attempt qualifies as a response.”
- The *completion rate* “is calculated as the proportion of those who have started, qualified, and then completed the survey.”

The terminology used by different companies and organizations varies, and often the same term is used with a different meaning. Alternative concepts, evaluation criteria, and rates are also provided within different contexts. Each guidelines quoted in this section uses its own terminology.

EXAMPLE 12.6 ESOMAR evaluation criteria

ESOMAR (2005) suggests the following possible indicators for use in evaluation criteria:

- Percentage response based on the total amount of invited individuals (% of full number) per sample drawn (country, questionnaire).
- Percentage of questionnaires opened.
- Percentage of questionnaires completed (including screen-out).
- Percentage in the target group (based on quotas).
- Percentage validated (the balance is cleaned up, if applicable).

Another example of an alternative evaluation criterion is the *initial response rate* (i.e., the percentage of people who initially agree (for example, on the telephone) to become a member of the panel and complete a second in-house interview). This definition has been proposed by Saris and applied in the analysis of the prerecruited probability-based Dutch Telepanel. These concepts are derived from computer-assisted panel research (Saris, 1991, 1998).

One has to bear in mind that, although a variety of criteria and rate definitions exist, it is of vital importance to use standardized concepts and definitions when comparing the performance and quality of different panels.

Before introducing and describing several indicators, the following points are stressed:

- There are different stages in building a web panel. In each stage, the focus is on different kinds of response rates and indicators.
- Cumulative response rates over consecutive stages can be used to evaluate the performance and quality of studies using probability-based panels.
- Computation of indicators should be restricted to only the active part of the panel. For the computation of response rates over time, the availability of panel members at a given point in time must be taken into account.

Callegaro and DiSogra (2008) propose a systematic framework of concepts and indicators related to response in panels. The basic idea is that the computation of response rates for a probability-based panel has to take into account all steps in the recruitment and maintenance. At each step, different response rates can be computed and each of them provides insight into different aspects of the quality and success of the survey. An example of the response rates for each of the different stages of recruitment in the Dutch CentERpanel can be found in the study by Sikkel and Hoogendoorn (2008).

Several indicators are introduced and described in the following subsections: the recruitment rate, the profile rate, the attrition rate, the completion rate, the break-off rate, the cumulative rate, the screening completion rate, and the study-specific rate.

12.2.4.1 The Recruitment Rate. Recruitment is the first step in building up a panel. Recruitment is also required to maintain the panel. Recruitment is only well defined for probability-based panels. It comes down to selecting a sample from the target population and to inviting the selected individuals to become a member of the panel. The recruitment rate is an indicator of the success of this activity.

If a sample of households is selected, the researcher has a choice to just invite one (randomly selected) member per household or to invite all eligible members of the household. Different choices for selection procedures may lead to different selection probabilities of individuals. For example, if just one person per household is selected, persons in larger households have a smaller probability of selection. The proper selection probabilities should be taken into account when computing estimates.

The recruitment rate cannot be computed for self-selection panels. The reason is that the target population is undefined. It is also unclear how many people are invited to participate as it is unknown how many people see the invitation.

The *recruitment rate* (RECR) (see Callegaro and DiSogra, 2008) is defined by

$$(12.1) \quad \text{RECR} = \frac{\text{IC}}{\text{IC} + (\text{R} + \text{NC} + \text{O}) + e(\text{UH} + \text{UO})},$$

where

- IC = number of initial consent cases
- R = cases directly and actively refusing
- NC = noncontacts
- O = other cases
- UH = unknown if household is occupied
- UO = unknown other
- e = estimated proportion of cases of unknown eligibility that are eligible

The estimated proportion of eligible cases among those of unknown eligibility cannot be computed in practice, because information with respect to UH and UO is usually not available. If panel members are recruited during the first contact with the household, there are no separate recruitment rates. Instead, both steps are combined and one indicator is used; see the study by Couper et al. (2007).

If first contact and recruitment for the panel are separate steps, and thus take place at two different points in time (see Arens and Miller-Steiger (2006)), then the RECR represents only the second step (i.e., the consent to join the web panel).

If all eligible members of the household are recruited, the RECR can be computed at either the household or the person level. Note that.

- *At the household level*, the total number of eligible households is in the denominator. Each household must have at least one potentially eligible member to be recruited.
- *At the person level*, the total number of eligible persons across all households needs to be known. The denominator describes all eligible persons, and the numerator refers to all recruited persons. The factor e would then be a multiplier that gives the estimated number of eligible persons expected from the number of “unknown” or “other” households.

The RECR computed at a household level is the same as the RECR computed at a person level if a within-household selection at the recruitment stage is applied where only one member per household is recruited for the web panel.

When several members per household have been recruited, but a member-level sample for a given study is drawn in which only one random member per household is selected (among all eligible members, if there is an eligibility criterion) and no substitutions are allowed, a household-level RECR measure can be used because it is similar to recruiting only one person per household.

EXAMPLE 12.7 Computing recruitment rate

Suppose that in a web panel 11,420 households were eligible for panel selection. In total, these households contain 26,490 persons. During the recruitment stage, the results from Table 12.2 are obtained.

TABLE 12.2 Household-and person-level recruitment data for a fictitious web panel

Household level		Person level	
IC	10,200	IC	23,460
R	400	R	1,200
NC	370	NC	800
O	200	O	500
UH	150	UH	380
UO	100	UO	150
Total eligible	11,420	Total eligible	26,490
Estimated proportion of unknown eligibility that are eligible	0.01	Estimated proportion of unknown eligibility that are eligible	0.06

The recruitment rate at household level is computed as follow:

$$\text{RECR} = \frac{10,200}{10,200 + (400 + 370 + 200) + 0.01(150 + 200)} = 0.91.$$

Thus, the recruitment rate is 91% of the eligible population. The recruitment rate at the person level can be computed analogously.

12.2.4.2 The Profile Rate. As described in Section 12.2.1, the second step of the recruitment process is to send an initial profile survey to all those who have agreed to become a member. Alternatively, the survey organization can redirect recruited respondents to an on-line registration page that functions as the profile survey. By responding to this short survey, respondents become part of the active panel.

There may be a difference in probability-based panels between the number of initially recruited panel members and the number active panel members. The cause is that recruited candidates could choose not to complete their profile survey and therefore drop out before being registered as panel members. To account correctly for this initial drop-out effect, an indicator based on the active web panel should be computed. This indicator is called the profile rate PROR.

No unknown eligibility or ineligible cases exist at this stage because these will have previously been removed. The *profile rate* is defined by:

$$(12.2) \quad \text{PROR} = \frac{(I + P)}{(I + P) + (R + \text{NC} + O)},$$

where

- I = number of complete cases in the profile survey
- P = number of partially complete cases in the profile survey
- R = refusals (direct or active)
- NC = noncontacts (including passive refusals)
- O = other cases

Noncontacts (NC) means that there is no reply from a respondent in the profile survey (Couper et al., 2007). Because these people were contacted at the recruitment stage, they could be contacted again to record their reasons for not completing the profile survey or, if relevant, to confirm that the e-mail invitation for the profile survey actually reached them. Note that noncontacts can be a form of passive refusal behavior. This phenomenon, if included in the NC, could contribute to membership bias.

As with the RECR, the PROR can be computed either at the household level or at the person level. If the sample for a given study is limited to selecting only one random panel member per household, the household-level and person-level *PROR*s are equal.

For self-selection panels, the profile rate has a different meaning because determining the number of refusals and noncontacts is not as straightforward as it is for probability-based panels. This information is unknown for single opt-in panels. When a double opt-in procedure is used, the profile rate may be computed as the number of people who confirmed at their second opt-in opportunity over the total number who initially opted in.

EXAMPLE 12.8 Computing the profile rate

In Example 12.7, 10,200 households had given their initial consent. Suppose they all respond in the profile survey, either totally or partially. In practice this almost never happens. Assume there was a complete response in 7,000 case and a partial response in 3,200 cases. The profile rate is now

$$\text{PROR} = \frac{(7000 + 3200)}{(7000 + 3200) + (400 + 370 + 200)} = 0.91.$$

This rate is exactly the same as the recruitment rate. In practice, however, some households that initially agree to become a member do not respond

in the profile survey. Suppose that only 7,800 households complete the survey (5,500 complete responses and 2,300 partial responses). Now the profile rate is:

$$\text{PROR} = \frac{(5500 + 2300)}{(5500 + 2300) + (400 + 370 + 200)} = 0.89.$$

12.2.4.3 The Absorption Rate. Once the panel is in operation, members of the panel are contacted for participation in specific surveys. This is usually done by means of e-mail. Not every selected member will respond. Pratesi, Manfreda, Biffignandi, and Vehovar (2004) distinguish four steps in the web survey participation process. Nonresponse may occur in each step:

1. *Sending an e-mail invitation.* In this contact step, some e-mails may not reach the members because of technical problems. This may happen in approximately 10% of the cases. For example, an e-mail invitation might end up in a spam filter and be deleted or moved to a spam map. In this case, the e-mail seems not to be “absorbed” by the panel member. There is no feedback indicating that the e-mail was treated as spam. This prevents the researcher from successfully contacting an invited person. This is a problem. The researcher has no means of measuring the relevance or magnitude of this problem. Other examples of e-mails not reaching the selected person are a wrong e-mail address, a full mailbox, or a network error.
2. *Access to the introductory page of the web survey.* The contacted person accesses the questionnaire. This action does not yet imply survey participation.
3. *Start of questionnaire completion.* The respondent starts answering questions. This means at least partial completion. There is always a risk that some questions are skipped, or that completion of the questionnaire interrupted.
4. *Completion of the survey questionnaire.* At this step, participation is completed, although some answers may have been skipped in the process.

