Nonresponse in Sample Surveys

Methods for Analysis and Adjustment

Fannie Cobben

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Printed by
OBT bv, The Hague

Cover
TelDesign, Rotterdam

Information
Telephone + 31 88 570 70 70
Telefax + 31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax + 31 45 570 62 68

Internet
www.cbs.nl

ISBN: 978-90-357-2028-2

6008209020 X-11

NONRESPONSE IN SAMPLE SURVEYS
Methods for Analysis and Adjustment

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 2 oktober 2009, te 10.00 uur

door

Fannie Cobben

geboren te 's-Hertogenbosch

Promotiecommissie

Promotor:	Prof. dr. J.G. Bethlehem
Co-promotor:	Prof. dr. J.G. De Gooijer
Overige Leden:	Prof. dr. R.J.M.M. Does
	Prof. dr. J.J. Hox
	Prof. dr. J.F. Kiviet
	Prof. dr. C.J. Skinner

Faculteit Economie en Bedrijfskunde

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

In the first place, I am very grateful for the skillful guidance and support of my promotor, Jelke Bethlehem, whose inspiring enthusiasm enabled me to develop an understanding of the nonresponse problem. You quickly became my role model of a successful researcher in the field of survey nonresponse. I am indebted to many of my colleagues for their encouragement during the past five years. Special thanks are there for Barry Schouten. Our brainstorm sessions contributed greatly to this thesis and without your invaluable support and guidance I would have been lost. Jacco, thank you for the much appreciated 'koffie, thee of limonade'. José, you offered me discipline when I needed it most. I really appreciate your inspiring interest in my work.

I am grateful to my co-promotor, Jan de Gooijer, who supported me during the completion of the thesis. Futhermore, I wish to acknowledge the members of my promotion committee: Chris Skinner, Jan Kiviet, Joop Hox and Ronald Does for agreeing to act as a member of the committee. The International Workshops on Household Survey Nonresponse were a constant factor during my PhD-period. It was always inspiring to participate in this workshop.

It would be nice to thank all my friends individually. By fear of leaving someone out though, I will simply say *thank you very much* to all of you. However, some of you are lucky: thank you Caroline, Fabienne, Janneke, Karin and Stacy for being the great friends that you are to me.

I cannot end without expressing gratitude to my family: 'aunt' Marian, my brother and sister Bram and Fleur, and especially my parents, Frans and Lieke, for their absolute confidence in me. And finally I thank Dras, who shares my happiness and makes me happy.

Fannie Cobben, The Hague, June 2009

Contents

Chapter 1

An Introduction to Sample Surveys and Nonresponse

Since 1899, Statistics Netherlands collects and produces information about aspects of the Dutch society (Erwich and Van Maarseveen, 1999). The information comes from persons, households and businesses. Traditionally, Statistics Netherlands used a census to collect the data for its social surveys. A census, or complete enumeration, is a very costly and time consuming approach. In 1895, the Director General of Statistics Norway held a speech at the annual meeting of the International Statistical Institute in favour of sample surveys. It took some time before the idea of sampling became widely accepted in statistical research, but eventually this led to the development of the classical sampling theory. It was only after the second World War that Statistics Netherlands started collecting data with sample surveys.

In this chapter, we first describe the ideal situation of sample surveys with full response. We discuss the classical idea behind sampling and describe estimators for statistics about population characteristics. Next, we describe what can go wrong when conducting a survey. We especially focus on missing data caused by unit nonresponse and we describe how nonresponse can be introduced into sampling theory by the concept of response probabilities. We end this chapter with an overview of this thesis.

1.1 Introduction

The theory behind sampling is well described by, amongst others, Cochran (1977) and Särndal et al. (1992). In section 1.2 of this chapter we give an introduction to the general idea behind sample surveys. We first outline how data can be collected by sample surveys and how sample surveys are designed. Next, we introduce estimators for population characteristics. These estimators use auxiliary information and can be used to estimate population characteristics if the survey data is not subject to errors.

In practice, however, surveys are potentially subject to a number of errors. Therefore, in section 1.3 we describe what can go wrong when conducting a survey. We discuss potential sources of error in sample surveys. We especially focus on missing data due to unit nonresponse. Furthermore, we describe the response process as a sequential process with different stages that result in different response types. Hence, we no longer make the unrealistic assumption of full response. Instead nonresponse is introduced in the classical sampling theory by means of the random response model. We end this chapter with an overview of this thesis in section 1.4.

1.2 Classical sampling theory

1.2.1 Sample surveys

A sample survey is an instrument to provide information about a specific population, based on the observation of only a small part of that population. Under specific conditions, it is possible to use the sample to make inference about the population as a whole. These specific conditions comprise that the sample is randomly selected from the population, i.e. the selection mechanism uses a random selection procedure. Furthermore, we have to know how the selection mechanism works to compute the probabilities of being selected in the sample. We also need procedures to use the observed data to make estimates of population characteristics. If these conditions are fulfilled, it is possible to make proper inference about the population as a whole. For now, we will make the assumption of full response to the survey.

The population for which information is collected, is the *target population* of the survey. In the surveys that we discuss, the sample elements are either persons or households from the general population of the Netherlands, but the target population of a survey can also consist of something else, like businesses,

or students of a particular school. The elements in the population must be identifiable. We denote the target population by U, consisting of N physically existing elements, $i = 1, 2, \ldots, N$ which we can observe and on which we can make measurements. Furthermore, U is assumed to be finite and N is supposed to be known.

1.2.1.1 Types of variables

There are two types of variables that play a role in survey sampling: survey items and auxiliary variables. *Survey items* are the variables of interest that are measured in the survey. A survey item can be the answer to one survey question, but it can also be composed of the answers to two or more survey questions. Every element $i = 1, 2, \ldots, N$ of the population U has a fixed, non random value for the survey item, denoted by Y_i. The only randomness comes from the selection of the sample. The $N \times 1$-vector of values for the survey items is denoted by $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$.

Auxiliary variables are in general variables that are available prior to sampling. They can be used in creating a sampling design or in the computation of the survey estimates. We denote auxiliary variables by X. Usually, more than one auxiliary variable is available. Suppose we have J auxiliary variables. Every person i can be associated with a $J \times 1$-vector of values for the auxiliary variables $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})'$. The auxiliary information for the entire population is denoted by the $N \times J$-matrix $\mathbf{X}$.

Usually the auxiliary variables are available for every element in the sample or the target population, for instance from a population register. When used to create a sampling design, the values of the auxiliary variables need to be known for every element in the population. However, in the estimation stage such detailed information is not always necessary. It may be sufficient to know the population total of the auxiliary variables, while the individual values are known for the respondents only.

Särndal and Lundström (2005) distinguish between auxiliary information on the sample level, denoted by vector $\mathbf{X}^\circ$ of dimension J°, and information on the population level, denoted by vector $\mathbf{X}^*$ of dimension J^*. The population total $\sum_U \mathbf{X}^*$ is known, for instance from an external source. However the population total $\sum_U \mathbf{X}^\circ$ is not known, but can be estimated based on the sample. In both cases, the values for the auxiliary variables are known for the responding elements in the sample. In case there is a population register, as in the Netherlands (see Chapter 3), values for the auxiliary variables are available for all elements i in the population U. We denote the auxiliary information vector for element

i by $\mathbf{X}_i$, and omit the superscript $*$ or $\circ$. Only when the level of information is not clear from the context, we distinguish between sample information and population information.

A special type of auxiliary variable is *paradata*. Paradata, also known as process data, is information about the data collection process. For example call history data or the number of contact attempts. Typically, paradata is auxiliary information that is available on the sample level. In Chapter 9 we describe paradata in more detail.

Survey items and auxiliary variables can be either quantitative, qualitative or indicator variables. *Quantitative variables* measure quantities, amounts, sizes, or values. Examples of quantitative variables are income and age. *Qualitative variables* divide the population into groups. The values denote categories. Sample elements in the same category belong to the same group. Examples of qualitative variables are ethnicity, religion and employment status. Qualitative variables are often referred to as categorical variables. Most of the social statistics that Statistics Netherlands produces are categorical variables. Business statistics are usually quantitative. *Indicator variables* measure whether or not a sample element has a certain property. They can only assume the values 0 and 1. The value is 1 if an element has the property, and otherwise it is 0. An example of an indicator variable is the response indicator. If a sample element responded to the survey, the value of the response indicator is 1. Otherwise its value is 0.

Paradata are always restricted to the sample level, whereas other auxiliary information can also be available on the population level. Survey items are identified for the target population, however we only observe them for the sample or, more precisely, for the respondents to the survey. For now, however, we do not regard nonresponse so the values for the survey items are assumed to be known for the selected sample elements.

1.2.1.2 Population characteristics

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

quantitative variable are the population total

$$Y_{tot} = \sum_{i=1}^{N} Y_i \tag{1.1}$$

and, related to the population total, the population mean

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i = \frac{1}{N} Y_{tot} \tag{1.2}$$

Another important population parameter is the population variance

$$S^2(\mathbf{Y}) = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \tag{1.3}$$

For indicator variables, the population total counts 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

$$P = 100\overline{Y} = \frac{100}{N} \sum_{i=1}^{N} Y_i = \frac{100}{N} Y_{tot} \tag{1.4}$$

Note that for indicator variables the population variance reduces to

$$S^2(\mathbf{Y}) = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 = \frac{N}{N-1} P(100 - P) \tag{1.5}$$

There are no specific population parameters for qualitative variables. Of course we can estimate totals, fractions or percentages of elements within categories. This comes down to replacing the qualitative variable by a set of indicator variables.

1.2.2 Sampling design

The aim of a survey is to provide information on population characteristics, like e.g. the population mean or total. These characteristics can be estimated based on a sample that is selected by a random selection procedure from the population. A random selection procedure assigns a known, fixed and non-zero probability of being selected to every element i in the population U. If the

sample is selected without replacement, elements can be selected only once and the random selection procedure results in a selection indicator $\delta_i \in \{0, 1\}$ for every element $i \in U$, i.e.

$$\delta_i = \begin{cases} 1, & \text{if element } i \text{ selected in sample} \\ 0, & \text{otherwise} \end{cases} \tag{1.6}$$

The sample size is equal to $n = \sum_{i=1}^{N} \delta_i$. The sample s consists of all n elements for which $\delta_i = 1$ and $s \subset U$. A sample that is selected under these conditions is a *probability sample*. If we assume that we obtain the true values of the survey items for all sample elements, we can estimate the values of population characteristics without bias based on these values.

The probability that an element is selected in the sample is $\pi_i = P(\delta_i = 1) = E(\delta_i)$. This is the first order inclusion probability for element i, $i = 1, 2, \ldots, N$. The $N \times 1$-vector of first order inclusion probabilities is denoted by $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)'$. The second order inclusion probability of elements i and k is $\pi_{i,k} = E(\delta_i \delta_k)$. By definition, $\pi_{i,i} = \pi_i$. The most straightforward way to select a random sample arises when each element has the same probability of being selected. Such a random sample is called a *simple random sample*. Sampling can be done with and without replacement. Throughout this thesis we assume that the sample is selected without replacement.

A *sampling design* is a set of specifications which defines the target population, the sampling elements and the probabilities attached to possible samples, see Särndal et al. (1992). Auxiliary information can be used in the design of the sample, leading for example to cluster sampling, stratified sampling or two-stage sampling.

1.2.3 Estimators using auxiliary information

Population characteristics can be estimated based on the sample. An estimator is a mathematical function of the observed data that attempts to approximate a population parameter as closely as possible. The estimator may also employ auxiliary information, either from the sampling frame or from external registers. In this section, we describe estimators for the population mean that use auxiliary information. Estimators for other population characteristics can be obtained in similar ways.

- *The Horvitz-Thompson estimator*
  The Horvitz-Thompson estimator uses inclusion probabilities. Auxiliary

information from the sampling frame can be used to compute these probabilities. Horvitz and Thompson (1952) show that it is always possible to construct an unbiased estimator for the population mean when the inclusion probabilities $\boldsymbol{\pi}$ are known and nonzero. The Horvitz-Thompson estimator is a mean of survey items, weighted by the inverse of the inclusion probabilities. This ensures that elements that are over-represented in the survey, i.e. that have a large inclusion probability, receive a smaller weight, and vice versa. In appendix 1.A we give a detailed description of the Horvitz-Thompson estimator.

- *The generalized regression estimator*
  Särndal (1980) and Bethlehem and Keller (1987) introduce the generalized regression estimator, or GREG-estimator. The GREG-estimator is equal to the Horvitz-Thompson estimator plus an additional adjustment term. The additional adjustment term is calculated with auxiliary information. By using auxiliary information in the estimator, a more accurate estimate of the population mean is obtained. Therefore, it is required that the auxiliary variables are available for the sample, and that their totals or means are known for the population of interest. Based on the sample, the values of the totals for the auxiliary variables in the population can be estimated. If there is a difference between the sample-based estimated totals and the true totals, the generalized regression estimator then adjusts the Horvitz-Thompson estimator for the observed difference. If the auxiliary variables are correlated with the survey item, i.e. it is plausible that a similar difference exists for the survey item in the sample and the population, the adjustment to the Horvitz-Thompson estimator will lead to a reduction of the variance of the estimator. A detailed description of the GREG-estimator is given in appendix 1.B.