To take possible absorption effects into account, Lozar Manfreda and Vehovar (2002) propose to compute the so-called absorption rate. According to Callegaro and DiSogra (2008), the *absorption rate* is defined as:

$$(12.3) \quad \text{ABSR} = \frac{\text{EI} - \text{BB} - \text{NET}}{\text{EI}},$$

where

- EI = number of e-mail invitations sent.
- BB = number of undeliverable e-mail invitations (bounce back).
- NET = network error – undeliverable e-mails.

The absorption rate is 1 (i.e., all e-mails are absorbed) if every selected member actually receives the e-mail. Therefore, this rate can be considered an indicator of the quality of the e-mail list of the sampled web panel members. It is a proxy indicator of how many units receive the e-mail because, as stated, it is impossible to check whether e-mails get lost in the system. In other words, it is impossible to determine exactly the number of selected members that received an initial e-mail.

12.2.4.4 The Completion Rate. The completion rate is the proportion of selected and invited eligible members who completed a specific web survey. Thus, this rate reflects the success of a specific study. The *completion rate* is defined by:

$$(12.4) \quad \text{COMR} = \frac{(I + P)}{(I + P) + (R + \text{NC} + O)},$$

where

- I = number of complete cases in the survey
- P = number of partially complete cases in the survey
- R = refusals (direct or active)
- NC = noncontacts (including passive refusals)
- O = other cases

The variables are the same as those found in the profile rate. The difference is that these variables are not determined for the profile survey, but for a specific survey.

As stated in Section 12.2.4.2, the completion rate can also be computed for volunteer opt-in panels, but with some modifications.

12.2.4.5 The Break-Off Rate. In expressions (12.2) and (12.4), the partially completed interviews (P) are considered successful cases. However, in some situations an incomplete questionnaire is not considered to be an acceptable case.

To measure the extent of this phenomena, the concept of “break-off” can be introduced. Break-off means that the questionnaire was started but not finished. The *break-off rate* is defined by

$$(12.5) \quad \text{BOR} = \frac{\text{BO}}{I + P + \text{BO}},$$

where

- I = number of complete cases in the survey.
- P = number of partially complete cases that are considered successful.
- BO = number of break-offs i.e., the number of unsuccessful partially completed cases (according to the criteria of the researcher).

12.2.4.6 The Screening Completion Rate and Study-Specific Rate. For each specific study, the target population need not coincide with the panel. Therefore, the researcher has to select eligible respondents. This is called *screening*. There are two options:

1. The screening process can be based on previously collected information (in the profile survey or in another specific survey).
2. If no screening variables are available, an e-mail invitation can be sent to all (or a large subset of) panel members. The first questions of the survey must aim at assessing the eligibility of the respondents. Routing instructions will see to it that only eligible respondents answer the remaining questions.

Bearing in mind the above-mentioned options, two indicators can be computed: the screening completion rate (S_COMP) and the study-specific eligibility rate (S_ELIG). The *screening completion rate* is defined by:

$$(12.6) \quad S_COMP = \frac{SCQ + SCNQ}{INV},$$

where

- SCQ = the number of people who were successfully screened and qualified for the study.
- $SCNQ$ = the number of people who were successfully screened and did not qualify for the study.
- INV = the number of survey invitations sent.

A potential problem is that nonresponse may hamper correct interpretation of this rate. If selected persons do not answer the screening questions, it cannot be established whether they qualify for the survey.

The *study-specific eligibility rate* is defined by

$$(12.7) \quad S_ELIG = \frac{SCQ}{SCQ + SCNQ},$$

where

- SCQ = the number of people who were successfully screened and qualified for the study.
- $SCNQ$ = the number of people who were successfully screened and did not qualify for the study.

12.2.4.7 Cumulative Response Rates. In the case of probability-based panels, cumulative response rates can be computed. These indicators take into account what happens in the different steps of a survey, from panel recruitment

to response in a specific study. These cumulative response rates are obtained by multiplying rates that have been obtained for each step in the process.

Note that an often adopted approach to panel maintenance is to add new cohorts of members to the current panel. Consequently, the response rates should be computed separately for each cohort. As a next step, the response rates for the panel as a whole are obtained by taking a weighted average of the cohort response rates. Here the focus is on computing the response rate for a single cohort. Two types of cumulative response rates can be calculated. They are called the cumulative response rate 1 and the cumulative response rate 2.

The cumulative Response Rate 1 (CUMRR1) is defined by:

$$(12.8) \quad \text{CUMRR1} = \text{RECR} \times \text{PROR} \times \text{COMR}.$$

This response rate reflects the percentage of cases left over after nonresponse in the recruitment phase, nonresponse in the profile survey, and nonresponse in the specific web survey.

The cumulative response rate 2 introduces a fourth component: the retention rate. The *retention rate* (RETR) is the proportion of an original cohort that remains in the active panel at the time the sample for the specific survey is drawn. Therefore, cumulative response rate 2 is defined with reference to a specific cohort. For a given cohort, this indicator is obtained by multiplying the cumulative response rate 1 by the retention rate, thus, *cumulative response rate 2* is defined by:

$$(12.9) \quad \text{CUMRR2} = \text{RECR} \times \text{PROR} \times \text{RETR} \times \text{COMR}.$$

12.2.4.8 The Attrition Rate. The concept of attrition was discussed in Section 12.2.3.2. It is defined as the proportion of active panel members that drops out of the panel in a specific time period. Attrition has an effect on cumulative response rates. The overall representativeness of the web panel also can be affected by the differential attrition rates. For example, subgroups with higher attrition rates than other subgroups will become underrepresented in the panel in the course of time. Therefore, these groups also will be underrepresented in specific surveys.

DiSogra et al. (2007) and Sayles and Arens (2007) state the importance of studying differential attrition rates for subgroups of the population. The attrition rate of group a from month t to month $t + 1$ is defined by:

$$(12.10) \quad \text{ATTR}_{M_t} = \frac{\text{Cohort}_a@Time_t - \text{Cohort}_a@Time_{t+1}}{\text{Cohort}_a@Time_t},$$

where

- $\text{Cohort}_a@Time_t$ = the size of the specific cohort at time (month) t
- $\text{Cohort}_a@Time_{t+1}$ = the size of the specific cohort at time (month) $t + 1$

Roughly speaking, attrition is measured by counting how many recruits stay in the web panel month after month (Clinton, 2001).

Web panel attrition is an informative indicator because, when it is high, it “could be the result of placing surveys that are too long or poor in question design” (European Society for Opinion and Marketing Research, 2008). Monitoring attrition is also crucial in assessing the representativeness of any panel, especially since attrition is rarely equal across all demographic subgroups.

12.2.4.9 Which Rate Should be Considered when Evaluating a Web Panel? Several indicators for the quality of probability-based web panels were presented in the previous sections. All these indicators may help to judge the reliability of the outcomes of a web panel. It is important to include indicators like the recruitment rate, the profile rate, the study-specific survey completion rate, and the final cumulative response rate (CUMRR1) in the survey report.

Other factors may play a role in the quality of the outcomes. Some of them are as follows:

- Field period (starting and closing dates, and length of the fieldwork period length).
- Number of reminders sent and follow-up mode (email, letter, IVR call, or personal call).
- Use of incentives.

For example, the use of incentives generally increases the response rates, but some literature suggests it may not help to improve the composition of the response, resulting in a larger bias. Another example is shortening the fieldwork period. It may be attractive to have timely information, but if this leads to an overrepresentation of frequent e-mail users and early respondents, this may also be the cause of a bias.

An obvious indicator for self-selection panels is the completion rate. This indicator can be interpreted as the respondent’s interest in the survey and/or the ability of the survey organization to maximize cooperation. Note that it is possible to increase the survey completion rate by just selecting the most cooperative web panel members. Unfortunately this can seriously affect the composition of the sample. Vonk, Van Ossenbruggen, and Willems (2006) conducted an experiment in which the same survey was carried out at the same time by 19 different self-selection web panels in the Netherlands. The completion rate turned out to vary between 8% and 77%. Thus, completion rate is an extremely volatile indicator.

The absorption rate is an interesting indicator for measuring ability in managing and updating the web panel database, whereas the break-off rate could suggest that problems exist either with respect to the design of the questionnaire (e.g., too long or too boring) or to technical aspects during the survey administration (e.g., streaming media or animations that may “break” a survey at some point).

The absorption rates and the break-off rates should be reported for both probability-based and volunteer panel research.

12.2.5 REPRESENTATIVITY

Many web panels are large. This sometimes leads to claims that therefore they are representative and the quality of the survey results is high. These claims are often not justified. The concept of representativity and its relationship to web panels is discussed in this section.

The concept of “representativity” plays a crucial role in the discussion about the foundations of survey sampling. This concept is often used in survey research, but usually it is not clear what it means. Kruskal and Mosteller (1979a, 1979b, 1979c) present an extensive overview of what representative is supposed to mean in the nonscientific literature, scientific literature excluding statistics, and current statistical literature. They found the following meanings for “representative sampling”:

- General acclaim for data.
- Absence of selective forces.
- Miniature of the population.
- Typical or ideal case(s).
- Coverage of the population.
- A vague term, to be made precise.
- Representative sampling as a specific sampling method.
- As permitting good estimation.
- Good enough for a particular purpose.

Kruskal and Mosteller (1979b) recommend not using the word “representative” but instead to specify what one means. The problem is that both in probability sampling and in other forms of sampling, claims are made that samples are representative, often with different meanings and sometimes with no concrete meaning at all besides conveying a vague sense of good quality.

In this chapter, the term “representative” is used to indicate that a sample is *representative with respect to a variable*. It means that the distribution of the variable in the sample is the same as its distribution in the population. The idea is that if a sample is representative with respect to many auxiliary variables, the hope is it will also be representative with respect to the target variables of the survey, therefore, allowing unbiased estimation of population characteristics. In case a sample is not representative with respect to several auxiliary variables, weighting adjustment can be applied in attempt to improve representativity.

Bethlehem and Stoop (2007) discuss the frequent misunderstanding about on-line research that large numbers make a sample better. Couper (2000) comments on the claims of a self-selected online survey: “We received more than 50,000 responses – twice the minimum required for scientific validity,–” where as the survey did not yield a random sample and the selection probabilities are unknown. Not surprisingly, despite the large number of respondents, they did not resemble the U.S. population on several key indicators.

Dillman and Bowker (2001) express a similar opinion about on-line surveys:

Conductors of such surveys have in effect been seduced by the hope that large numbers, a traditional indicator of a high quality survey (because of low sampling error), will compensate in some undefined way for whatever coverage and nonresponse problems that might exist. Large numbers of volunteer respondents, by themselves, have no meaning. Ignoring the need to define survey populations, select probability samples, and obtain high response rates, together provide a major threat to the validity of web surveys.

Couper (2001) also pointed to the misguided assumption that large samples necessarily mean more valid outcomes. Only in the case of probability samples does an increase of sample size lead to an increase of precision. Otherwise, no inference about the underlying population is possible, and larger samples do not necessarily produce better estimates than smaller samples.

Large web panels and web surveys have the advantage that specific subgroups can be identified. Information about such groups may be difficult to obtain in traditional surveys because few people belong to these groups, they are hard to identify, or unlikely to participate in surveys. The underlying assumption is that the elderly single women, low educated, ethnic minorities, or other usually underrepresented groups who participate in a web survey are similar to people with the same characteristics but who do not participate. In some cases, this might be a likely assumption; in others definitely not; and in most cases, it will be difficult to test.

An additional caveat is that self-selection in web panels may require heavy weighting adjustment because of vastly varying participation probabilities. Because of large weights, the effective sample size is likely to be much smaller than the number of participants in a survey (Duffy, Smith, Terhanian and Bremer, 2005).

When analyzing the data collected by means of a web panel, it is usually implicitly assumed that the panel is a small copy (a miniature) of the population it came from. So relationships found in the panel are not an artifact of the sample but are identical to relationships that exist in the whole population. However, even if there are a large number of respondents, they might not reflect the population structure. For example, Faas and Schoen (2006) have studied whether participants in self-selection web surveys represent Internet users in general. They conclude this is not the case. Hence, the conclusion must be that self-selection web surveys do not yield results representative of Internet users in general (in terms of marginal distributions of variables or in terms of relationships between variables). If the aim is to have survey results that are representative of Internet users, one must carefully select a sample of such users.

In considering the effects of the lack of representativity and sample size, it is important to distinguish the various types of sample selection mechanisms:

- *Probability sampling.* The sample is selected from the population using some kind of random selection mechanism. Each element in the

population must have a nonzero probability of selection, and all selection probabilities must be known. The simplest form is a simple random sample, in which all elements have the same selection probability. Correct inference from the sample to the population is possible. Unbiased estimates can be computed. The accuracy of estimates increases as the sample size increases.

- *Nonprobability sampling.* This comes in many forms. See, for example, the studies by Kalton (1987), Couper (2000), and Schonlau, Fricker, and Elliott (2002). Some forms are as follows:
 - *Convenience sampling.* Elements are drawn for such a sample because of their convenient accessibility or proximity to the researcher. Convenience sampling is fast, simple, and cheap. Self-selection samples can be considered a form of convenience sampling.
 - *Purposive sampling.* The researcher picks units that are “representative” in a subjective way. The sample is selected such that its characteristics resemble the characteristics of the population. Purposive sampling is similar to the “Representative Method” proposed by Kiaer (1895). See also Chapter 1. He constructed fairly large samples that were a miniature of the population with respect to the regional distribution. Because selection was not based on probability sampling, the accuracy of the estimates could not be computed.
 - *Quota sampling.* The population is divided into strata. The size of the strata is supposed to be known. A sample of predefined size (quota) is selected in each stratum. It is left to the judgment of interviewers to pick elements in a stratum. The basic assumption is made that the probability of being available for an interview is the same for each element within stratum. See the study by Sudman (1966). The problem is that often this assumption is not satisfied. For example, people with a larger probability of being at home may differ from those that are frequently not at home. Also for quota sampling, it is not possible to compute selection probabilities.

Because of the problems of nonprobability sampling described above, it is not possible to apply probability theory. This prohibits making proper inference to the target population of the panel or survey. It is also not possible to compute the indicators for survey data quality like response rates.

One may wonder whether it would help to apply some form of adjustment weighting in an attempt to make the sample at least representative with respect to several auxiliary variables. Chapter 10 shows there is absolutely no guarantee this solves the problems. If there are no strong relationships among the target variables of the survey, the selection mechanism, and the weighting variables, biases will remain.

Some alternative approaches have been proposed such as the use of propensity scores. See Chapter 11 and the studies by Lee (2006), Lee and Valliant (2009), Schonlau, Van Soest, Kapteyn, and Couper (2007), and Loosveldt and Sonck (2008). This also does not offer a solution for nonprobability sampling.

EXAMPLE 12.9 Quota sampling in election polls

George Gallup was one of the first to apply quota sampling. With his opinion poll, he predicted the winner of the 1936 presidential election to be Franklin Roosevelt.

Hundreds of interviewers all over the United States were given quotas for different types of respondents: so many middle-class urban women, so many lower-class rural men, and so on. With a sample of only 3,000 respondents, he could make a better prediction than the *Literary Digest* poll. This magazine had a sample of 10 million respondents. They were recruited from vehicle registrations and telephone directories. The sample was not representative because at that time cars and telephones were more often owned by the middle and upper classes, and they preferred the Republican candidate Alf Landon. The poorer people favored Roosevelt. As a result, *Literary Digest* predicted Landon to be the winner. Gallup turned out to be right: Roosevelt was reelected.

However, quota sampling failed to predict the winner in the 1948 elections. George Gallup incorrectly predicted the victory of New York Governor Thomas Dewey over President Harry Truman. Quota sampling turned out to be ineffective. As a result, use of quota sampling was abandoned as a valid method for opinion polls. Starting in 1956, it was replaced by probability sampling.

Groves (1989) states that a statistical paradigm in which sample size and response rates affect the precision and accuracy of estimates is rooted in probability sampling. Therefore, this paradigm cannot be transferred to nonprobability sampling.

Web panels are often claimed to have high response rates. Such claims are based on the high response rates for samples selected from the panel for specific surveys. This ignores the fact that the response rate in the recruitment phase of the panel may have been very low, and as a result, the estimates using the survey data may substantially be biased.

When initial recruitment is based on self-selection or when initial nonresponse is high, high response rates in a specific survey may hide a wide range of other survey errors. For an overview of survey errors, see the study by Bethlehem (2009).

It is always possible in a self-selection panel to generate a high response rate by selecting only those members who always participate when they are invited. This increases the sample size, but it is unlikely to improve survey estimates, as this special group might produce highly biased results. This shows again that high response rates in a panel will generally not be able to compensate for low initial cooperation or an unknown selection bias. Defining the response rate as the response probabilities of willing respondents or boosting response rates by preselecting the most cooperative panel members makes response rates difficult

to compare with those of probability samples. For this reason, response rates in nonprobability samples do not have the same meaning as those in probability samples.

Morton-Williams (1993) shows the claims that panels can represent the population or can be made representative by weighting adjustment based on two assumptions: (1) the behavior and attitudes to be measured are related primarily to the variables used as quota controls, and (2) they are not associated, independently of these controls, with factors underlying nonresponse or with the characteristics of those likely to require more than one call to obtain an interview.

Panels become richer in the course of time, as more and more data are added that are collected in each specific survey. This information may help to explore the possible reasons why some panel members participate in a specific survey, and why other panel members do not. However, this rich data set will not help much in investigating the reasons for becoming a panel member, as these data are only available for panel members and not for nonmembers.

Response rates are becoming more and more of a major concern for web panels as well. As an illustration, Schillewaert, Verhaeghe, Weijters, and De Wolf (2006) reported low response rates for various methods of data collection on a comparative study. These rates are reproduced in Table 12.3.