- *The calibration framework*
  The calibration framework is introduced by Deville and Särndal (1992) and Deville et al. (1993). The calibration estimator is a weighted mean of survey items. The calibration weights are determined by calibrating the auxiliary variables in the sample to known totals in the population. Hence, the generalized regression estimator also is a calibration estimator. In fact, Särndal and Lundström (2008) show that most familiar estimators can be expressed as calibration estimators. In appendix 1.C we describe calibration estimators in more detail.

For further reading on these estimators, see Cochran (1977), Särndal et al. (1992) and Bethlehem (1988).

1.3 Surveys in practice

1.3.1 Sources of error

In the previous sections, we outlined the requirements for sample surveys to properly estimate population characteristics. Potentially, the estimated population characteristics are subject to error. This error can have many causes. The ultimate result of all these errors is a discrepancy between the survey estimate and the true population characteristic. This discrepancy is called the total survey error, see Biemer and Lyberg (2003). Bethlehem (1999) gives a taxonomy of survey errors, displayed in figure 1.1. Two broad categories can be distinguished contributing to this total error: sampling errors and non-sampling errors. Sampling errors can be divided into estimation errors and selection errors. *Estimation errors* arise because only a subset of the population is surveyed. This type of error disappears if the entire population is surveyed. *Selection errors* occur when the actual inclusion probabilities differ from the true values of the inclusion probabilities, for instance when an element has multiple entries in the sampling frame. Consequently, sampling errors vanish when the entire population is surveyed and can be controlled for by using a correct sampling design.

Non-sampling errors do not vanish when the entire population is surveyed. Non-sampling errors can be divided in errors caused by erroneous observation of sampled elements and non-observation errors. *Erroneous observation* arises due to overcoverage or measurement errors. In case of *overcoverage*, the error is caused by a discrepancy between the sampling frame and the actual population. Overcoverage occurs when sample elements that are not in the population, have an entry in the sampling frame. These sample elements are observed whereas they should not have been. A *measurement error* is a discrepancy between the reported or measured value and the true value. This discrepancy can be caused by the respondent, the interviewer or the questionnaire design. Measurement errors also arise when the concept implied by the survey question, and the concept that should be measured in the survey are different. Biemer and Lyberg (2003) refer to this type of error as a specification error.

Non-observation errors are errors that affect the inclusion probabilities. They are caused by undercoverage or nonresponse. In case of *undercoverage*,

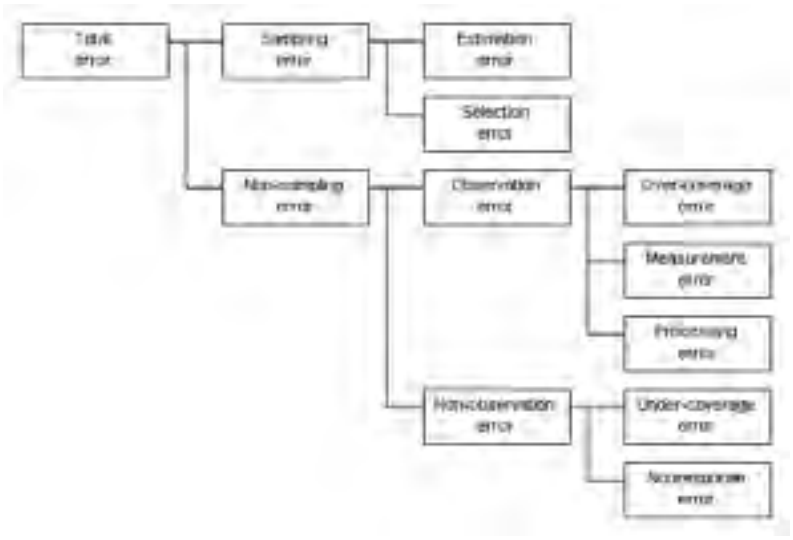

Figure 1.1: *A decomposition of the total survey error*

the sample element does not have an entry in the sampling frame whereas it is part of the population. The sample element cannot be observed and the selection probability is zero. There are two types of *nonresponse*. The first type is item nonresponse, which occurs when a sample element participates in the survey but refuses to answer some of the questions. The second type is unit nonresponse. Unit nonresponse occurs when selected sample elements do not provide the requested information. Due to nonresponse, the inclusion probability is lower than the design based inclusion probability. How much lower is unknown.

For a particular survey item Y, the set of elements that is missing consists of unit nonresponse and an additional set of item nonresponse. Although item nonresponse may have implications for survey estimates, in this thesis we focus on unit nonresponse alone. We thus disregard missing observations due to item nonresponse. For a discussion on item nonresponse we refer to Little and Rubin (2002).

1.3.2 Missing data in surveys

Nonresponse is a source of missing data. The estimators that adjust for nonresponse bias rely on assumptions about the mechanism behind the missing data. In this section we discuss missing data mechanisms, see also Little and Rubin (2002) and Schafer and Graham (2002). The missing data mechanism concerns the reasons why survey items are missing and the relationship between these reasons and the auxiliary variables. The missing data mechanism plays a crucial role in the analysis of data with missing values due to nonresponse. The properties and the success of adjustment techniques for nonresponse bias strongly depend on the nature of the dependencies in these mechanisms. Any analysis of data involving unit nonresponse requires an assumption about the missing data mechanism.

We can formalise the missing data mechanism as follows. We partition the dataset $\mathbf{Y}$ in an observed part, $\mathbf{Y}_{obs}$, and a missing part, $\mathbf{Y}_{mis}$. So,

$$\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) \tag{1.7}$$

The response indicators $R_1, R_2, \ldots, R_N$ indicate which survey items are available and which are missing. In case of unit nonresponse, R_i is a binary variable that indicates for each sample element i whether survey items are observed ($R_i = 1$) or missing ($R_i = 0$).

The distribution of the missingness is characterised by the conditional distribution of $\mathbf{R} = (R_1, R_2, \ldots, R_N)'$ given $\mathbf{Y}$, that is

$$P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) \tag{1.8}$$

When the conditional distribution of $\mathbf{R}$ given $\mathbf{Y}$ does not depend on the data at all, then $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R})$. The data are missing-completely-at-random (MCAR). When the conditional probabilities of missingness depend on the observed data, but not on the missing values, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs})$, the missing data are said to be missing-at-random (MAR). And finally, when (1.8) cannot be simplified any further and the distribution of the missingness depends on both observed and missing data, the mechanism is called not-missing-at-random (NMAR). When it comes to likelihood-based inference, MCAR and MAR are considered to be ignorable missing data mechanisms, whereas NMAR is nonignorable, see Little and Rubin (2002).

Let $\mathbf{X}$ represent a set of auxiliary variables that are completely observed and $\mathbf{Y}$ a survey item that is partly missing. $\mathbf{Z}$ represents causes of missingness unrelated to $\mathbf{X}$ and $\mathbf{Y}$, and $\mathbf{R}$ is the response indicator that indicates the missingness.

Figure 1.2: *Missing data mechanisms*

Figure 1.2 gives a graphical representation of the missing data mechanisms. In case of MCAR, missingness is caused by a phenomenon **Z** that is completely unrelated to **X** and **Y**. Estimates for **Y** are unbiased. In case of MAR, missingness is caused partly by an independent phenomenon **Z** and partly by the auxiliary variables **X**. So, there is an indirect relationship between **Y** and **R**. This leads to biased estimates for **Y**. Fortunately, it is possible to adjust for such a bias by using a technique that takes advantage of the availability of all values of **X**, both for respondents and nonrespondents. In case of NMAR, there may be a relationship between **Z** and **R** and between **X** and **R**, but there is also a direct relationship between **Y** and **R** that cannot be accounted for by **X**. This situation also leads to biased estimates for **Y**. In general, adjustment techniques using **X** will not be able to remove the bias. So, in case the missing data mechanism is MCAR or MAR it is still possible to compute unbiased estimates for population characteristics. However, a different approach is required when the mechanism is NMAR.

Example 1.3.1 Missing data mechanisms
Consider the case of unit nonresponse in a Labour Force Survey. Suppose there is one auxiliary variable X, being educational level, that has been measured for every sample element. The survey item Y is employment status and is missing for nonrespondents. MCAR means that the probability of a response does not depend on the educational level, nor on the employment status. So, MCAR means that the probability that persons provide information on their employment status is the same for all persons, irregardless of their educational level or their employment status. When the response is unrelated to the employment status, but does depend on the educational level, then the response mechanism

is MAR. For instance when persons with a high educational level are less inclined to participate in the survey than lower educated persons. The missing data then is missing-at-random with respect to education. When the missing data mechanism is NMAR, it means that the probability of a response depends on the values of the employment status as well. For instance, if persons do not have time to participate in the survey because they are at work. ■

In the survey literature, most methods that deal with unit nonresponse assume that the missing data mechanism is MAR. Throughout this thesis we will make the same assumption.

1.3.3 Different response types

Nonresponse is a common feature of sample surveys. A part of the sampled elements cannot be contacted, refuses participation or does not participate in the survey for other reasons. It is important to distinguish between these types because they require different measures to reduce nonresponse and may differently affect survey estimates.

There are many ways to classify nonresponse according to cause. This complicates computing response rates and comparing nonresponse for different surveys. The American Association for Public Opinion Research (AAPOR) has published a comprehensive list of definitions of possible survey outcomes, see AAPOR (2006). These definitions apply to household surveys with one respondent per household, samples selected by means of Random Digit Dialling (RDD) and mail- or Internet surveys of specific persons. Lynn et al. (2002) have proposed a classification that follows the possible courses of events when sample elements are approached with the survey request.

The response process can be seen as a hierarchical (nested) process with three stages. First, the eligibility of the sample elements is determined. Eligible elements are approached for participation in the survey. Once contacted, this results in either an interview, a refusal or another form of non-interview. However, sometimes the non-eligibility can only be determined when contact is made. Bethlehem et al. (2006) further extend the hierarchical representation of the response process by including a separate stage for being able to participate. They assume that it is possible to determine the eligibility of all contacted elements. We extend this representation with an additional stage before contact is attempted, to indicate whether a sample element is being processed or not.

We restrict ourselves to the eligible sample elements. The different outcomes of the survey process are:

Unprocessed cases Also referred to as administrative nonresponse. Unprocessed cases are sample elements that have not been approached in the field. This can be caused by a lack of interviewer capacity due to a high interviewer workload, or interviewer unavailability (illness or holiday).

Non-contact Once processed, sample elements are contacted during the fieldwork period. Contact (AAPOR 2006) can be defined as 'reaching some responsible housing unit member'. So this involves human contact to either the head of the household or the partner. Luiten (2008) defines contact for Statistics Netherlands as: either a face-to-face meeting, intercom talk (CAPI) or a telephone conversation (CATI) with the sample element. But contact is also obtained when the sample element has contacted the help desk from Statistics Netherlands.

Not-able Once processed and contacted, a longtime illness or a language problem can prevent sample elements from being able to participate.

Refusal Once processed, contacted and found able to participate, the sample element refuses to answer the survey questions.

Response A selected sample element is processed, contacted and able to participate and does not refuse the survey request. The sample element hence provides the requested information.

We make a distinction between not-able due to language problems and not-able due to a longtime illness because the underlying reasons for these two groups are distinct.

1.3.3.1 The sequential structure of the response process

For the different stages in the response process, we introduce individual probabilities for proceeding to the next stage of the response process, see table 1.1. The participation probability ϱ is not to be interpreted as the response probability. The final response probability ρ is the product of the probabilities of the different stages of the process, i.e. $\rho = \xi \times \gamma \times \theta \times \varrho$. In words, the response probability ρ is the product of the individual probabilities for being processed ξ, contacted γ, able to participate θ and participation ϱ. The response process can be graphically displayed as in figure 1.3. This figure shows the sequential (nested) structure of the response process, i.e. the outcome of one stage is conditional on the outcome of all previous stages.

Table 1.1: Probabilities in the response process

Process	Indicator	Values	Probability
Being processed	U	1 processed, 0 unprocessed	ξ
Contact	C	1 contact, 0 no contact	γ
Being able (language)	L	1 able, 0 not able	θ
Participation	P	1 participation, 0 refusal	ϱ

1.3.4 Implications of unit nonresponse

The quality of survey estimates can be expressed in terms of the Root Mean Squared Error (RMSE). The RMSE of an estimator for some population parameter for Y, denoted $\hat{y}$, can be decomposed as

$$RMSE(\hat{y}) = \sqrt{B^2(\hat{y}) + S^2(\hat{y})} \tag{1.9}$$

where $B(\hat{y})$ denotes the bias of the estimator. The variance of estimators is determined by two components: variance due to sampling and variance due to nonresponse. Both sampling and nonresponse lead to a smaller number of observations and thus to less accurate estimators. In case of full response, the variance of the estimator approaches zero when the sample size of the survey increases.

Besides that nonresponse increases the variance, it also introduces a bias. For large surveys, like most surveys conducted by national statistical offices, the increase in variance due to nonresponse is usually small compared to the variance due to sampling. The nonresponse bias is of greater concern. It does not vanish when the sampling size is increased.

To give insight in the implications of nonresponse in sample surveys, we employ the Random Response model to incorporate nonresponse in the sampling theory. Based on this model, we can show the effect of nonresponse on estimates for population characteristics.