In the past, the structure and composition of the panel was determined in advance (as in quota samples) and weighting for underrepresented groups (attributed to sampling errors, undercoverage, and nonresponse) was considered unnecessary for a sample from a panel. Because of decreasing response rates, control over the final composition of the survey sample is more and more difficult and can only be achieved by taking into account unequal response probabilities in many different groups, based on information about earlier participation in similar surveys. This comes down to assuming that nonresponding panel members in a specific group are similar to respondents.

Response rates are not affected by the topic of the survey because the decision to participate in a panel is a general one and does not depend on future topics. Response rates in individual surveys are still high. However, with declining response rates in web panels, interest in the topic may again have become an important determinant of survey participation. Of course, in web surveys based on self-selection, the topic of the survey is likely to be the most important determinant of participation, possibly resulting in highly biased results.

According to Stoop (2005), acceptable response rates are still possible. In a study on the relationship between nonresponse rates and nonresponse bias,

TABLE 12.3 Response rates by data collection mode and sampling technique

Data collection mode	Sampling technique	Response rate
Face-to-face	Random walk	54%
Web	Self-selection	21%
Mail	Random	15%
Telephone	Random	12%

Groves (2006) states that, despite low response rates, probability sampling retains the value of unbiased sampling procedures from well-defined sampling frames. The coverage error of well-defined sampling frames can be evaluated relative to a desired target population, prior to the survey being launched. Probability sampling from a frame permits use of auxiliary variables in the frame to improve the estimation. Self-selection panels do not have these advantages. Bethlehem (2009) shows that in the worst case, bias in self-selection surveys is much larger than in probability-based surveys.

A sensible thing to do would be to consider building a panel based on random sampling and to include both the Internet and the non-Internet population. See the study by Scherpenzeel (2008) for an example of such a panel. A panel that conforms to strict methodological specifications can function as a starting point for a wide range of cross-sectional and longitudinal studies.

A final recommendation when using self-selection web panels is to follow the advice of Fowler (2002) and be transparent. If a researcher decides to use a nonprobability sample, readers should be told how the sample was drawn, the fact that it is likely to be biased in the direction of availability and willingness to be interviewed, and that the normal assumptions for calculating sampling errors do not apply. This may help to avoid the results of nonprobability samples to be seriously misrepresented, thereby contributing to a loss of credibility of social science research.

Surveys can be conducted much more quickly on a self-selection panel than for probability-based surveys. In face-to-face surveys and telephone surveys, several attempts may have to be made to contact sample persons and to convert initial refusers. In mail surveys, questionnaires have to be printed and sent to sample persons, followed by reminders. Fieldwork may take weeks or even months. In on-line surveys, where the emphasis is more on mass than on response rates, the preparation takes less time and fieldwork takes far less time. There are examples of web panels where surveys are designed, carried out, and analyzed within a day. According to Day, Risk, Koo, and Martin (2006), the first 12 to 24 hours are the most important in any on-line project, with approximately two thirds of panelists responding in this period.

The difference in turnaround times reflects the different aims of traditional surveys and on-line research. In the first case, a lot of effort should be devoted to obtaining the participation of every sample member. This is necessary for producing reliable and accurate estimates. However, this is at the expense of a longer fieldwork period and increased costs. In the second case, it is possible to collect information on a current issue within a very short time if the focus is less on minimizing survey errors and more on the number of respondents, speed, and low costs. These two different survey paradigms are hard to compare because the aims and strategies differ.

The short turnaround time is one of the great advantages of on-line research. It contrasts sharply with the lengthy periods and much greater effort spent in contacting hard-to-contact sample persons in face-to-face surveys (see, e.g., Lynn, Clarke, Martin, and Sturgis, 2002; Stoop, 2005) to enhance response rates and reduce nonresponse bias.

Web surveys and surveys based on samples from web panels are usually much less expensive than random samples. In the former case, most costs are fixed costs, supplemented with some minor costs for respondents' remuneration. As shown, however, the theoretical underpinning of traditional survey research and online research seem to differ substantially. Costs are very important, but costs should always be considered against the background of the purpose of a survey. Almost half a century ago, Deming (1960) said about this: "cost has no measure without a measure of quality, and there is no way to appraise objectively the quality of a (quota) sample as there is with a probability sample."

12.3 Application

Using the Internet as a means for conducting cross-national research offers many challenges and opportunities. The European *WageIndicator Web Survey* is a survey that makes use of the Internet as a means to collect data on wages. It is an international web-based operation, which provides free information on wages. In return for this information, website visitors are invited to complete a web survey, comprising an internationally comparable questionnaire on wages and work.

This survey, which is accessible to all participating countries, 24 hours a day and seven days a week, produces novel data on wages, labor standards, working conditions, and other work-related issues that are available for cross-national research.

WageIndicator is owned by the WageIndicator Foundation, a Dutch non-profit organization. The project was launched in the Netherlands in the year 2000, and it has been quickly enlarged and consolidated as a result of the interest at the European level it was awarded a three-year grant by the European Commission (6th Framework Programme) for the Work Life Web (WOLIWEB) project. Collaboration has been extended to many countries such as Belgium, Denmark, Germany, Finland, Italy, Poland, Spain, the United Kingdom, Hungary, Brazil, India, South Africa, South Korea, Argentina, Mexico, and the United States. See also www.WageIndicator.org.

The WageIndicator survey consists of a questionnaire aimed at collecting information on wages and working conditions. More specifically, the questionnaire is divided into six sections: occupation, place of work, employment history, working hours, employment contract and salary, as well as personal questions.

The target population of the web survey is the labor force. On the whole, the data set contains more than 500 variables. It constitutes the first WorldWide Web survey aimed at gathering data from different countries in a consistent and uniform manner.

Sample recruitment is based on self-selection. There is a self-selection bias, and therefore, the data are not representative for the whole population. Some studies are in progress to explore the selection bias in this panel and to find methods for correcting estimates. For example, an analysis based on the German (*Lohnspiegel*) and Dutch (*Loonwijzer*) data sets from the *WageIndicator Survey* (WIS) has been carried out (Steinmetz, Tijdens, and Pedraza, 2009). Since 2004, this survey has

collected information on a wide range of subjects including basic demographics, wages, and other work-related topics as well as health and job satisfaction, which can be considered “webographic” variables. These variables are expected to be correlated with participation behavior. The WIS data set contains large samples (90,000 in the Netherlands and 70,000 in Germany). Even though the number of observations of the WIS is extremely large, the samples seem to fail to be representative of the population because of the self-selectivity in sample recruitment.

To correct for bias, weighting adjustment techniques are applied in which reference surveys supply estimates of the population distributions of the weighting variables. In the case of Germany, this is the *Socio-Economic Panel* (SOEP; a representative longitudinal panel study of private households collects data on occupational biographies, employment, earnings, as well as health and job satisfaction indicators). In case of the Netherlands, the *Labor Supply Panel* (OSA) is used. This survey collects data about the (potential) labor force in the Netherlands. The panel is a face-to-face survey based on a representative sample of approximately 2,000 households that are sampled from all households in the Netherlands. The survey includes a large variety of information on labor market position, level of education, family status, and attitudes toward work and health.

By comparing the population distribution of a variable with its sample distribution, it has been established that the sample is not representative for the population with respect to specific variables. In both countries, there was a bias with respect to job satisfaction, part-time work, and age. Less large selection biases have been found for gender and education. Some country-specific selection biases have been detected with respect to education and nonmanual work.

To correct for selection bias, adjustment weights can be computed. Two techniques were applied in this study: poststratification weighting and propensity score weighting. The impact of poststratification weighting was very limited. The bias of estimates for attitude variables remained. The only advantage was that it made the distributions of the auxiliary variables comparable. This finding is in line with other literature such as Loosveldt and Sonck (2008). With respect to the propensity score weighting, the findings of this study showed minimal changes and sometimes correction in the wrong direction. Moreover, the inclusion of additional “webographic” variables did not improve adjustment.

Summing up, the applied weighting techniques did not make web survey data more comparable with the general population in the case of the German as well as of the Dutch WIS. Because none of the applied weights consistently adjusted the web survey estimates in the appropriate direction, it was concluded that it seemed wiser to use the unweighted web data.

It should be underlined that the findings of this study also could be attributed to mode effects (see Chapter 5). In the case of propensity scores weighting, correction could have been more effective if more variables were included into the models for the propensity scores.

It is not easy to draw a simple and clear conclusion from this application with respect to the impact of the lack of representativity. The only thing that can be said is that the tested correction techniques did not have a sufficient effect. With more research, better correction techniques may be developed.

12.4 Summary

A *web panel* (also called an *on-line panel*, *Internet panel*, or *access panel*) is a survey in which the same individuals are interviewed via the web at different points in time. Information is therefore collected in a longitudinal way using the same group of individuals. Panels differ from cross-sectional surveys, which if repeated over time select a new sample for each survey release.

Before a web panel can be used, it has to be built up. In this recruitment phase, individuals are invited to join the panel. There are two ways to do this:

- Draw a probability sample from a sampling frame representing the target population and ask the selected individuals to become members of the web panel. This can be done, for example, by means of a face-to-face survey, telephone survey, or mail survey.
- Let individuals select themselves for the web panel. Invitations can take the form of banners or pop-up windows on websites or of advertisements in other media (radio, TV, and newspapers).

Probability-based web panels have the advantage that they allow for proper statistical inference from the panel to the population. The lack of representativity of self-selection web panels may hinder reliable inference. Weighting adjustment may be applied to improve representativity, but there is no guarantee that this will be successful.

It is more expensive and time-consuming to develop a probability-based panel than a self-selection panel. A well-designed recruitment campaign may cause many people to become a member of a self-selection panel. Such a large panel has the advantage that it will also contain people belonging to special groups in the population. The inference problems caused by the lack of representativity of self-selection panels are not solved by a large panel size.

During recruitment, new members complete a demographic or profile questionnaire. The collected data can be used for weighting adjustment and to select special groups from the panel.

For a specific survey, all panel members can be asked to complete the questionnaire. Often only a random sample is selected from the panel, or a special group that is selected using the available variables.

The response rates of specific surveys are often high. This is not surprising as the people in the panel have already agreed in the recruitment phase to complete questionnaires regularly. This response rate is not a good indicator of the quality of the response in a specific survey. Also, the response rate in the recruitment phase should be taken into account. This results in so-called cumulative response rates.

It is not possible to compute response rates for self-selection panels as the target population is not defined and no sampling frame is used. There are other indicators, like the completion rate, that provide some insight into the quality of the response.

KEY TERMS

Access panel: See **self-selection panel**.

Attrition: The phenomenon that panel members drop out of the panel.

Attrition rate: The proportion of members who drop out of the panel in a defined time period.

Completion rate: The proportion of selected and invited eligible panel members who started and completed the, survey questionnaire.

Cumulative response rate: The response rate obtained by multiplying the response rate in the recruitment phase by the response rates in the subsequent waves or specific surveys.

Internet population: The subpopulation of the target population consisting of only elements that have access to the Internet.

Opt-in panel: See **self-selection panel**.

Probability-based panel: A panel for which members are recruited by means of a probability sample.

Reference survey: A survey conducted with the objective of obtaining unbiased estimates of the population distributions of auxiliary variables.

Representative: The (weighted) survey response is representative with respect to a variable if the (weighted) response distribution is equal to its population distribution.

Self-selection panel: A web panel for which people select themselves in response to a banner, pop-up window, or advertisement in other media (radio, TV, and newspapers).

Volunteer panel: See **self-selection panel**.

EXERCISES

Exercise 12.1. Using the person-level data in Example 12.7, compute the recruitment rate for persons.

Exercise 12.2. A list of eligible households has been contacted with the invitation to become a member of a web panel. Overall, 7,000 households are on the list and have been sent e-mail invitations. A total of 200 e-mails have bounced back as undeliverable, and 250 e-mails can be characterized as network error-undeliverable emails. Compute the absorption rate.

Exercise 12.3. In web panel recruitment of the KnowledgePanel, the following rates have been calculated:

- Profile rate (PROR) = 0.568.
- Completion rate (COMR) = 0.845.
- Break-off rate (BOR) = 0.0056.
- Retention rate (RETR) = 0.390.

Compute cumulative response rate 1 and cumulative response rate 2.

Exercise 12.4. A large self-selection web panel is made representative with respect to gender and age by removing members of overrepresented groups from the panel until the gender by age distribution in the panel is the same as this distribution in the population. The researcher claims that the resulting panel is representative. Is this correct?

- a. Yes, surveys from the panel can now be treated as equal probability samples.
- b. Yes, but weighting adjustment by gender and age should be applied to specific surveys from the panel.
- c. No, but it can be made representative by repeating this for other auxiliary variables.
- d. No, the panel will never be completely representative because a specific part of the population is always missing.

Exercise 12.5. Why is it not possible to compute the response rate for the recruitment phase of a self-selection web panel?

- a. There are only respondents, not nonrespondents.
- b. It is not possible to distinguish nonresponse from overcoverage.
- c. Recruitment is continuous activity.
- d. No initial sample has been selected.

Exercise 12.6. What is an advantage of a self-selection panel over a probability-based panel?

- a. It is less expensive and less time-consuming to construct a web panel.
- b. A much wider population can be covered because no sampling frame is used.
- c. Only people who are really interested become members.
- d. More surveys per month can be offered.

Exercise 12.7. In which situation is it wise to use a reference survey for adjustment weighting?

- a. To improve the accuracy of estimates after weighting adjustment.
- b. If “webographic” variables are unrelated to the target variable of the survey.
- c. If the specific survey lacks representativity and effective weighting variables cannot be retrieved from another source.
- d. Only if the panel was recruited by means of probability sampling.

Exercise 12.8. The recruitment sample for a longitudinal study is obtained by means of probability sampling. The response rate is 50%. There are three waves of interviewing after recruitment. In each wave, 10% of the participants decide to stop. What is the cumulative response rate in the last wave?

REFERENCES

- American Association for Public Opinion Research, AAPOR. (2006), *Final Dispositions of Case Codes and Outcomes Rates for Surveys*, 4th ed. AAPOR, Lenexa, KS.
- Arbeitskreis, D. M. (2001), *Standards for Quality Assurance for Online Surveys*. ADM, Bonn, Germany.
- Arens, Z. & Miller-Steiger, D. (2006). Time in Sample: Searching for Conditioning in a Consumer Panel. *Public Opinion Pros*. www.publicopinionpros.norc.org.
- Bethlehem, J. G. (2009), *Applied Survey Methods, A Statistical Perspective*. John Wiley & Sons, Hoboken, NJ.
- Bethlehem, J. G., Cobben, F., & Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons, Hoboken, NJ.
- Bethlehem, J. G. & Stoop, I. A. L. (2007), Online panels — A theft of paradigm? *The Challenges of a Changing World*, Proceedings of the Fifth International Conference of the Association of Survey Computing, Southampton, U.K. pp. 113–132.
- Callegaro, M. & DiSogra, C. (2008), Computing Response Metrics for Online Panels, *Public Opinion Quarterly*, 72, pp. 1008–1032.
- Clinton, J. D. (2001), *Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel*. Paper presented at 56th Annual Conference of the American Association for Public Opinion Research, Montreal, Canada.
- Comley, P. (2007), Online Market Research. In: ESOMAR (ed.), *Market Research Handbook*, John Wiley & Sons, Hoboken, NJ, pp. 401–20.
- Couper, M. P. (2000), Web Surveys. A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, pp. 464–494.
- Couper, M. (2001), The Promises and Perils of Web Surveys. In: Westlake, A., Sykes, W., Manners, T., & Riggs, M. (eds.), *The Challenge of the Internet*. Proceedings of the ASC International Conference on Survey Research Methods. Chesham, U.K.
- Couper, M. (2007), Issues of Representation in Health Research (with a Focus on Web Surveys). *American Journal of Preventive Medicine*, 32, pp. 83–89.
- Couper, M. & Dominitz, J. (2007), *Using an RDD Survey to Recruit Online Panel Members*. Paper presented at 2007 Biennial Conference of the European Survey Research Association, Prague, Czech Republic.
- Couper, M., Kapteyn, A., Schonlau, M., & Winter J. (2007), Noncoverage and Nonresponse in an Internet Survey. *Social Science Research*, 36, pp. 131–48.
- Day D., Risk, R., Koo, J., & Martin, B. (2006), Ensuring Data Integrity for Business Decisions. An In-depth Analysis of the Components that Affect Data Quality. *Panel Research 2006*, ESOMAR Publication Series 317, pp. 253–269.
- Deming, W. (1960), *Sample Design in Business Research*. John Wiley & Sons, New York.
- Dillman, D. A. & D. K. Bowker, (2001) The Web Questionnaire Challenge to Survey Methodologists. In: Batinic, B., Reips, U. D. & Bosnjak, M. (eds.), *Dimensions of Internet Science*, Pabst Science Publishers, Lengerich, Germany. pp. 159–177.
- DiSogra, C., Slotwiner, D., Clinton, S., Chan, E., Hendarwan, E., & Zheng, W. (2007), *Nonresponse Bias in Two Methods of Panel Recruitment*. Paper presented at Joint Statistical Meetings (JSM), Salt Lake City, UT.
- Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005), Comparing Data from Online and Face-to-Face Surveys. *International Journal of Market Research*, 47, pp. 615–639.