1.3.4.1 The Random Response model

We show how nonresponse can be introduced in the classical sampling theory by means of the Random Response model. In the Random Response model, every element $i \in U$ has a nonzero, unknown response probability, which we will denote by ρ_i. If element i is selected in the sample, a random mechanism

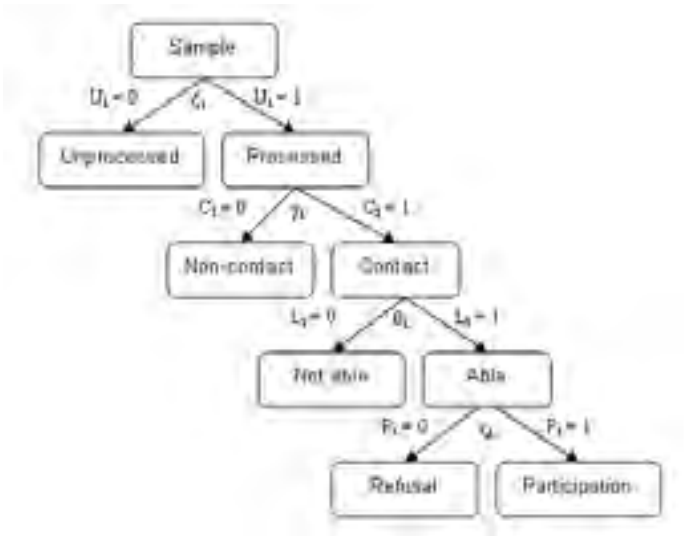

Figure 1.3: *The nested response process*

is activated that results with probability ρ_i in response and with probability $(1 - \rho_i)$ in nonresponse.

The response probabilities ρ_i are theoretical quantities. The definition of a 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 held. For example, the number and timing of contact attempts and the interviewer characteristics. If these circumstances change, it is very likely that the individual response probabilities also change. In addition, response probabilities vary over time. However, the more conditions are fixed, the less random the response probabilities become. In example 1.3.2 we discuss the Fixed Response model that arises as a special case of the Random Response model when the response probabilities are viewed conditional on very detailed circumstances. No variation is left and the response probabilities become deterministic. In case no conditions are imposed, the response probabilities are stochastic. The Random Response model involves response probabilities that are conditional only on characteristics of the sample elements.

Under the Random Response model, we can introduce a vector of response

indicators $\mathbf{R} = (R_1, R_2, \ldots, R_N)'$, where $R_i = 1$ if the corresponding element i responds, and where $R_i = 0$ otherwise. So, $P(R_i = 1) = \rho_i$, and $P(R_i = 0) = 1 - \rho_i$. The response probabilities ρ_i are unknown. Furthermore, the indicators R_i are observed for sample elements only.

These probabilities can be estimated based on the sample. By using an appropriate model based on auxiliary information, we can compute sample-based estimates of the response probabilities, i.e.

$$\hat{\rho}_i = \rho(\mathbf{X}_i) = P(R_i = 1 | \delta_i = 1; \mathbf{X}_i) \tag{1.10}$$

for $i = 1, 2, \ldots, n$. Hence, $\hat{\rho}_i = \hat{\rho}_j$ if $\mathbf{X}_i = \mathbf{X}_j$, i.e. person i has the same probability of response as all other persons in the same strata defined by $\mathbf{X}$. We refer to $\hat{\rho}_i$ as the *response propensity*. The response propensity is the estimated response probability conditional on the sample and the individual characteristics $\mathbf{X}_i$. In the appendix to Chapter 5 we describe methods that can be used to compute response propensities.

Example 1.3.2 The Fixed Response model
The Fixed Response model is a special case of the Random Response model. It occurs when the response probabilities ρ_i are dependent on very detailed circumstances of the survey. Then there is no variation left and the response probabilities become deterministic instead of stochastic. Under the Fixed Response model, sample elements are assumed to have a fixed response behaviour. Either they always cooperate, or they never do. The population can be divided into two strata: a response stratum and a nonresponse stratum. These strata are mutually exclusive and exhaustive, i.e. every element belongs to exactly one of these strata. Beforehand, we do not know to which stratum an element belongs. A sample selected from the population contains elements from both strata. This ultimately results in a subsample of respondents and a subsample of nonrespondents. Associated with every element i is a response indicator R_i. $R_i = 1$ in case i belongs to the response stratum and 0 otherwise. The size of the response stratum equals $N_r = \sum_{i=1}^{N} R_i$ and the size of the nonresponse stratum $N_{nr} = \sum_{i=1}^{N} (1 - R_i)$. The total size of the population thus equals $N = N_r + N_{nr}$. ■

1.3.4.2 Nonresponse bias

Under the Random Response model, we observe the survey items for every element i with $R_i = 1$. Now suppose we select a simple random sample without

replacement of size n from the population U. The response only consists of those elements i for which $\delta_i = 1$ and $R_i = 1$. Hence, the size of the subsample of respondents is equal to

$$n_r = \sum_{i=1}^{N} \delta_i R_i \tag{1.11}$$

Note that this realised sample size is a random variable. The size of the subsample of nonrespondents is equal to

$$n_{nr} = \sum_{i=1}^{N} \delta_i (1 - R_i) \tag{1.12}$$

The total sample size equals $n = n_r + n_{nr}$. We only obtain survey items for the n_r responding elements. The mean of these survey items is denoted by

$$\overline{y}_r = \frac{1}{n_r} \sum_{i=1}^{N} R_i Y_i \tag{1.13}$$

Theoretically, it is possible that no observations at all become available. This happens when none of the sample elements responds. In practical situations, this event has a very small probability of happening. Therefore, we will ignore it. Bethlehem (1988) shows that the expected value of the response mean is approximately equal to

$$E(\overline{y}_r) \approx \tilde{Y} \tag{1.14}$$

where

$$\tilde{Y} = \frac{1}{N} \sum_{i=1}^{N} \frac{\rho_i}{\overline{\rho}} Y_i \tag{1.15}$$

and

$$\overline{\rho} = \frac{1}{N} \sum_{i=1}^{N} \rho_i \tag{1.16}$$

is the mean of all response probabilities in the population.

From expression (1.14) 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. We refer to this bias as the *nonresponse bias*. Bethlehem (1988) shows that it is approximately equal to

$$B(\overline{y}_r) = \tilde{Y} - \overline{Y} = \frac{R(\boldsymbol{\rho}, \mathbf{Y}) S(\boldsymbol{\rho}) S(\mathbf{Y})}{\overline{\rho}} \tag{1.17}$$

where $R(\boldsymbol{\rho}, \mathbf{Y})$ is the correlation between survey items $\mathbf{Y}$ and response probabilities $\boldsymbol{\rho}$, $S(\mathbf{Y})$ is the standard deviation of survey items $\mathbf{Y}$, and $S(\boldsymbol{\rho})$ is the standard deviation of the response probabilities. From this expression of the nonresponse bias we can draw a number of conclusions:

- The bias vanishes if there is no relationship between the survey item and response behaviour. Then $R(\boldsymbol{\rho}, \mathbf{Y}) = 0$. The stronger the relationship between survey item and response behaviour, the larger the bias will be.
- The bias vanishes if all response probabilities are equal. Then $S(\boldsymbol{\rho}) = 0$. In this situation the nonresponse is not selective. It just leads to a reduced sample size.
- The bias vanishes if all survey items are equal. Then $S(\mathbf{Y}) = 0$. In this situation, there is no difference between respondents and nonrespondents with respect to Y, and hence there is no bias.
- The magnitude of the bias increases as the mean of the response probabilities decreases (assuming the relationship between survey item and response behaviour is unaltered). Translated in practical terms, this means that lower response rates will lead to larger biases.
- Unlike the variance, the nonresponse bias does not approach zero when the sample size is increased.

Under the Fixed Response model, the nonresponse bias can be expressed as (Groves and Couper, 1998)

$$B(\overline{y}_r) = \frac{N_{nr}}{N}(\overline{Y}_r - \overline{Y}_{nr}) \tag{1.18}$$

where capital letters denote population equivalents, $\overline{Y}_{nr}$ is the population equivalent of the mean of survey item Y among nonrespondents. This bias is composed of the response rate and the difference between respondents and nonrespondents with respect to Y. If the response rate decreases and the difference between respondents and nonrespondents is unaltered, the bias increases as can be seen from expression (1.18).

1.3.5 The importance of auxiliary information

We defined auxiliary variables in section 1.2.1 as variables that are in general available prior to sampling. The Netherlands, as well as the Scandinavian countries and some other West-European countries, is privileged with the availability of many reliable registers. The population register serves as the backbone, or sampling frame. In Chapter 3 we describe the population register in more detail. Statistics Netherlands can link this register to other registers, thereby exploiting coherent information on the registered population. For surveys conducted by Statistics Netherlands, the auxiliary variables are usually available for every element of the sample as well as the target population. They can be used in creating a sampling design or in the computation of survey estimates as we have shown in section 1.2.3.

Other organisations, for example universities or market research organisations, usually do not have access to these registers. However, Statistics Netherlands publishes population distributions for specific domains on the internet database Statline. This information is publicly available. Other organisations can use this information in the estimation stage where it may be sufficient to know population characteristics of auxiliary variables, while individual values are known for respondents only.

If used efficiently, auxiliary information can increase the accuracy of estimates of population characteristics, as we have discussed in section 1.2.3. The GREG-estimator is one example of such an estimator. But the auxiliary information can also serve another purpose; to adjust for the effects of nonresponse. In fact, the key to successfully adjusting for nonresponse bias lies in the use of powerful auxiliary information. In Chapter 6 and 8 we describe methods that use auxiliary variables to adjust for nonresponse bias.

In the most ideal situation, the vector of auxiliary variables $\mathbf{X}$ has three features (Särndal and Lundström, 2008): It is related to the response indicators $\mathbf{R}$, it is related to the survey item Y, and it identifies the domains on which population characteristics are published. If the auxiliary variables used in the estimation of population characteristics have a strong relationship with the phenomena to be investigated, as well as the response behaviour, the weighted sample will be approximately representative with respect to these phenomena, and hence estimates of population characteristics will be more accurate and less biased. Algorithms to select auxiliary variables are described in Schouten (2007) and Särndal and Lundström (2008).

1.4 Overview of the thesis

In this chapter we have described the classical sampling theory and we have incorporated nonresponse in the theory of sample surveys. Nonresponse is a common feature of sample surveys and efforts to keep nonresponse rates acceptable have increased, see for example De Heer (1999), Stoop (2004). The increasing nonresponse is caused by a deteriorating survey climate, combined with a decreased willingness to participate in surveys. From the survey literature it is well known that nonrespondents are different from respondents with respect to demographic and socio-economic information. Moreover, the reason for nonresponse and the topic of the survey are often related.

Nonresponse reduces the sample size and the accuracy of survey estimates. Furthermore, selective response with respect to the survey topic increases the risk of a systematical nonresponse bias in survey estimates. Nonresponse hence seriously threatens the quality of survey statistics. Over the years, the treatment of nonresponse has therefore received a lot of attention in survey methodology. Broadly two aspects of nonresponse can be distinguished: the prevention of nonresponse before it occurs in the field, and estimation techniques that aim at adjusting for nonresponse bias after the data have been collected.

The prevention of nonresponse, or nonresponse reduction, is aimed at achieving a higher response rate. The higher the response rate, the lower the risk of a nonresponse bias in survey estimates, see section 1.3.4. Research in this area is for a large part based on behavioural theories. Measures to enhance survey participation can be taken before and during the data collection period. A recent trend in this area is a shift from a focus on increasing response rates to a focus on obtaining a more balanced composition of the response. A balanced composition refers to a composition of characteristics within the response that is similar to the composition in the population. More effort is spend at obtaining response from difficult respondents, instead of easy respondents which results in 'more of the same' and does not improve the composition of the response.

The tendency of research on nonresponse reduction to balance the composition of the response resembles nonresponse adjustment, where the difference is in the timing. Nonresponse adjustment balances the composition after the data have been collected. Traditionally, nonresponse adjustment research focussed on the development of statistical methods to adjust for nonresponse bias. The basic idea of adjusting is to assign weights to respondents in such a way that they represent nonrespondents as well. This involves calibration of respondents' characteristics to known characteristics of the population. A large part of this research is devoted to modelling response behaviour. Thus far, the attempts of

modelling this behaviour have not been very successful. The influence of the survey design and topic complicate consistent prediction of response behaviour. For that reason, nonresponse adjustment research gained interest in the data collection process. It has become common practice to include information that is collected during the data collection process in models to predict response behaviour.

Both nonresponse reduction and the adjustment for nonresponse bias profit from the analysis of response behaviour. Hence, the usefulness of nonresponse analysis is twofold. In the first place, it reveals how respondents differ from nonrespondents. This information is useful for the construction of weighting models to adjust for nonresponse bias, as well as for the construction of models to predict response behaviour. At the same time, analysis of nonresponse points at characteristics of the sample that have become unbalanced due to nonresponse. The nonresponse can be efficiently reduced by concentrating nonresponse reduction efforts on these unbalanced groups. Through the analysis of nonresponse, research on nonresponse adjustment hence approaches research on the reduction of nonresponse, and vice versa.