- European Federation of Associations of Market Research Organizations, EFAMRO. (2004), Quality Standards for Access Panel (QSAP). www.efamro.com/shortprint2.html.
- European Society for Opinion and Marketing Research, ESOMAR. (2005), Conducting Market and Opinion Research Using the Internet. www.esomar.org/uploads/pdf/.
- European Society for Opinion and Marketing Research, ESOMAR. (2008), 26 Questions to help research buyers of online samples ESOMAR Codes & Guidelines Conducting Research using Internet. www.esomar.org/uploads/pdf/professional-standards/26questions.pdf.
- Eysenbach, G. (2004), Improving the Quality of Web Surveys: The Checklist for Reporting Results from Internet E-Surveys (Cherries). *Journal of Medical Internet Research*, 6, e34.
- Faas, T. & Schoen, H. (2006), Putting a Questionnaire on the Web is Not Enough: A Comparison of Online and Offline Surveys Conducted in the Context of the German Federal Election 2002. *Journal of Official Statistics*, 22, pp. 177–190.
- Fowler, F. J. Jr. (2002), *Survey Research Methods*, 3rd ed. Sage, Thousand Oaks, CA.
- Fricker, R. D. & Schonlau, M. (2002), Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Social Science Computer Review*, 14, pp. 347–67.
- Görizt, A. S. (2007), Using Online Panels in Psychological Research. In: Joinson, A. N., McKenna, K. Y. A., Postmes, T., & Reips, U. D. (eds), *The Oxford Handbook of Internet Psychology*. Oxford University Press, New York, pp. 473–485.
- Groves, R. M. (1989), *Survey Errors and Survey Costs*. John Wiley & Sons, New York.
- Groves, R. M. (2006), Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70, pp. 646–675.
- Groves, R. M. & Peytcheva, E. (2008), The Impact of Nonresponse Rates on Nonresponse Bias. A Meta-Analysis. *Public Opinion Quarterly*, 72, pp.167–189.
- Hoogendoorn, A. W. & Daalmans, J. (2008), *Nonresponse in the Recruitment of an Internet Panel Based on a Probability Sample*. Discussion Paper 08007, Statistics Netherlands, Voorburg/Heerlen, the Netherlands.
- Huggins, V. & Eyerman, J. (2001), *Probability Based Internet Surveys: A Synopsis of Early Methods and Survey Research Results*. Paper presented at Federal Committee on Statistical Methodology Research Conference, Arlington, VA.
- Interactive Marketing Research Organization, IMRO. (2006), IMRO Guidelines for Best Practices in Online Sample and Panel Management. www.imro.org/pdf/.
- Kalton, G. (1987), *Introduction to Survey Sampling*. Sage, Thousand Oaks, CA.
- Kiaer, A. N. (1895), Observations et Expériences Concernant des Dénombrements Représentatives. *Bulletin of the International Statistical Institute*, IX, Book 2, pp. 176–183.
- Kruskal, W. & Mosteller, F. (1979a), Representative Sampling, I: Non-scientific Literature. *International Statistical Review*, 47, pp. 13–24.
- Kruskal, W. & Mosteller, F. (1979b), Representative Sampling, II: Scientific Literature. Excluding Statistics, *International Statistical Review*, 47, pp. 111–127.
- Kruskal, W. & Mosteller, F. (1979c), Representative Sampling, III: The Current Statistical Literature. *International Statistical Review*, 47, pp. 245–265.
- Lee S., (2006), An Evaluation of Nonresponse and Coverage Errors in a Pre-recruited Probability Web Panel Survey. *Social Science Computer Review*, 24, pp. 460–475.

- Lee, S. & Valliant, R. (2009), Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research* 37, pp. 319–343.
- Loosveldt, G. & Sonck, N. (2008), An Evaluation of the Weighting Procedures for Online Access Panel Surveys. *Survey Research Methods*, 2, pp. 93–105.
- Lozar Manfreda, K. & Vehovar, V. (2002), *Survey Design Features Influencing Response Rates in Web Surveys*. Paper presented at International Conference on Improving Surveys, Copenhagen, Denmark.
- Lozar Manfreda, K. & Vehovar, V. (2008), Internet Surveys. In: De Leeuw, E., Hox, J., & Dillman D. A. (eds.), *International Handbook of Survey Methodology*. Lawrence Erlbaum, New York.
- Lynn, P., Clarke, P., Martin, J., & Sturgis, P. (2002), The Effects of Extended Interviewer Efforts on Nonresponse Bias. In: Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (eds.), *Survey Nonresponse*. John Wiley & Sons, New York.
- Miller, J. (2006), Online Marketing Research. In: Grover, R. & Vriens, M. (eds.), *The Handbook of Marketing Research. Uses, Abuses and Future Advances*. Sage, Thousand Oaks, CA.
- Morton-Williams, J. (1993), *Interviewer Approaches*. Dartmouth Publishing, Aldershot, U. K.
- Office of Management and Budget, OMB. (2006), *Questions and Answers When Designing Surveys for Information Collections*. Office of Management and Budget, Washington, DC.
- Pineau, Nukuljij & Tang (2006). Assessing Panel Bias in the Knowledge Networks Panel: Updated Results from 2005 Research. In: *Joint Statistical Meeting 2005 Proceedings*, pp. 3480–3486, Alexandria, VA, American Statistical Association.
- Postoaca, A. (2006a), *The Anonymous Elect. Market Research through Online Access Panels*. Springer, Berlin, Germany.
- Postoaca, A. (2006b), Response Rates. Avoiding the Red Herrings. *Panel Research 2006*, ESOMAR. Amsterdam, the Netherlands.
- Pratesi, M., Lozar Manfreda, K., Biffignandi, S., & Vehovar, V. (2004), List-Based Web Surveys: Quality, Timeliness, and Nonresponse in the Steps of the Participation Flow. *Journal of Official Statistics*, 20, pp. 451–465.
- Reips, U. D. (2007), The Methodology of Internet-based Experiments. In: Joinson A. N., McKenna, K. Y. A., Postmes, T., & Reips, U. D. (eds.), *The Oxford Handbook of Internet Psychology*. Oxford University Press, New York.
- Saris, W. E. (1991), *Computer Assisted Interviewing*. Sage, Newbury Park, CA.
- Saris, W. E. (1998), Ten Years of Interviewing without Interviewers: The Telepanel. In: Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls, W. L., & O'Reilly, J. M. (eds.), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.
- Sayles H., & Arens, Z. (2007), *A Study of Panel Member Attrition in the Gallup Panel*. Paper presented at 62nd Annual Conference of the American Association for Public Opinion Research, Anaheim, CA.
- Scherpenzeel, A. (2008), An Online Panel as a Platform for Multi-Disciplinary Research. In: Stoop, I. & Wittenberg, M. (eds.), *Access Panels and Online Research, Panacea or Pitfall?* Aksant, Amsterdam, the Netherlands.

- Schillewaert, N., Verhaeghe, A., Weijters, B., & De Wolf, K. (2006), Social Class and Life Style Differences Between Modes of Data Collection. *Panel Research 2006*, ESOMAR Publication Series. 317, pp. 174–193.
- Schlegel, W. E., Caddell, J. M., Ebert, L., Jordan, K. B., Rourke, K. M., Wilson, D., Thalji, T., Dennis, M. J., Fairbank, J. A., & Kulka, R. (2002), Psychological Reactions to Terrorist Attacks. Findings from the National Study of American's Reactions to September 11. *Journal of the American Medical Association*, 288, pp. 581–588.
- Schonlau, M., Fricker, R. D., & Elliott, M. N. (2002), *Conducting Research Surveys Via E-Mail and the Web*. Rand, Santa Monica, CA.
- Schonlau, M., Van Soest, A., Kapteyn, A., & Couper, M. P. (2007), Are “Webographic” or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring? *Survey Research Methods*, 1, pp. 155–156.
- Schonlau, M., Van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection Bias in Web Surveys and the Use of Propensity Scores. *Sociological Methods and Research*, 37, pp. 291–318.
- Sikkel, D. & Hoogendoorn, A. (2008), Panel Surveys. In: De Leeuw, E., Hox, J., & Dillman, D. A. (eds.), *International Handbook of Survey Methodology*. Lawrence Erlbaum, New York.
- Steinmetz S., Tijdens, K. & Pedraza P. (2009), Comparing Different Weighting Procedures for Volunteer Web Surveys, WP 9–76. AIAS, Amsterdam.
- Stoop, I. A. L. (2005), *The Hunt for the Last Respondent*. Social and Cultural Planning Office, The Hague, the Netherlands.
- Sudman, S. (1966), Probability Sampling With Quotas. *Journal of the American Statistical Association*, 61, pp. 749–771.
- Tortora, R. (2009), Attrition in Consumer Panels. In: Lynn, P. (ed.), *Methodology of Longitudinal Surveys*. John Wiley & Sons, Hoboken, NJ.
- Tourangeau, R. (2003), Web-Based Data Collection. In: Cork D. L., Cohen, M. L., Groves, R., & Kalsbeek, W. (eds.), *Survey Automation, Report and Workshop Proceedings*, National Academies Press, Washington, DC.
- Vonk, T., Van Ossenbruggen, R., & Willems P. (2006), The Effect of Panel Recruitment and Management on Research Results. A Study across 19 Online Panels. *Panel Research 2006*, ESOMAR Publications Series, Amsterdam, the Netherlands.