In this thesis, we discuss methods for nonresponse analysis and adjustment for nonresponse bias. We thereby focus on different sources of nonresponse, i.e. being processed, contact, being able and participation as described in section 1.3.3. The first chapters describe approaches to obtain information on nonrespondents. This can be done by linking data from external registers to the survey to analyse response behaviour. Another way to obtain information on nonrespondents, is by asking them for it. There are several strategies to re-approach nonrespondents that can be employed for this purpose and which we have applied to the Dutch LFS. In addition, we have used the linked data to evaluate these strategies. However, despite efforts to prevent nonresponse, or to obtain information about or from nonrespondents, it is commonly accepted that some nonresponse will always remain. To judge the quality of the obtained response, we propose to consider the representativeness of the response by means of a representativity indicator, or R-indicator. Additionally, we present methods to analyse nonresponse as well as estimation techniques to adjust for nonresponse bias that account for different sources of nonresponse. We show how the adjustment techniques can also be applied to simultaneously adjust for errors due to undercoverage and nonresponse. Furthermore, we describe methods that can be used to combine data from different modes, thereby focusing on nonresponse bias adjustment.

This thesis is outlined as follows. Chapters 2, 3 and 4 focus on the analysis of nonresponse, with an emphasis on the distinction between different types

of nonresponse. In Chapter 5 we present an indicator for survey quality that employs information from the analysis of nonresponse and that can be used to reduce nonresponse. Methods to adjust for nonresponse bias are described in Chapters 6 to 9.

In Chapter 2 we summarize the behavioural theories underlying survey response. The survey literature deals almost exclusively with contact and participation, therefore in Chapter 2 we distinguish between these two types of response. The Netherlands has access to a large number of reliable registers. Based on linked data from these registers, it is possible to analyse the survey response and to identify demographic and socio-economic correlates of response that distinguish between different types of response. In Chapter 2, we provide an overview of the causes and correlates of contact and participation as identified by the survey literature. We then use the wealth of information from registers to analyse the response behaviour in the Dutch LFS in Chapter 3. Thereby, we distinguish between the four different types of response. The LFS is performed in many other countries in ways similar to the Dutch LFS. The analysis of the Dutch LFS therefore provides insight into response behaviour that can be informative for other countries that conduct the LFS but that cannot link their survey to information from registers.

In Chapter 4, two strategies to re-approach nonrespondents are described, namely the basic-question approach and the call-back approach. These approaches have been applied to the Dutch LFS in a pilot from July to October 2005. We extensively analyse the data from these two approaches with the use of linked data. Here, the linked data serve to evaluate the approaches in their effectiveness of obtaining a more balanced composition of the response.

In Chapter 5 we provide a mean to use information from the analysis of nonresponse for nonresponse reduction. We present a measure that describes how balanced the composition of the response is compared to the composition of the sample, for a set of prescribed characteristics. Therefore, we give a definition of representativeness and we introduce the representativeness indicator or R-indicator. The R-indicator is a result of the recent developments in the two areas of nonresponse research. It serves as a counterpart of the response rate. Furthermore, it is possible to use the R-indicator as a management tool to point at unbalanced groups in the response.

The second approach to deal with nonresponse is the adjustment for nonresponse bias in survey estimates. Traditional methods are based on the idea of adjustment weighting. The respondents are assigned weights, using auxiliary data from the sampling frame or external registers. The weights are constructed in a clever way, so that the respondents represent the nonrespondents as well.

In Chapter 6 we give an overview of traditional nonresponse adjustment techniques. In Chapter 7, we make a little side step from the problem of nonresponse by showing how these traditional methods can be applied to another missing data problem, namely the missing data in telephone surveys. In telephone surveys undercoverage of sample elements without a listed telephone leads to an additional source of missing data besides nonresponse. We show how the techniques from Chapter 6 can be adapted to adjust for undercoverage, as well as a simultaneous adjustment for undercoverage and nonresponse bias. The adapted techniques are then applied to the Integrated Survey on Household Living Conditions.

The techniques in Chapters 6 and 7 do not account for the sequential structure of the response process that arises due to the different types of response. For instance, the nesting of contact and participation that arises due to the fact that participation can only be obtained for sample elements that have been contacted first. Chapter 8 is therefore devoted to the development of methods that distinguish between different response types. In Chapter 8 we describe both methods to analyse response behaviour and estimation methods that adjust for nonresponse bias. The methods differ in the way that they include different response types and account for the sequential structure of the response process.

Chapter 9 describes data collection with multiple modes, or mixed-mode. Mixed-mode data collection offers attractive ways to reduce survey costs, as well as possibilities to enhance response. In Chapter 9 we discuss issues regarding mixed-mode data collection, with an emphasis on the estimation of population characteristics based on data collected by different modes.

Finally, in Chapter 10 conclusions and directions for further research are given.

Appendices

1.A The Horvitz-Thompson estimator

Horvitz and Thompson (1952) show that, when the inclusion probabilities $\boldsymbol{\pi}$ are known and nonzero, an unbiased estimator for the population mean is

$$\overline{y}_{ht} = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i Y_i}{\pi_i} \tag{1.19}$$

The estimator (1.19) uses the auxiliary information from the sampling frame through the inclusion probabilities π_i. Only selected elements are included ($\delta_i = 1$), and the value for the survey item Y_i is weighted by the inverse of the first order inclusion probability π_i. This ensures that elements that are over-represented in the survey, i.e. that have a large inclusion probability, receive a smaller weight, and vice versa. Let $d_i = \frac{1}{\pi_i}$ be the design weight. Since $\delta_i = 0$ for elements not in the sample, we can rewrite (1.19) as

$$\overline{y}_{ht} = \frac{1}{N}\sum_{i\in s} d_i Y_i \tag{1.20}$$

That the Horvitz-Thompson estimator (1.19) is an unbiased estimator for the population mean $\overline{Y}$ in case of full response follows by taking the expectation

$$E(\overline{y}_{ht}) = \frac{1}{N}\sum_{i\in U} d_i Y_i E(\delta_i) = \overline{Y} \tag{1.21}$$

If the sample elements are sampled without replacement, the variance is equal to

$$\begin{aligned} var(\overline{y}_{ht}) &= E(\overline{y}_{ht} - \overline{Y})^2 = E(\overline{y}_{ht})^2 - \overline{Y}^2 \\ &= \tfrac{1}{N^2}\Big(\textstyle\sum_{i\in U}\frac{Y_i^2}{\pi_i^2}(E(\delta_i))^2 + \sum_{i\neq k}\frac{Y_i Y_k}{\pi_i \pi_k}E(\delta_i\delta_k)\Big) - \overline{Y}^2 \\ &= \tfrac{1}{N^2}\Big(\textstyle\sum_{i\in U} Y_i^2 + \sum_{i\neq k}\frac{Y_i Y_k}{\pi_i \pi_k}\pi_{i,k})\Big) - \overline{Y}^2 \\ &= \tfrac{1}{N^2}\textstyle\sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_{i,k} - \pi_i\pi_k)\frac{Y_i}{\pi_i}\frac{Y_k}{\pi_k} \end{aligned} \tag{1.22}$$

Furthermore, in case of a fixed sample size n, (1.22) can be rewritten as (Bethlehem, 1988)

$$var(\overline{y}_{ht}) = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_{i,k} - \pi_i\pi_k)\left(\frac{Y_i}{\pi_i} - \frac{Y_k}{\pi_k}\right)^2 \tag{1.23}$$

It follows from this expression that the variance can be reduced if the first order inclusion probabilities $\boldsymbol{\pi}$ are chosen proportional to the values of $\mathbf{Y}$. In practice $\mathbf{Y}$ is unknown. If, however, certain auxiliary variables $\mathbf{X}$ are related to $\mathbf{Y}$, the first order inclusion probabilities can also be chosen proportional to the values of $\mathbf{X}$ to reduce the variance of estimator (1.20).

1.B The generalized regression estimator

Särndal (1980) and Bethlehem and Keller (1987) introduce the generalized regression estimator, or GREG-estimator, to increase the precision of estimated population characteristics. Särndal (1980), as well as Särndal et al. (1992), follows a model-assisted approach whereas Bethlehem and Keller (1987) follow a design-based approach. As stated in section 1.2.1, we would like to use the sample to make inference about the population as a whole. Two different types of inference can be distinguished, see Särndal et al. (1992). The first type is inference about the finite population U itself, which we have described in the previous sections. This type of inference is referred to as *design-based*. The objective is to estimate characteristics of the population U. The second type is inference about a model or a superpopulation that is thought to have generated U. This type of inference is referred to as *model-assisted*. The interest lies in the process that underlies the finite population U. In the design-based approach, the finite population U is fixed. In the model-assisted approach, the population is generated by the superpopulation model described by

$$\begin{array}{ll} E(Y_i) & = \mathbf{X}_i'\boldsymbol{\beta} \\ var(Y_i) & = \sigma_i^2 \end{array} \tag{1.24}$$

for $i = 1, 2, \ldots, N$. Under this model, the values $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$ for the finite population U are realised values for $\mathbf{Y}$ based on the superpopulation model (1.25). $\mathbf{X}$ is an $N \times J$-matrix of values for the auxiliary variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)'$ is an $J \times 1$-vector of regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)'$ is an $N \times 1$-vector of error terms. The error terms $\boldsymbol{\epsilon}$ are independent random variables, i.e. $E(\epsilon_i) = 0, var(\epsilon_i) = \sigma_i^2$ and $E(\epsilon_i \epsilon_j) = 0; i \neq j$. It is often assumed that $\sigma_i^2 = 1$, $\forall i$.

Both the design-based approach and the model-assisted approach lead to the same GREG-estimator. The GREG-estimator can be derived from a linear regression of the values $\mathbf{Y}$ on $\mathbf{X}$, i.e.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.25}$$

Under the model-assisted approach, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)'$ is an $N \times 1$-vector of error terms. Under the design-based approach, the ϵ's are not random but $\boldsymbol{\epsilon}$ is a vector of residuals.

The general regression estimator can be computed with the sample observations and is commonly expressed as

$$\overline{y}_{gr} = \overline{y}_{ht} + (\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht})'\hat{\boldsymbol{\beta}} \tag{1.26}$$

where $\overline{y}_{ht}$ is the Horvitz-Thompson estimator of the survey item mean, $\overline{\mathbf{X}}$ is the $J \times 1$-vector of the population means for the auxiliary variables, $\overline{\mathbf{x}}_{ht}$ is the $J \times 1$-vector of the estimated sample means and $\hat{\boldsymbol{\beta}}$ is the Ordinary Least Squares estimator for $\boldsymbol{\beta}$, i.e.

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_J)' \\ &= \left(\sum_{i=1}^N \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \sum_{i=1}^N \mathbf{X}_i Y_i \\ &= (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{Y} \end{aligned} \tag{1.27}$$

We turn to the sample observations to estimate $\hat{\boldsymbol{\beta}}$, denoted by $\hat{\boldsymbol{\beta}}^{(s)}$

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(\mathbf{s})} &= (\hat{\beta}_1^{(s)}, \hat{\beta}_2^{(s)}, \ldots, \hat{\beta}_J^{(s)})' \\ &= \left(\sum_{i=1}^n d_i \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \sum_{i=1}^n d_i \mathbf{X}_i Y_i \\ &= (\mathbf{X}\boldsymbol{\Pi}^{-1}\mathbf{X}')^{-1}\mathbf{X}\boldsymbol{\Pi}^{-1}\mathbf{Y} \end{aligned} \tag{1.28}$$

where $\boldsymbol{\Pi}$ denotes the diagonal matrix of inclusion probabilities, i.e.

$$\boldsymbol{\Pi} = \begin{bmatrix} \pi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \pi_n \end{bmatrix} \tag{1.29}$$

and, hence, $\boldsymbol{\Pi}^{-1}$ denotes the matrix of design weights.

As follows from (1.26), the generalized regression estimator is equal to the Horvitz-Thompson estimator plus an additional adjustment term. The generalized regression estimator adjusts the Horvitz-Thompson estimator for the observed differences between the true values and the sample-based estimated values of the totals of the auxiliary variables. If the auxiliary variables are correlated with the survey item, i.e. there is a strong linear relationship, the additional term will be negatively correlated with the error term from the Horvitz-Thompson estimator thus leading to a reduction of the variance. Bethlehem (1988) shows

that the generalized regression estimator can be expressed as a mean of fitted values in the population

$$\overline{y}_{gr} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i = \overline{\mathbf{X}}'\hat{\boldsymbol{\beta}}^{(s)} \tag{1.30}$$

under the condition that

$$\boldsymbol{\lambda}'\mathbf{X}_i = 1 \tag{1.31}$$

for all $i = 1, 2, \ldots, N$, where $\boldsymbol{\lambda}$ is a $J \times 1$-vector of constants. This condition is satisfied if the regression model contains a constant term.

1.C The calibration framework

Deville and Särndal (1992) and Deville et al. (1993) introduce the calibration framework to estimate population characteristics. The calibration estimator is a weighted mean of survey items

$$\overline{y}_w = \frac{1}{N}\sum_{i=1}^{n} w_i Y_i \tag{1.32}$$

where $w_i = d_i g_i$. The g-weights are the correction weights, often referred to as g-weights. The weights $w_i = d_i g_i$ for the estimation of $\overline{y}_w$ are based on known auxiliary information. The weights have to satisfy the calibration equation

$$\sum_{i=1}^{n} w_i \mathbf{X}_i = \sum_{i=1}^{N} \mathbf{X}_i \tag{1.33}$$

When the weight system $w_i = d_i g_i$ for $i \in s$ is applied to auxiliary vector $\mathbf{X}_i$, and summed over the sample, an estimate for the population total of $\mathbf{X}$ is obtained. This estimate agrees exactly with the known value of that total, which leads to the concept of calibration to known population characteristics. The weights w_i are not uniquely defined by (1.33). Therefore, the difference between the weights w_i and the design weights d_i usually minimises some distance function.

Dependent on the type of auxiliary information, Särndal and Lundström (2008) show that most familiar estimators can be expressed as calibration estimators. The GREG-estimator, discussed in appendix 1.B can be expressed as a

calibration estimator. To derive an expression of the g-weights for the GREG-estimator, consider (1.26) based on the sample

$$\begin{aligned}\overline{y}_{gr} &= \overline{y}_{ht} + (\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht})'\hat{\boldsymbol{\beta}}^{(s)} \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^{n} d_i Y_i + (\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht})'\Big(\sum_{i=1}^{n} d_i \mathbf{X}_i \mathbf{X}_i'\Big)^{-1} \sum_{i=1}^{n} d_i \mathbf{X}_i Y_i \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^{n} \Big[1 + \Big(\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht}\Big)'\Big(\sum_{i=1}^{n} d_i \mathbf{X}_i \mathbf{X}_i'\Big)^{-1} \mathbf{X}_i\Big] d_i Y_i\end{aligned} \tag{1.34}$$

Consequently, the g-weights are expressed as

$$1 + \Big(\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht}\Big)'\Big(\sum_{i=1}^{n} d_i \mathbf{X}_i \mathbf{X}_i'\Big)^{-1} \mathbf{X}_i \tag{1.35}$$

When condition (1.31) holds, the g-weights can be simplified to

$$g_i = \Big(\sum_{i=1}^{N} \mathbf{X}_i\Big)'\Big(\sum_{i=1}^{n} d_i \mathbf{X}_i \mathbf{X}_i'\Big)^{-1} \mathbf{X}_i \tag{1.36}$$

From expressions (1.35) and (1.36) it becomes clear that the final weights in estimator (1.32) are not dependent on the survey item. The weights w_i, once computed, can be used for all survey items. This is convenient in practice, the procedures to estimate the survey characteristics only have to be performed once. However, it is also inflexible as different survey items all receive the same weight whereas not every auxiliary variable is as predictive for each survey item. Preferably, a different weighting model is constructed for each survey item.

Other calibration estimators are the post-stratification estimator, the ratio estimator and the regression estimator. The post-stratification estimator is obtained when the auxiliary variables are qualitative variables. In ratio estimation the weighting model consists of one quantitative weighting term. This quantitative variable can be crossed with other (qualitative) variables. The regression estimator is obtained when the auxiliary information consists of a constant 1 and a quantitative variable X_i.

Chapter 2

Causes and Correlates of Survey Nonresponse

How can we assess whether nonresponse causes estimators to be biased, taken into account that nonrespondents do not provide information? There are a number of ways to obtain information on the nonrespondents:

- *Linked data.* Sometimes it is possible to link the survey sample to one or more external registers, for instance a population register. From these registers, data can be retrieved for both respondents and nonrespondents.
- *Call-back approach.* A sample is selected from the nonrespondents, and re-approached by specially trained interviewers.
- *Basic-question approach.* Nonrespondents are persuaded to cooperate by offering them a much shorter questionnaire containing just a few basic questions.

The call-back approach and the basic-question approach are the focus of Chapter 4. In this chapter we give an overview of the survey literature on correlates of nonresponse that become available from linked data.

2.1 Introduction

In Chapter 1, we introduced four different types of response; being processed, contacted, able to participate and finally, participation. The corresponding types

of nonresponse are: unprocessed cases, non-contact, not able to participate and refusal. The response process can be displayed as a sequential process, see figure 2.1. A lot of literature is devoted to the identification of different sources of nonresponse. Most of the authors thereby focus on non-contact and refusal, e.g. Kish (1965), Wilcox (1977), O'Neill (1979), Stinchcombe et al. (1981), Goyder (1987), Groves and Couper (1998), Lynn et al. (2002), Nicoletti and Peracchi (2005), Stoop (2005). Other sources of nonresponse can be identified too, for example Bethlehem and Schouten (2004) consider nonresponse due to unprocessed cases, and inability due to language problems or longtime illness. In those cases, it is more straightforward what caused the nonresponse (e.g. interviewer unavailability, ethnicity, age).

According to Kish (1965), refusals should be considered separately from non-contacts, because they are less open to reduction with call backs and less tractable to generalizations. Wilcox (1977) brings up another justification for the independent treatment of the errors: non-contacts result from a lack of physical availability whereas refusals are attributable to a negative mental state. But, the combination of both is also possible; in which case they cannot be regarded separately. Lynn et al. (2002) find evidence that ease of contact and reluctance to participate are two distinct processes.

Groves and Couper (1998) present a general framework for participation in surveys. They provide theoretical models for contact and participation. Stoop (2005) presents an extensive overview of the causes and correlates of contact, and cooperation given contact, based on the survey methodology literature. She follows the theoretic models of Groves and Couper (1998).

In this chapter we provide an overview of the existing literature on the causes and correlates of nonresponse. As the literature deals almost exclusively with contact and participation, we focus on these two response types. This literature comprehends theories on the causes of contact and survey participation (e.g. Groves and Couper 1998, Stoop 2005). Furthermore, it identifies variables that are correlated with contact and participation. For instance, a cause of contact is the at-home behaviour of the household. A correlate of contact is a variable that is related to the at-home behaviour, for example the labour force status since working persons tend to be at-home less frequently.

In section 2.2 we discuss the theory on contact and participation put forward by survey methodologists. There exists a broad strand of disciplines from where theories can be applied to survey participation. The theory for contact is more straightforward than for participation. These theories can help to inform us what variables should be used in the analysis of different types of nonresponse. We focus on interviewer assisted surveys, namely Computer Assisted Personal

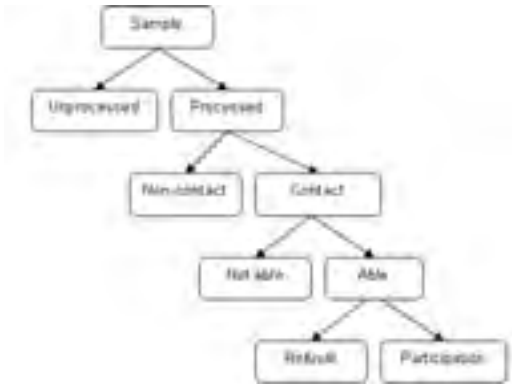

Figure 2.1: *Hierarchical representation of the response process*

Interviewing (CAPI), face-to-face interviews and Computer Assisted Telephone Interviewing (CATI). In self-administered surveys such as a paper or a web survey there is less information available regarding the different nonresponse types, see also Chapter 9. Section 2.3 summarises the main findings in the literature for correlates related to contact and participation.

2.2 Causes of contact and participation

2.2.1 Contact

Groves and Couper (1998) present a model that relates contact to three causes:

- *At-home pattern of the household*
- *Presence of physical impediments*
- *Call pattern of the interviewers*

When it comes to the causes of contact, the literature seems to be unambiguous. The probability of making contact is strongly related to the probability that some member of the household is at home (1) when the interviewer calls (3). Physical impediments (2) make life harder for the interviewer. Both in face-to-face surveys (e.g. a locked central entrance to an apartment building, or an intercom) and telephone surveys (e.g. answering machines or number identification) physical impediments result in lower contact rates. In a face-to-face

survey, the impediments can be observed prior to the first contact. Households that have some sort of impediment have a lower probability of contact on the first attempt, and more households with impediments remain non-contacted at the end of the fieldwork period.

The call pattern concerns the timing and the frequency of calls made by the interviewer. Research on at-home patterns of households shows that during the evening, the probability of finding someone at home is much larger than during daytime (Groves and Couper, 1998). For various reasons however, the percentage of calls that is made during the evening is smaller than during daytime. Of course, evening time is a shorter period than daytime. In a telephone survey where all calls are distributed to the interviewers by a central calling system it is logical that the number of evening calls is smaller than the number of daytime calls. In a face-to-face survey the interviewer makes the decision when to visit a household. A possible reason for the smaller proportion of visits made during the evening might be that interviewers feel less safe during evenings and therefore prefer daytime visits.

2.2.2 Survey participation

Theories that can be used to explain survey participation are found in both the sociological and the psychological literature. Goyder (1987) describes two psychological views on the phenomenon of nonresponse in survey research: behavioural and voluntary. Survey researchers with a behavioural view of nonresponse believe that survey participation can be influenced by the survey design, and depends on socio-demographic characteristics, or social location, of the respondents. In the voluntary view, the respondents themselves decide whether or not they want to participate. This decision is hardly influenced by the survey researcher. In extreme, the voluntary theory resolves into a view that persons act for reasons best known to themselves. The behavioural extreme would be that there is no conscious decision-making left within the social context of a survey request.

Blau (1964) describes a theory of social exchange that both Dillman (1978) and Goyder (1987) translate to the survey participation context. The social exchange theory (Blau, 1964) 'considers the perceived value of the equity of long-term associations between persons, or between a person and societal institutions'. Social exchange theory assumes that relationships between humans are based on a subjective cost-benefit analysis, and the comparison of alternatives.

Goyder (1987) finds a synthesis in the social exchange theory for both the behavioural and the voluntary viewpoint. He describes it as a number of obliga-

tions and expectations over a longer period of time between an individual and various institutions of society. Individuals balance cost and reward for survey participation, in which their behaviour is based on rational decisions. Dillman (1978) takes a more narrow view on the social exchange by describing the social exchange between a survey organization and a householder.

Groves et al. (2000) have developed a theory of interaction between survey design features and socio-economic characteristics of the sample person. Thereby, they adhere to the behavioural viewpoint described by Goyder (1987). Their theory stipulates that individual persons assigns different values to survey design features. These values are determined by the socio-economic characteristics of the sample person. The interviewer tries to make some of the survey design features more salient to the sample person. The final response decision is determined by the value that the sample person attributes to the survey design features, multiplied by the saliency that the features has for the sample person. Groves et al. (2000) refer to this theory as the 'leverage-salience' theory.

Example 2.2.2.1 The leverage-salience theory
Roose et al. (2007) analysed the effect of reminders and replacement questionnaires in the second step of a two step data collection for a survey amongst audience. They found that the positive leverage of reminders and replacement questionnaires decreases when the interest in the survey topic increases. Groves et al. (2000) analysed the different effects of incentives on survey response, depending on the degree of involvement in the community. Highly involved persons in the community will be more inclined to participate in a survey. Incentives also increase the chances of participation. However, when combining these correlates the leverage-salience theory stipulates that the positive effect of the incentive on survey participation will be smaller for persons with a high degree of community involvement. ■

In an earlier article, Groves et al. (1992) use theories from an area of social psychology that describes the compliance with requests to explain survey participation. They identify several components that influence survey participation: influences of socio-demographic and survey design features, the interaction between the interviewer and the sample person, tailoring and maintaining interaction. They use the compliance principles identified by Cialdini (1988), that serve as heuristic rules for compliance with a request within society. These principles can be used to form hypotheses about survey participation behaviour that are more specific than the social exchange theory from Blau (1964). The six compliance principles are:

- *Reciprocation.* People tend to respond with similar behaviour as being approached with. Thus, a sample person should be more willing to participate in a survey when the compliance is seen as the repayment of a gift or favour. This principle underlies the positive effect of incentives on survey participation. Advance letters and brochures also enhance compliance.

- *Consistency.* People have a strong tendency to be consistent within their attitudes, beliefs, words and deeds. If the interviewer can make a connection between the attitude of the respondent and survey participation, for instance the sharing of knowledge, this will enhance survey participation.

- *Social validation.* This principle implies that people like to validate their behaviour by comparing it to the behaviour of similar other persons. With respect to survey participation; persons will be more inclined to participate when they believe that similar persons do so too.

- *Authority.* A request that comes from (a representative of) a legitimate authority will be more likely to obtain high participation.

- *Scarcity.* Scarce opportunities are perceived as more valuable. Thus, suggesting that participating in a survey is an opportunity that is limited ('only one in every 300.000 persons is contacted' or 'we only have a couple of days left to do the interview') will stimulate people to comply with the request.

- *Liking.* People will be more inclined to comply with the request of a person that they like. In case of survey participation, this may also be the organisation the interviewer represents.

These principles can be used in the leverage-salience framework to make certain survey features more salient to sample elements. For instance, if a person is very sociable then the interviewer may choose to make the liking principle more salient. Tailoring and maintaining interaction can be best explained in the leverage-salience framework. Tailoring implies that the interviewer identifies which survey features are valued most by the sample person, and, consequently, tries to increase the saliency of these features to increase the response probability. Maintaining interaction gives the interviewer more clues which survey design features are important to the sample person. Furthermore, this appeals to the principle of reciprocation; the interviewer invests time into persuading the sample person to participate in the survey. This should be paid back.