Index

- 21minuten.nl, 19, 304
- Ability, 225
- Absolute maximum bias, 312
- Absorption rate, 438
- Access panel, see web panel
- Accessibility, 150
- Accuracy, 149, 150, 183
- Acquiescence, 108, 135, 136, 138, 151, 154, 161, 167, 171, 183, 245
- Active panel, 428
- Adjusted population variance, 63
- Adjustment weight, 335
- Adjustment weighting, see weighting adjustment
- Administrative burden, 264
- Administrative nonresponse, 255
- Advance letter, 154, 159
- Allocation, 79, 92
 - Neyman, 79
 - optimal, 79
 - proportional, 79
- American Life Panel, 425
- Answer
 - space, 211
 - type, 200
- Arbitrary answer, 116, 151, 155, 167, 172, 245
- Archival analysis, 235
- ARPANET, 13
- Asymptotically design unbiased (ADU), 86
- Attitude, 104
- Attitude toward behavior, 221
- Attrition, 452
 - panel, 431
 - rate, 441, 452
- Audit trail, 217, 218
- Auxiliary variable, 62, 335, 336, 349, 379, 394
- Average man, 3
- Behavioral approach, 220
- Behavioral belief, 221
- Behavioral intention, 222
- Bias, 98, 130, 293, 333, 334, 342, 345
 - absolute maximum, 312
 - relative, 290, 313
 - self-selection, 314, 356
- Blaise, 21, 23, 26, 30, 31, 174, 183
- Book of the Year Award, 306
- Break-off rate, 439
- Broadband, 284
- Browser, 13, 171, 196
- Business survey, 262
- CADI, 22
- Calibration estimator, 336, 361
- Call
 - attempt, 160
 - management, 11, 161, 181
- Calling number identification (CNID), 161
- CAPI (Computer-assisted personal interviewing), 2, 11, 26, 32, 37, 56, 148, 183
- CASI (Computer-assisted self-interviewing), 12, 32, 148, 184
- Categorical variable, 62, 349, 397

- CATI (Computer-assisted telephone interviewing), 2, 11, 27, 32, 37, 56, 148, 184
- CAWI (Computer-assisted web interviewing), 56, 148, 184
- Census, 1, 3, 31, 59
- CentERpanel, 425
- Check, 24, 121, 152, 156, 164, 168, 173, 177, 182
- Check box, 18, 116, 200, 208, 228
- Check-all-that-apply question, 18, 117, 155, 162, 167, 172, 198, 208, 228
- Clarity, 150
- Client-side paradata, 220
- Closed question, 17, 197, 229
- Cluster sampling, 80, 92
- Cognitive
 approach, 223
 efforts, 247
 interview, 225
- Cognitive Aspects of Survey Methodology (CASM), 52, 221
- Coherence, 149
- Collapsing strata, 348
- Color, 214
- Comparability, 149
- Complete enumeration, 1, 21, 59
- Completion rate, 195, 433, 439, 442, 452
- Comprehension, 247
- Computer-assisted data input (CADI), 22
- Computer-assisted interviewing (CAI), 1, 10, 11, 31, 148, 183, 281
- Computer-assisted personal interviewing (CAPI), 2, 11, 26, 32, 37, 56, 148, 183
- Computer-assisted self-interviewing (CASI), 12, 32, 148, 184
- Computer-assisted self-administered questionnaires (CSAQ), 12, 148, 183
- Computer-assisted telephone interviewing (CATI), 2, 11, 27, 32, 37, 56, 148, 184
- Computer-assisted web interviewing (CAWI), 56, 148, 184
- Concurrent mixed-mode data collection, 238, 251, 274, 291
- Conditional independence, 386, 394
- Confidence
 interval, 6, 70, 98, 290, 308, 313
 level, 70, 290, 308
- Consumer Satisfaction Survey, 236
- Contact attempt, 160
- Continuous variable, 62
- Contrast, 289, 299, 333
- Control belief, 221
- Convenience sampling, 445
- Core survey, 426
- Costs, 10, 47, 153, 158, 165, 170, 236, 244
- Cover letter, 166
- Coverage problem, 67, 150, 153, 246
- Cramér's V, 399
- Cross-sectional
 research, 426
 survey, 41, 44, 56
- CSAQ, 12, 148, 183
- Cumulative response rate, 441, 452
- Current Employment Statistics (CES), 19, 262
- Data collection, 147, 235
 mixed-mode, 20, 182, 294
 mode, 149, 240
 off-line, 44, 263
 on-line, 44, 45, 263
- Data collection instrument, 238
- Data editing, 10, 103, 138, 158, 165
- Data Editing Research Project, 21, 22
- Data entry
 heads-down, 23
 heads-up, 22
- Design weight, 330, 335
- Design-based inference, 261
- Disparate modes, 250
- Do-not-call register, 159, 161
- Domesday Book, 3
- Don't know, 113, 114, 115, 151, 155, 162, 167, 172, 177, 182, 191, 201, 205, 245
- Double opt-in enrollment, 423
- Drop-down box, 107, 200, 207, 229
- Dummy variable, 349, 351
- Dutch Online Panel Study, 420
- Dynamic shading, 112
- E-mail, 13, 39, 170, 189, 267, 282
- E-mail survey, 2, 13, 14, 15, 32, 304
- EenVandaag Opinion Panel, 372
- Effective sample size, 369
- Electronic Data Reporting (EDR), 16

- Eligible, 138
- Endorsing the status quo, 109, 151, 155, 162, 167, 171, 185, 245
- Error
- total, 51, 97, 140
 - checking, 26
 - message, 173, 179
- Estimate, 69
- Estimation error, 99, 138
- Estimator, 69
- generalized regression, 85, 336, 349, 350, 361, 379, 402
 - Horvitz-Thompson, 69, 70, 79, 82, 93, 303, 307, 330, 390
 - poststratification, 84, 93, 338
 - ratio, 82, 93
 - regression, 83, 93
- Exogeneity, 394
- Face-to-face
- interviewing, 8, 32
 - survey, 152, 184
- Factual question, 104
- Filter question, 115, 167
- First-order inclusion probability, 69, 71, 289, 299, 307, 323, 330, 379, 390
- Follow-up, 45
- Form-based approach, 30, 119, 173, 181
- Frame population, 284, 299
- Gallup Panel, 424, 430
- Generalized regression estimator, 85, 336, 349, 350, 361, 379, 402
- Gestalt principle
- of similarity, 214
 - of simplicity, 214
- Graphic language, 199
- Grid question, see matrix question
- Heads-down data entry, 23
- Heads-up data entry, 22
- Homogeneous, 77, 317, 337, 340, 379
- Horvitz-Thompson estimator, 69, 70, 79, 82, 93, 303, 307, 330, 390
- modified, 391
- Household survey, 267
- HTML, 17, 106, 107, 200
- Hypertext, 13
- Ignorable treatment, 394
- Image, 215
- Incentive, 161, 166, 273
- Inclusion probability
- first-order, 69, 71, 289, 299, 307, 323, 330, 379, 390
 - second-order, 69, 71, 307, 330
- Indicator variable, 62
- Inference, 258, 261
- design-based, 261
 - model-based, 261
- Instruction
- question, 192
- Interactive Voice Response (IVR), 241
- Internet, 2, 12, 170, 171, 267, 284
- coverage, 41, 48, 49, 267, 285
 - panel, see web panel
 - population, 287, 294, 299, 309, 324, 379, 448, 452
 - survey, 40, 56
- Internet Voice, 19
- Internet and Computer Technology Survey, 42, 252, 270
- Interactional information system
- approach, 196
- Interview
- approach, 29
 - effect, 102
- Interviewer-assisted survey, 190
- Item nonresponse, 124, 138
- Judgment, 247
- Kauffman Firm Survey (KFS), 49
- Key-from-image processing, 211
- Keystroke file, 217
- Knowledge Panel, 424, 428
- L'homme moyenne, 3
- Language
- graphic, 199
 - numeric, 199
 - symbolic, 199
 - visual, 199
- Large weight, 363
- Likert scale, 54, 203, 226, 229
- Linear model, 395
- Link function, 395
- LISS Panel, 65, 127, 132, 170, 290, 425, 427, 430

- Logfile, 218
- Logistic regression model, 394
- Longitudinal
 - research, 426
 - survey, 41, 56
- Mail survey, 13, 32, 37, 38, 164, 184
- Margin of error, 74
- Marginal distribution, 355
- Matching assumption, 396
- Matrix question, 55, 111, 138, 177, 207, 209
- Mean square error, 98
- Measurement error, 102, 103, 134, 139, 151, 269
- Memory
 - effect, 102
 - error, 122, 139
- Memory-based model, 105
- Middle option, 110, 167
- Missing at Random (MAR), 131, 139, 295, 299, 317, 320, 324, 365, 394
- Missing Completely at Random (MCAR), 131, 139, 299, 311, 320, 324
- Mixed-mode
 - data collection, 20, 182, 294
 - concurrent, 238, 251, 274, 291
 - sequential, 239, 244, 251, 275, 291
 - survey, 20, 54, 184, 235, 275, 299, 324
 - system, 251, 275
- Mobile telephone, 160
- Mode, 238
 - effect, 245, 250, 275
 - of data collection, 149, 240
- Model-based inference, 261
- Modified Horvitz-Thompson estimator, 391
- Monograph study, 3
- Motivation, 196, 225
- Multi-mode system, see mixed-mode system
- Multiplicative weighting, see raking ratio estimation
- Negative weight, 361, 363
- Nested options, 202
- Neyman allocation, 79
- No opinion, 205
- Non-Internet population, 287, 294, 299, 309, 324, 448
- Noncontact, 126, 255
- Noncoverage bias, 290
- Nondifferentiation, 111, 135, 139, 151, 155, 162, 167, 171, 184, 245
- Nonfactual question, 104
- Nonobservation error, 102, 139
- Nonprobability sampling, 445
- Nonresponse, 67, 103, 124, 127, 136, 139, 150, 159, 160, 166, 170, 184, 340, 364, 379, 386
 - item, 124, 138
 - selective, 125
 - unit, 124, 140
- Nonsampling error, 51, 101, 139
- Normative belief, 221
- Not able, 126, 256
- Not Missing at Random (NMAR), 132, 139, 295, 300, 320, 324
- Numeric language, 199
- Numerical question, 229
- Observation error, 101, 139
- Off-line data collection, 44, 263
- On-line data collection, 44, 45, 263
- On-line panel, see web panel
- Online processing model, 104
- Open Debate, 19
- Open question, 198, 229
- Opinion, 104
- Opinion poll, 75, 304
- Opt-in panel, 422, 452
- Optimal allocation, 79
- Ordinal question, 229
- Overcoverage, 67, 283, 300
- Overcoverage error, 101, 139
- PAADEL panels, 426
- Panel, 44
 - opt-in, 422, 452
 - self-selection, 422, 442, 444, 452
 - volunteer, 422, 452
- Panel attrition, 431
- Panel maintenance, 431
- Paradata, 46, 51, 56, 191, 217, 220
 - client-side, 220
 - server-side, 218
- Partial investigation, 3
- Perceived behavioral control, 221
- Personalization, 15, 165
- Perspektive Deutschland, 20, 304