Groves and Couper (1998) have tested the opportunity cost hypothesis and

the social isolation hypothesis. The opportunity costs hypothesis implies that the prospective respondent weighs the pros and the cons of participation, resulting in a decision one way or the other. This rational decision making relates to the voluntary view of Goyder and the leverage-salience theory of Groves et al. (2002). Costs of participation would be (Groves and Couper, 1998) 'the time required to complete the interview, the lost opportunity to perform other activities, the cognitive burden incurred in comprehending and answering the survey questions and the potential embarrassment of self-revelations that the questions require.' Benefits of participation could for instance be the satisfaction of contributing to a socially useful enterprise, the joy of having a conversation with the interviewer, or avoiding more unpleasant tasks, or the satisfaction of fulfilling a perceived civic duty.

The social isolation theory hypothesizes that persons who are alienated from society will be less inclined to participate with surveys that come from important institutions of that society, e.g. universities or government. Furthermore, they will have less feeling of civic duty, and consequently feel less obligated to participate in surveys. Besides lower participation in surveys, these groups are posited to have a lower social participation in general which for instance results in lower voting rates. Voogt (2004) shows that voting behaviour and survey participation are highly correlated. Persons that vote, are more prone to participate in surveys.

2.3 Correlates of contact and participation

2.3.1 Literature review

In sections 2.2.1 and 2.2.2 we discussed the general theories behind making contact and obtaining survey participation. These theories can be used to form hypotheses concerning the causes of either contact or participation. In this section, we summarise the main findings of the survey literature on the correlates related to contact and survey participation.

The main focus lies on face-to-face surveys of a general population of households. This overview is based to a large extent on the books by Goyder (1987) and Groves and Couper (1998). They have attempted to present a general overview of the correlates influencing contact and participation whereas most empirical literature only presents results from one particular study.

Following these books, a lot of authors have used these theories. The theories have been tested and extended by Durrant and Steele (2007). They analyse a

number of surveys merged into one dataset with a survey indicator. Stoop (2005) presents an overview of the factors that are hypothesized to influence contact and participation. Furthermore, she also presents empirical results testing these hypotheses, based on one survey. These two references are therefore also used as input for the overview.

These studies employ a multivariate model with a (binary) indicator for either contact or participation as the dependent variable, and socio-economic and demographic variables as the explanatory variables so that also relationships between the explanatory variables are accounted for.

In table 2.1 we show the common correlates of survey response. A distinction is made between correlates of contact, and correlates of participation once contacted. '+' indicates that a positive relation is found, '-' implies a negative relation and '+/-' states that in some studies a positive, and in other studies a negative relation was found. We discuss the correlates of contact and participation separately in section 2.3.2 and 2.3.3 respectively. The studies that we report on, use a number of surveys for their analysis, thereby excluding the topic of specific surveys and generalizing the relationships found to hold for surveys from a general population. The results may be confounding, however, due to different topics, country-specific influences or time of the survey.

There has been a lot of research into the correlates of nonresponse for particular surveys, for example Bethlehem and Kersten (1986). We chose however, to summarise findings that are less survey-specific. It is important to note that the correlates that we report on here are not exhaustive. The overview is based on the available data that the authors used in their research.

2.3.2 Contact

The reported correlates for contact are

- *the number of persons in the household*
- *the composition of the household*
- *the labour force status of the household members*
- *age*
- *house ownership*
- *urbanicity* and
- *Social Economic Status*, or SES.

Table 2.1: Correlates of survey response - contact and participation

Variable	*Contact*	*Participation*
Number of persons	+ more persons	
Household composition	+ young children	+ children
	+ elderly persons	+ elderly persons
	- singles	+ presence of carer
Labour force status	- working	+ working
		- head of hh is unemployed
Age	+ higher age	-/+ higher age
		+ younger age
House ownership	+ owner	
Urbanicity	- high urbanicity	- high urbanicity
Social Economic Status (SES)		
- Occupation	- students	+ high SES
- Housing costs		+ lower SES
		- higher SES
- Education		+ lower education
Number of households	- multiple	

The larger the number of persons, the higher the probability that at least one member of the household is present when the interviewer calls. This is found in the studies by Groves and Couper (1998), who included the number of persons in the model, and Goyder (1987) who found a positive correlation between large families and contact. A negative correlation for singles and contact is reported by Goyder (1987) and Durrant and Steele (2007).

The composition of the household, the labour force status of its members and the occupational SES are related to the daily occupation of the household members. This provides an indication of the at-home behaviour of the household.

Households with younger children or elderly persons have a higher probability of contact. Households with children tend to have someone at home to take care of the young ones, whereas elderly people are at home more often because of retirement. Groves and Couper (1998), Stoop (2005) and Durrant and Steele (2007) all find a positive correlation between contact and the presence of children. Furthermore, Goyder (1987), Groves and Couper (1998) and Durrant and Steele (2007) find a positive effect of the presence of elderly.

Households where all members are active in the work force are more difficult to contact. That is, during working hours. The daily occupation of household

members provides an indication of the at-home behaviour. Stoop (2005) and Durrant and Steele (2007) find a negative effect on contact if the sample person is employed. Goyder (1987) finds that students have a lower probability of contact. He relates this to the occupational SES of sample elements by noting that students have an irregular at-home behaviour.

The positive effect of age on contact is found in the studies by Stoop (2005) and Durrant and Steele (2007). Both studies show that older persons (aged 50 - 79) have a significantly higher probability of contact.

The positive effect of house ownership is found in the studies of Goyder (1987) and Durrant and Steele (2007). Goyder (1987) hypothesizes that 'owners are significantly more likely to be 'available' for surveys'.

The negative effect of urbanicity is reported by Groves and Couper (1998). They relate urbanicity to a lifestyle with a smaller at-home probability. The studies of Stoop (2005) and Durrant and Steele (2007) also incorporate an indicator for urbanicity; Stoop (2005) for large cities (Amsterdam, The Hague, Utrecht and Rotterdam) and Durrant and Steele (2007) for London. These two studies also find a strong negative effect of urbanicity on contact. Note that this effect is corrected for the effects of the other variables in the multivariate model.

2.3.3 Survey participation

For the explanation of survey participation, the results are less easy to interpret as there are a number of theories that can be used to motivate the relationships found. Furthermore, the subject of the survey will have an influence on participation, as well as other features of the fieldwork that are not included in the analysis (e.g. fieldwork strategy, mode of data collection). As shown in example 2.2.2.1, the leverage-salience framework acknowledges that features of the survey can have a different influence on survey participation for different subgroups of sample elements. We will come back to this in chapter 3. The reported correlates for survey participation are:

- *age*
- *the labour force status of the household members*
- *the Social Economic Status*
- *household composition* and
- *urbanicity*

In the following subsections, we discuss all correlates separately.

2.3.3.1 Age

The findings reported on survey participation with respect to age are not consistent. A higher age is found to be negatively correlated with survey participation in the study by Goyder (1987). He finds that 'older people resist surveys'. In the study of Groves and Couper (1998), they find that age has a curvilinear relationship with participation. Households with younger and older age composition have a higher participation rate than households in the middle range or with mixed ages. Groves and Couper (1998) hypothesize that 'the higher rates among young households reflects an interest in social participation because of greater social engagement in general. The higher rates among the elderly may reflect stronger norms of civic duty'. Durrant and Steele (2007) also find a positive effect of higher ages on participation, positing that 'older people might feel a stronger obligation to contribute to the good of society'. An effect of age on participation is not found in the study by Stoop (2005).

2.3.3.2 Labour force status

Labour force status, or work status, holds implications for both contact and participation. As we saw earlier, being active in the workforce is a good indication of the at-home behaviour and thus highly related to (non)contact. For survey participation it is stipulated that employed persons have less time to participate in surveys. In the study by Groves and Couper (1998), they test for work status as an indicator of the opportunity cost hypothesis. As a result from this hypothesis, employed persons would have less time to participate in a survey and consequently view the burden of the interview larger than someone with more time. Following this line of reasoning, employed persons will be less inclined to participate. However, they find no evidence for this hypothesis in their data. Durrant and Steele (2007) report a significant positive effect on the probability of participation if the head of the household is employed, or retired. If the head is self-employed or unemployed, the probability of participation decreases. They motivate this finding by the theory of social exchange, that hypothesizes that 'individuals receiving fewer services from government and those feeling disadvantaged may also feel less obligated to respond to a government request'.

2.3.3.3 Social Economic Status

Contrary to the finding of Durrant and Steele (2007) that lower educated SES groups have lower participation rates, Groves and Couper (1998) find higher participation rates for lower socio-economic groups (with respect to housing costs)

and lower educated groups. However, as they motivate, 'the social exchange theory can lead to two different hypothesis about differences in cooperation by different SES groups.' The second hypothesis would be that lower SES groups have the greatest indebtedness to government because they receive public assistance whereas higher SES groups have less need for government services and do not feel like they owe the government a repayment. Groves and Couper (1998) thus find evidence for this latter hypothesis, however the difference in findings between the two studies can also be contributed to a country effect. It remains unclear how the relationship between SES and survey participation should be interpreted.

2.3.3.4 Household composition

Household composition plays an important role in explaining survey participation. Groves and Couper (1998) use composition of the household as an indirect indication for social isolation. They find that single households have a lower participation rate and that households with children present have a higher participation rate. Single households are suspected to be less socially integrated, whereas households with children have a high level of social integration through school and contact with other children/parents. Durrant and Steele (2007) also find that children in the household imply a higher rate of participation. They find no evidence, however, for the effect of single households after having controlled for all the other factors. What they do find, is a positive effect on participation for households with a carer present. They relate this to the helping tendency (Groves et al. (1992), section 2.2.2) or feelings of civic duty. In none of the other studies this effect was investigated. Also, a positive effect of households with elder persons is found. This may be motivated by the opportunity cost hypothesis, because elderly persons are often retired and thus have more time available. Also feelings of civic duty can play a role, elderly persons may feel a stronger obligation to contribute to society. The social isolation theory would hypothesize that this group has a lower probability of participation, but this effect is not found by any of the studies.

2.3.3.5 Urbanicity

The last factor that is put forward by the literature, is the degree of urbanicity. In the studies of Stoop (2005) and Durrant and Steele (2007), this factor has been included in the models as an indicator for large cities and London respectively. For both studies, a significant effect of this indicator on participation was

found. The degree of urbanicity can be interpreted as an 'aggregate-level effect on overall levels of cooperation' (Groves and Couper, 1998). There are different measures for the concept of urbanicity, e.g. large cities versus small towns, rural versus urban, or a classification based on the population density. According to Groves and Couper (1998) 'the trend is clear: residents of inner-city areas of large metropolitan areas exhibit the lowest levels of cooperation, while those in rural areas have the highest.' When interpreted in terms of population density, urbanicity can be hypothesized to decrease participation due to 'crowding' (Groves and Couper, 1998). It is not so much the densely populated areas itself, but the excess of brief, impersonal social encounters with strangers which can lead to a social overload, which in itself can lead to a lower participation rate in surveys. Furthermore, urban life can be associated with higher crime rates and a larger degree of social disorganization.

2.4 Classifying sample elements

In the previous sections we have focussed on theories that explain contact and survey participation from the perspective of the sampled persons or households. In this section, we take a different approach and look at the sample elements from the survey researchers' perspective. This means that we attempt to classify sample elements based on their behaviour with respect to the survey request.

2.4.1 Classes of participants and continuum of resistance

There are two views on how to classify sample elements: the continuum of resistance theory, and the classes of participants theory. The continuum of resistance theory indexes respondents on a linear scale from easy cooperative to fully resistant. According to the classes of participants theory, sample elements are placed in different subclasses according to some criterion. For a description and a comparison of these two theories, see for example Lin and Schaeffer (1995).

The implications of these two views for the classification of nonrespondents are quite different. According to the continuum of resistance, the nonrespondents are all situated at the end of the continuum. With every additional contact attempt respondents become more resistant to survey participation. This is different in the classes of participants theory. According to this theory, respondents as well as nonrespondents can be divided into subclasses based on some criterion. This means that nonrespondents are classified in different classes too, as opposed to the continuum of resistance where nonrespondents are simply

placed at the end of the continuum.

Both theories face the difficulty of how to define the measure that distinguishes sample elements. Even though the concept of resistance is intuitively appealing, it is very difficult to define and to measure. In the continuum of resistance theory resistance has to be defined by some linear measure. The classes of participants theory uses a categorical measure to distinguish between different types of sample elements.

It is not unusual for the concept of resistance to be defined in terms of both contact and participation, for example by regarding accessibility (contact) and amenability (participation) of the sample elements as described in Stoop (2005). Respondents that participate already after the first contact attempt are both easy accessible and very amenable. With every other contact attempt, respondents become either less accessible or more resistant to survey participation. At the end of the fieldwork period, the final nonrespondents are either difficult to reach or persistent in their refusal. Accessibility and amenability can also be used as criteria for the classes of participants theory. In this view, sample elements are placed in different classes depending on their accessibility and amenability.

Hence, the difference between the two theories vanishes when resistance is measured by the same concepts that distinguish between the classes of participants. When more than one concept is used to measure resistance, the continuum of resistance becomes multi-dimensional. Likewise, when only one criterion is used to form classes of participants, the measure can be placed on a continuous scale. So, although the two views are often presented as different theories, in practice they lead to similar conclusions.

2.4.2 A multi-dimensional continuum based on response probabilities

In Chapter 1, we introduced the different response types and their corresponding probabilities in the response process. These probabilities can be used to define the dimensions or classes according to which sample elements are classified. We illustrate this by the most commonly used classification based on accessibility and amenability. In terms of the response process probabilities, we can interpret accessibility as the contact probability γ and amenability as the participation probability ϱ. Both γ and ϱ are defined on the $[0, 1]$-interval. If γ_i respectively ϱ_i is close to 0, this implies that the accessibility respectively the amenability of sample element i is low. Sample elements with a low accessibility are hard-to-reach. Likewise, if γ_i respectively ϱ_i is close to 1 the accessibility respectively

amenability of sample element i is high.

Now, sample elements can be classified according to their values of γ and ϱ. A classification is obtained by choosing a certain threshold t, $t \in (0,1)$, for both γ and ϱ to distinguish between different classes. It makes sense to vary t for γ and ϱ. After all, the refusal rate is usually higher than the non-contact rate. Therefore the same threshold t for both dimensions possibly leads to empty cells. Let $t_1 \in (0,1)$ denote the threshold for the contact probability γ, and $t_2 \in (0,1)$ denote the threshold for the participation probability ϱ. It is also possible to extend this classification to include other types of response, for example being processed and being able to participate. Based on these two probabilities and thresholds t_1 and t_2, we can identify four classes of participants along the two dimensions of resistance, see table 2.2. In the class of sample elements that are easy accessible and have a high amenability, there will be a lot of respondents. These respondents are referred to as *easy respondents*. In the class that consists of sample elements that are hard-to-reach and have a low amenability, there will be a lot of nonrespondents. The respondents in this class are usually referred to as *hard-to-reach, persistent refusals*. In the classes with either a low accessibility or a low amenability, there will be less difference in the number of respondents and nonrespondents. The respondents in the class of sample elements that are easy accessible and have a low amenability are usually referred to as *persistent refusals*. The respondents that are hard-to-reach but have a high amenability are referred to as *hard-to-reach respondents*.

The hard-to-reach, persistent refusals are often merged with the persistent refusals, which results in the three classes of respondents commonly used in survey literature (e.g. Ellis et al. (1970), Filion (1976), Lynn et al. (2002b), Stoop (2004)): hard to reach respondents, persistent refusals and easy respondents.

Both the contact probability and the participation probability are unknown characteristics of the sample elements. They can be estimated with some ap-

Table 2.2: *A possible classification of sample elements*

		Accessible	
		$t_1 \leq \gamma < 1$	$0 < \gamma \leq t_1$
Amenable	$t_2 \leq \varrho < 1$	easy accessible high amenability	hard-to-reach high amenability
	$0 < \varrho \leq t_2$	easy accessible low amenability	hard-to-reach low amenability

propriate model using the available auxiliary information, see Chapter 1. In Chapter 3 we model the different probabilities in the response process for the Dutch Labour Force Survey. The contact probability, or accessibility, can be modelled by characteristics of the sample element and for example the number of contact attempts needed for the first contact. It is however possible that, although a respondent is easy to reach, due to chance it takes many contact attempts before contact is actually made. Although the number of contacts is not a good measure it is often available. Besides, the field has yet to propose better measures of accessibility. The participation probability, or amenability, can be modelled by characteristics of the sample element and for example the interviewer effort after the first contact has been made. However, interviewer effort is difficult to measure. More research should be directed at identifying measures for accessibility and amenability.

2.5 Concluding remarks

It is clear from the theories presented in section 2.2.2 that survey participation behaviour is a complicated social phenomenon. As opposed to the theory on survey contact, there does not seem to be one true, accepted model that describes survey participation. This is also reflected in the overview on the correlates of contact and participation in section 2.3. The survey literature agrees to a large extent on correlates of contact. However, the findings reported on survey participation are not consistent. Some effects are reported by one study but not confirmed by others.

One possible explanation could be that the effects of correlates on response differ according to the type of nonresponse, and the composition of these types may be different over studies, either in surveys or countries. For instance, elderly persons have a lower non-contact rate, but have a higher probability to be ill. If the non-contact rate is high in one of the countries, this may imply that the elderly are over-represented and so there will also be more persons not able to participate due to illness. In another study/country, if the non-contact rate is low, this effect might not be so strongly present. In that case, the relationship between age and participation will be different in the two studies.

Another explanation is that we do not observe some correlates of response behaviour. Some proportion of the response decision is simply random and can never be explained. But some proportion of the response decision may be influenced by personal factors that we will never know nor observe. Furthermore, there may be factors that play a role that are out of the influence of the survey

organization, such as the survey climate (that differs by country). The topic of the survey or the way the survey is presented may have a strong impact on participation. Clearly this changes from survey to survey which makes it difficult to generalize over different surveys.

Another attempt to classify sample elements based on their response behaviour is described in section 2.4. We propose a classification of sample elements based on their probabilities in the response process. These probabilities are unknown, however, they can be estimated by some appropriate model. Besides characteristics of the sample elements, it seems profitable to also include in these models some measures from the fieldwork. For example, the number of contact attempts untill first contact as a measure of accessibility, and the interviewer effort to measure amenability. However, better measures are needed and more research should be directed at identifying these measures.

Finally, we note that the fact that we are not able to completely understand and explain survey participation behaviour is not necessarily troublesome but can also be interpreted positively. If survey response behaviour is completely random, nonresponse causes a decreasing sample size but introduces no additional bias in survey estimates. We will come back to this in Chapter 5.

Chapter 3

Analysis of Nonresponse in the Dutch Labour Force Survey

In this chapter we give an example of the information on nonrespondents that becomes available from linking the sample survey to external registers. We demonstrate how the linked data can be used to analyse the response behaviour in the Dutch Labour Force Survey (LFS).

3.1 Introduction

In this chapter, we first describe the Dutch Labour Force Survey (LFS). The Dutch LFS is used for illustrational purposes throughout this thesis. In section 3.2 some important aspects of the LFS are discussed. We provide a short review of the sampling design and the data collection structure. Furthermore, the objective of the LFS, a short history of the LFS at Statistics Netherlands and some historic response rates are presented.

We analyse the response behaviour in a selected sample from the 2005 Dutch LFS. This data is presented in section 3.3. In section 3.3.2 we describe the linked data that we used to analyse the response behaviour in the Dutch LFS. Section 3.4 presents the analysis of the nonresponse to the Dutch LFS. We distinguish between the different types of nonresponse. In addition, we show

how this analysis is different from a straightforward analysis of nonresponse that does not distinguish between these different types. In section 3.5 we make some concluding remarks.

3.2 Aspects of the Labour Force Survey

3.2.1 Sampling design

The Labour Force Survey is one of Statistics Netherlands' most important household surveys. The Labour Force Survey is a continuous sample survey amongst residents of the Netherlands aged 15 years and older, with the exception of persons living in institutions, asylums and old people's homes, i.e. the institutionalized population. The main objective of the survey is to provide statistics about the employment status of the Dutch population (see section 3.2.3).

The sample is selected by means of a stratified two-stage sample. Municipalities are selected in the first stage. In this stage, the number of addresses in each of the municipalities is determined. The municipalities are stratified based on a crossing of 40 regional areas and interviewer districts. This classification corresponds to level 3 in the European NUTS[1]. Proportional allocation is used to distribute the sample among the strata. Within strata, the municipalities are selected with probabilities proportional to their size. Furthermore, the number of months that a municipality is selected in the sample is also proportional to its size. The large municipalities are self-selecting; due to their large size they are always selected in the sample.

In the second stage, a systematic sample is drawn from each of the municipalities selected in stage one. For the large municipalities, the number of addresses is drawn proportional to the size of the population of the municipality. For the other municipalities, groups of 12 addresses (clusters) are selected from the municipalities. The clustering has been done to facilitate the fieldwork, and research has shown that it has little influence on the variance of the estimates (De Ree, 1989), i.e. there is no cluster effect.

Every month, a sample of approximately 10,000 addresses is selected from the population register, in Dutch: Gemeentelijke Basis Administratie or GBA. The sample is enriched with information from the Geographic Register, abbreviated GBR, or in Dutch: Geografisch Basisregister. The register contains all addresses in the Netherlands with an indication of their location and is composed by the Dutch postal services TNT Post. The only information that is retrieved from the

[1] Nomenclature of Territorial Statistical Units

GBR, are the number of delivery points per address. An address with more than one delivery point receives more entries in the sampling frame to ensure that the inclusion probabilities are correct. Addresses with institutions, asylums or old people's homes are removed from the sample as these addresses are no part of the target population. Addresses with only residents aged 65 and older are depleted (because most statistics based on the LFS concern persons aged 15 - 64). Furthermore, in July and August, the sample is halved due to interviewers' holidays. All this resulted in a total of 71,000 addresses that were approached for the Labour Force Survey in 2005[2].

3.2.2 Data collection

The LFS was first conducted by Statistics Netherlands in 1987. Before 1987, it was a Pen And Paper Interview (PAPI) carried out by employees of the municipalities under a different name. The first number of years, the LFS was a cross-sectional survey. Every month, 10,000 households were selected and interviewed by the means of CAPI. From October 1999, the LFS is performed as a rotating panel. Every month a new sample enters, and another sample leaves the panel, see figure 3.1.

The first round is a Computer-Assisted Personal Interview, or CAPI. Sample elements are visited at home by one of the face-to-face interviewers of Statistics Netherlands. The duration of the interview is approximately half an hour per household, depending on the number of persons in the household and their work situation.

Sometimes more than one household is found on one address and/or delivery point. In that case, up to four households per address are interviewed, and up to eight persons per household. Proxy interviewing is allowed by the *household core*, i.e. the head of the household and his/her partner. At the end of this interview, respondents are asked if they want to participate in four consecutive interviews over the telephone (CATI, Computer-Assisted Telephone Interviewing). The telephone interviews take place with time lags of three months so that the last telephone interview occurs 12 months after the face-to-face interview. The telephone interviews are considerably shorter in duration than the personal interviews, taking only a couple of minutes. The data from the first wave is fed to the interviewer, and the CATI- questionnaire is designed to survey changes in

[2]A detailed description (in Dutch) of the methods and definitions in the Labour Force Survey 2005 can be found at www.cbs.nl: Methoden en Definities Enqête Beroepsbevolking 2005

Figure 3.1: *The rotating panel structure of the LFS*

(work) situation. Also for the telephone interviews, proxy interviewing is allowed by a member of the household core.

3.2.3 Objective of the LFS

The LFS provides monthly, quarterly and yearly statistics on participation in the labour market, employment, unemployment, and working hours among the population aged 15 years and older. The most important statistical objective of the LFS is to divide the population of working age (15 years and above) into three mutually exclusive and exhaustive groups - employed persons, unemployed persons and inactive persons.

The European Union provides a classification of the labour force, see Eurostat (2003):

Employed persons Persons aged 15 year and over, who during the reference week performed work for pay, profit or family gain. Even when this is just one hour a week. Employed persons are also persons who were not at work but had a job or business from which they were temporarily absent because of, e.g., illness, holidays, industrial dispute and education and training.

Unemployed persons Persons aged 15-65 who were without work during the reference week, were currently available for work and were either actively seeking work in the past four weeks or had already found a job to start within the next three months.

Inactive persons Those persons who neither classified as employed nor as unemployed.

The labour force classification of Statistics Netherlands is slightly different, the difference being the number of hours that a person has to work in order to be classified as employed. For the EU this is at least 1, whereas Statistics Netherlands uses a minimum of 12 hours a week. Furthermore, there are some minor differences regarding the reference time of the survey, and the definition of job search behaviour.

3.2.4 Response rates

Over the years, there have been some changes in the organisation of the fieldwork that affect the interpretation of the response rate. These changes concern the labour status of the interviewers, the questionnaire and the mode of data collection. The changes complicate comparisons between the present and the past as changes in response rate may reflect changes in the fieldwork organisation and not in the underlying response behaviour.

Statistics Netherlands started carrying out the LFS in 1987. In 1992, the LFS questionnaire was subject to a re-design. In October 1999, the design of the LFS was changed from a cross-sectional survey, to a continuous sample survey with a rotating panel design (see figure 3.1). The interviewers became employed by Statistics Netherlands in 2002, before that they were employed as freelancers. That development paved the way for a new fieldwork strategy, that became effective in 2004 (see also the next section). Furthermore, the calculation of the response rate was changed in 2004. The response rates for the LFS are displayed in figure 3.2. The response rate fluctuates around 60% over the years. Since the implementation of the new fieldwork strategy in 2004, the response rate seems to increase slightly. It remains to be seen in the future whether this uprising trend will continue.

3.2.5 Fieldwork strategy

The new fieldwork strategy, introduced in 2004, optimizes the probability of finding sample elements at home. Accordingly, interviewers are instructed to

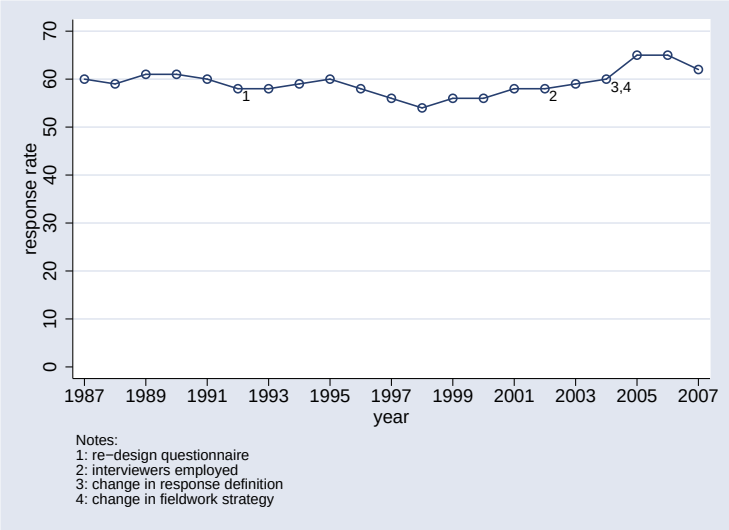

Figure 3.2: *The response rate for the LFS from 1987 - 2007*

visit all their addresses for the first time in the first half of the month. Contact attempts are preferably made at the end of the day, beginning of the evening or on Saturdays. When nobody is found at home, the subsequent attempt should be made at a different time, still preferably end day/beginning evening or Saturday.