- Physical impediments, 154, 160
- Pictures, 199
- Population
 - mean, 63, 76
 - percentage, 63, 72
 - register, 43, 64, 282
 - total, 62
- Popup, 189
- Postal Delivery Points, 64
- Poststratification, 292, 300, 315, 319, 324, 336, 337, 351, 380, 404
 - estimator, 84, 93, 338
- Precision, 69, 71, 74, 98, 184
- Preselection shading, 112
- Pretest, 192
- Primacy effect, 106, 139, 166, 171, 184
- Probability
 - sample recruitment, 422
 - sampling, 1, 6, 32, 60, 184, 303, 306, 307, 314, 329, 444
- Probability-based panel, 452
- Probit model, 394
- Processing error, 102, 139
- Profile rate, 436, 437
- Profiling survey, 426
- Progress indicator, 194, 195
- Propensity
 - score, 52, 318, 414
 - score matching, 386
 - score method, 414
 - score weighting, 318
- Proportional allocation, 79
- Psychographic variable, 388
- Punctuality, 149
- Purposive sampling, 6, 32, 445

- Qualitative interview, 46, 56
- Quality, 10, 149, 421
- Question
 - check-all-that-apply, 18, 117, 155, 162, 167, 172, 198, 208, 228
 - closed, 17, 197, 229
 - factual, 104
 - filter, 115, 167
 - numerical, 229
 - open, 198, 229
 - ordinal, 229
 - sensitive, 162, 171, 190, 240
 - yes/no, 200
 - instruction, 192
 - layout, 190
 - order, 123
 - wording, 196, 200
 - wording effect, 102
- Question-based approach, 119, 173, 181
- Questionnaire
 - design, 191, 196, 211, 223, 248
 - design effect, 152, 156, 168, 172
 - drop-off and pick-up, 166
- Quipu, 3
- Quipucamayoc, 3
- Quota sampling, 7, 32, 445, 446

- Radio button, 17, 106, 116, 200, 201, 229
- Raking ratio estimator, 336, 358, 361, 380, 402
- Raking, *see* raking ratio estimation
- Random digit dialing (RDD), 10, 32, 159, 160, 184
- Random
 - error, 103
 - response model, 129, 139, 332, 390
 - sampling, 6, 7, 47, 288, 304
 - simple, 68, 71, 93, 308, 330, 389
- Rate
 - absorption, 438
 - attrition, 441, 452
 - break-off, 439
 - completion, 195, 433, 439, 442, 452
 - cumulative response, 441, 452
 - profile, 436, 437
 - recruitment, 434, 435
 - response, 46, 48, 51, 127, 140, 153, 159, 161, 165, 166, 236, 243, 244, 248, 254, 257, 270, 432, 433, 447
 - screening completion, 440
 - study-specific eligibility, 440
- Ratio estimator, 82, 93
- Recency effect, 106, 135, 140, 161, 184
- Recruitment, 66, 189, 240, 283, 422, 427, 428
 - probability sample, 422
 - self-selection, 422
- Recruitment
 - questionnaire, 423
 - rate, 434, 435
- Reference
 - date, 67
 - survey, 318, 363, 364, 366, 369, 380, 388, 450, 452

- Refusal, 126, 255
- Regression estimator, 83, 93
- Relative bias, 290, 313
- Relevance, 149
- Reliability, 98, 149, 150, 184
- Reminder, 273
- Representative method, 4, 5, 6, 32, 336
- Representativity, 4, 52, 305, 334, 336, 380, 420, 421, 443, 452
with respect to a variable, 335, 443
- Response, 247
burden, 47
order effect, 106, 140, 151, 154, 161, 166, 171, 184, 245
probability, 129, 310, 311, 390, 392, 393, 408, 414
propensity, 385, 387, 388, 394, 396, 408, 414
propensity estimator, 404
propensity strata, 405
propensity stratification, 404, 411, 414
propensity weighting, 401, 411, 414
rate, 46, 48, 51, 127, 140, 153, 159, 161, 165, 166, 236, 243, 244, 248, 254, 257, 270, 432, 433, 447
- Retrieval, 247
- Road pricing, 61, 101, 304
- Route control, 26, 152, 156, 164, 168, 179, 180, 182
- Safety Monitor, 81, 133, 237, 253
- Sample
selection bias, 386
size, 69, 71, 74, 75, 307
effective, 369
survey, 1, 21
- Sampling
cluster, 80, 92
stratified, 77, 93
two-stage, 81, 93
unequal probability, 79, 93
error, 51, 99, 140
fraction, 71
frame, 42, 63, 64, 282, 283, 286, 300, 386
with replacement, 68
without replacement, 68
- Satisficing, 105, 140, 155, 162, 167, 172, 185
strong, 105, 224
weak, 105, 224
- SCI Survey, 264
- Screening completion rate, 440
- Second-order inclusion probability, 69, 71, 307, 330
- Selection
effect, 136, 140, 246
on variables, 394
- Selective nonresponse, 125
- Self-administered survey, 190, 256
- Self-selection, 56, 101, 303, 305, 306, 333, 345
bias, 314, 356
panel, 422, 442, 444, 452
recruitment, 422
sampling, 309, 324, 380, 392
- Sensitive question, 162, 171, 190, 240
- Sequential mixed-mode data collection, 239, 244, 251, 275, 291
- Server-side paradata, 218
- Show card, 8
- Simple random sampling, 68, 71, 93, 308, 330, 389
- Single opt-in enrollment, 423
- Social Statistics Database (SSD), 397
- Socially desirable answer, 118, 140, 151, 156, 164, 168, 172, 184, 240, 245, 275
- Specific survey, 426
- Specification error, 100, 140
- Standard error, 70
- Straightlining, 112, 140, 162, 167, 171, 185, 206
- Stratification, 336, 380
variable, 317
response propensity, 404, 411, 414
- Stratified sampling, 77, 93
- Straw poll, 7, 32
- Strong ignorability, 387
- Strong satisficing, 105, 224
- Study-specific eligibility rate, 440
- Subjective norm, 221
- Survey, 32, 60
interviewer-assisted, 190
mixed-mode, 20, 54, 184, 235, 275, 299, 324

- self-administered, 190, 256
- administrator, 41, 42
 - Central, 19
 - sampling, 4, 7
 - topic, 43
- Survey 2000, 19
- Symbolic language, 199
- Systematic error, 103

- Target population, 40, 41, 60, 283, 303, 306, 386
- Target variable, 61, 287, 307
- Task difficulty, 225
- Technical aspects, 192
- Technical implementation, 41, 44
- Telepanel, 12
- Telephone
 - coverage, 9
 - interviewing, 9, 32, 185
 - survey, 158
- Telescoping, 123
- Text
 - area, 209, 210
 - box, 209, 210
- The Library Study, 193
- Theory of planned behavior (TPB), 52, 222, 229
- Timeliness, 149, 152, 157, 164, 173
- Total error, 51, 97, 140
- Touchtone Data Entry (TDE), 266
- Two-stage sampling, 81, 93

- Unbiased, 69, 97
- Unconfoundedness, 394
- Undercoverage, 41, 43, 66, 102, 127, 140, 159, 160, 170, 185, 246, 283, 300, 333, 342, 380, 386
- Unequal probability sampling, 79, 93
- Unimode approach, 248, 259, 275
- Unit nonresponse, 124, 140

- Unit of measurement, 213
- Usability testing, 46, 56

- Validity, 150, 185
- Variable
 - auxiliary, 62, 335, 336, 349, 379, 394
 - categorical, 62, 349, 397
 - continuous, 62
 - dummy, 349, 351
 - indicator, 62
 - psychographic, 388
 - stratification, 317
 - target, 61, 287, 307
 - webographic, 388
- Variance, 308, 339, 365, 388
- Visual language, 199
- Volunteer panel, 422, 452

- WageIndicator Web Survey, 449
- Weak satisficing, 105, 224
- Web panel, 56, 419, 422
- Web survey, 2, 32, 37, 38, 49, 56, 59, 169, 185, 189, 281
- Webographic variable, 388
- Website, 190
- Weight
 - adjustment, 335
 - large, 363
 - negative, 361, 363
 - coefficient, 350, 358
 - factor, 358
- Weighting
 - adjustment, 260, 292, 297, 315, 332, 336, 450
 - propensity score, 318
 - response propensity, 401, 411, 414
- World Wide Web, 2, 18

- Yes/no question, 200