When contact is made, interviewers are prompted to schedule an appointment rather than conducting the interview instantly. That is, unless the interviewer suspects that the sample element will not keep the appointment; then it is better to do the interview straight away.

Another ingredient of the fieldwork strategy are so-called visiting cards. These are cards that can be left behind by the interviewer when nobody is at home, or when an appointment has been made. There are four different visiting cards. The card 'first visit, no contact' only indicates that the interviewer visited the address. On the second card, 'second visit, no contact' the telephone number of the help desk of Statistics Netherlands is displayed. The third time the interviewer does not find anybody at home, the visiting card 'request for appointment' displays the telephone number of the interviewer and a proposed date for the interview. After the third attempt, the interviewers are allowed to

contact respondents by telephone but only to make an appointment. When an appointment has been scheduled, the visiting card 'appointment' is left behind with the date of the appointment and the interviewer's telephone number.

3.3 Data for analysis

3.3.1 ELFS sample

For the analysis in this thesis, we use data from the face-to-face Labour Force Survey in the months July - October 2005. During these months, samples of nonrespondents were re-approached. This re-approach of nonrespondents is the topic of Chapter 4. The availability of this additional data on the nonrespondents is the motivation for choosing this particular time period.

In 2005 the Netherlands was divided into thirteen interviewer districts for face-to-face surveys. Eleven districts participated in the re-approach study. Two districts were not able to assist because of a lack of available interviewers. In the participating districts two or three of the best interviewers were selected depending on the geographical size of the district. We needed at least two interviewers per district to be able to assign each address to a new interviewer. In five districts a number of municipalities was excluded so that travelling distances were acceptable. We exclude from the LFS the same districts and municipalities that have been excluded in the re-approach study. This may compromise the generality of the results from this analysis. However, since the selected municipalities cover the full range from rural to strongly urbanized areas and make up about 75% of the population we assume that the results are generalizable to the whole population.

In the remaining districts and municipalities, the total sample size for the LFS in July - October 2005, available for analysis, comprised 17,628 households. In the remainder of this thesis, we refer to this data as the experimental LFS sample, or *ELFS sample*, to distinguish the data from the regular LFS. In table 3.1 an overview of the composition of the ELFS sample is provided. We linked the ELFS sample to a number of external registers. In section 3.3.2 we describe the auxiliary data obtained by this linkage.

Table 3.1: *Summary of ELFS sample July - October 2005*

	Size	*Percentage*
Total sample	17,628	100%
Processed	17,274	98.0%
Contacted	16,190	91.8%
Able to participate	15,835	89.8%
Participated	11,012	62.5%

3.3.2 Linked data

3.3.2.1 Social Statistical Database

We linked the ELFS sample to the Social Statistical Database. Furthermore, we linked the data to the register from the Dutch telephone company KPN to determine whether the address has a listed land-line telephone.

The Social Statistical Database of Statistics Netherlands, denoted by its Dutch acronym SSB or Sociaal Statistisch Bestand, consists of administrative information on persons, households, jobs, benefits and pensions. It covers the entire Dutch population, including persons living abroad but working in the Netherlands or receiving a benefit or pension from a Dutch institution. The population register, denoted in Dutch by GBA or Gemeentelijke Basis Administratie, serves as the backbone for the SSB. The GBA contains mostly demographic information for all the persons that are or were registered in one of the municipalities since 1995. Changes in the demographic situation are also processed, when registered with the municipality.

Various sources with data on jobs are integrated for the SSB. These sources are among others insurance data, tax data, and data gathered from the Dutch Survey on Employment and Wages (EWL). Furthermore, there is information from the Centre for Work and Income (CWI), Tax Administration (in Dutch Belastingdienst) and the institute for employees insurances (in Dutch Uitkeringsinstituut WerknemersVerzekeringen (UWV)). A detailed description of the SSB can be found in Arts and Hoogteijling (2002).

3.3.2.2 Linked data used for analysis

The available variables are displayed in table 3.2. We have geographical, demographic and socio-economic information on different levels. The lowest level that we use in our analysis is the household level. All personal variables are therefore

aggregated to the household-level, based on information about the household core. Because of this aggregation, the variables ethnic group, gender and type of household have an additional category to indicate a mixture of the categories on the personal level. The next level comprises information at the postal code level[3]. The postal code was introduced in 1987, by the Dutch postal company PTT (nowadays TNT Post) to facilitate automatic sorting of the mail. The postal code consists of four digits and two letters. For example, the postal code of Statistics Netherlands in The Hague is 2490 HA. The first two digits indicate the region, the last two digits the neighbourhood. The two letters are a more specific indication of the neighbourhood and the street. Individual streets often have a different postal code for odd and even numbers. Together with the house number, the postal code is unique for every address. Finally, we also have information on the region, being mostly a distinction between different regions such as provinces and part of the country.

Table 3.2: *Linked data to the experimental Labour Force Survey*

Variable	*Categories*	*Source*
Household level		
Ethnic group	Native, Morrocan, Turkish, Suriname / Netherlands Antilles, other non-Western, other Western or mixed	GBA
Gender	All male, all female or mixed	GBA
Average age	15 - 34, 35 - 54, 55 +	GBA
Listed telephone	yes, no	KPN
Type of household	Single, unmarried no children, married no children, unmarried with children, married with children, single parent, other, more than one household	GBA
Number of persons	1,2, . . . ,5, 6 +	GBA
CWI registration	yes, no	CWI
Paid job	yes, no	Job register
Disability allowance	yes, no	UWV
Self employed	yes, no	Tax Office
Social allowance	yes, no	Municipalities
Unemployment allowance	yes, no	UWV
Postal code area level		
Degree of urbanization	very strong, strong, moderate, low, not	GBR
Percentage non-natives	very high, high, average, low, very low	GBA
Average house value	missing, 0 - 50, 50 - 75, . . . ,	Municipalities

Continued on next page

[3]www.postcode.nl (in Dutch)

Table 3.2: *Linked data to the experimental Labour Force Survey - continued*

Variable	*Categories*	*Source*
(in 1000's of Euros)	125 - 150, 150 - 200, ..., 250 - 300, 300 - 400, $>$ 400	
Regional level		
Part of country	north, east, west, south	GBR
Province and large cities	12 provinces plus Utrecht, The Hague, Rotterdam and Amsterdam	GBR

We discuss the linked information in more detail, providing definitions and a description of the source and measurement if possible.

Ethnic group

The first variable is ethnic group, which is based on ethnicity. Unfortunately, there is no straightforward definition of ethnicity. Statistics Canada[4] provides a definition of the concept, and notes that it is complex to develop appropriate concepts and constructs of ethnicity and also to collect unambiguous data, i.e. to measure ethnicity. Their definition reflects these difficulties:

> The concept of ethnicity is somewhat multidimensional as it includes aspects such as race, origin or ancestry, identity, language and religion. It may also include more subtle dimensions such as culture, the arts, customs and beliefs and even practices such as dress and food preparation. It is also dynamic and in a constant state of flux. It will change as a result of new immigration flows, blending and intermarriage, and new identities may be formed.

Statistics Canada identifies three ways of measuring ethnicity: *origin or ancestry, race* and *identity*.

Origin or ancestry Origin or ancestry attempts to determine the roots or ethnic background of a person. The concept of origin or ancestry is ambiguous because it does not specify a reference point in history. Furthermore, it may be difficult for a respondent to answer a question about origins.

Race Conceptually, race is not without difficulties in terms of measurement. The concept is based primarily upon genetically imparted facial features among which skin colour is a dominant, but not the sole, attribute.

[4] http://www.statcan.ca/english/concepts/definitions/ethnicity.htm

Identity Identity has a certain appeal because it attempts to measure how people perceive themselves rather than their ancestors. Nevertheless, it retains certain dimensions of not only origin but race as well. In addition, it may include aspects of citizenship. A typical question might be, With which ethnic group do you identify? Some respondents may associate the question with citizenship, others with origin, and still others might see it as involving both citizenship and origin. Others might see racial dimensions and report as black. Furthermore, in some contexts, ethnicity might be implied but the reference is actually to language.

Statistics Canada uses the race to measure ethnicity. Statistics Netherlands adheres to the concept of origin or ancestry. To overcome the difficulties in measurement and interpretation, the country of origin of the respondent and its parents are used as a proxy for origin or ancestry. This factual information can be retrieved from the population register. The variable ethnic group thus indicates with which country a person affiliates. A distinction is made between persons that are born abroad (first generation) and persons that are born in the Netherlands but of whom at least one of the parents is born abroad (second generation). When a person is born abroad (first generation) the ethnic group is determined by the country of birth. For persons from the second generation, the ethnic group is determined by the country of birth from the mother, unless the mother is born in the Netherlands. In that case, the country of origin from the father is regarded.

The classification distinguishes between natives and non-natives. Furthermore, it distinguishes between Western non-natives and non-Western non-natives. This last distinction is made because of large differences in socio-economic and socio-cultural position. Non-natives from Turkey, Africa, Latin-America and Asia with the exception of Indonesia and Japan are considered non-Western non-natives. Within the non-Western non-native group four groups are distinguished based on the policies of the Dutch government with regard to minorities. These are: Turkey, Morocco, Suriname and the Netherlands Antilles. Western non-natives are persons from all the European countries (except for Turkey), North-America, Oceania, Japan and Indonesia. Natives are persons from whom both parents are born in the Netherlands, regardless of their own country of birth.

Example 3.3.2.1 Some exotic cases of ethnicity
The definition of ethnicity and its adhering measurement at Statistics Netherlands leaves room for some exotic cases. For instance, the child of parents that are both born in the Netherlands, is by definition native. Even when the child

is adopted and was born in Colombia. Also, the Dutch Royal family to a large extent consists of non-native persons. Our crown prince Willem-Alexander is married to Maxima Zorreguieta who is born in Argentina. As a consequence, their three daughters are second generation non-Western non-natives. In its turn, prince Willem-Alexander is second generation Western non-native because his father, prince Claus, was born in Germany. ■

Gender and average age

Information on gender is recorded in the population register. The gender for the household core can be all male, all female or a mixture. The average age of the persons belonging to the household core is derived from the registered date of birth in the population register. If the core consists of two persons, their ages are averaged.

Type of household and number of persons

The information concerning the size and the composition of the household is also based on the population register. The population register contains information on the number of persons that live on an address (or delivery point when linked to the GBR), and whether persons are related. For approximately 7% of the population, the information from the GBA cannot be used to determine whether persons belong to the same household. When persons are married or registered as partners, the marital status and thus the household composition is easily derived. However, when persons are living together without being registered, it is unclear whether they belong to the same household. In that case, the household type is imputed based on the probability that they do not belong to the same household. These probabilities are based on data from the LFS.

Listed telephone

This information is obtained by linking the sample to the register from the Dutch telephone company KPN. This register contains information on all households with a listed land-line telephone. Households with an unlisted (secret) number and mobile-only households are not registered by the KPN.

Information from the SSB

A detailed description of the registers in the SSB can be found in Arts and Hoogteijling (2002). Here, we shortly describe what linked variables from the SSB we use in our analysis. These are: CWI registration, paid job, self-employed, social-, unemployment-, or disability allowance.

A registration at the CWI implies that at least one person of the household core is looking for a job, and/or applying for an unemployment- or social al-

lowance. If at least one member of the household core appears in the job register (in Dutch: banenbestand), this implies that there is a paid job in the household. The Tax Office has information on the self-employed from the declaration of income tax. From this register we can deduct whether at least one member of the household core was self-employed. The information on disability for employment comes from the institute for employees insurances UWV. The information on social allowances is obtained from the municipalities, they are responsible for the allocation of these allowances. Only one person per household can receive a social allowance. Information about unemployment allowances is also obtained from the UWV.

Degree of urbanization

This variable is defined on the postal code area level. The degree of urbanization is based on the classification of surrounding address density in five categories: Very strong urbanized: 2,500 addresses or more per square kilometre; strongly urbanized: 1,500 to 2,000 addresses per square kilometre; moderately urbanized: 1,000 to 1,500 addresses per square kilometre; low or hardly urbanized: 500 to 1,000 addresses per square kilometre; not urbanized: fewer than 500 addresses per square kilometre.

The surrounding address density is the average number of addresses within a one kilometre radius. The information is based on the Geographical Basic Register, GBR.

Percentage non-natives

This variable is also defined on the postal code area level. The percentage of persons with a non-native background is derived from the population register. On a postal code area, 20% or more non-native households is regarded as a very high percentage of non-natives, 15–20% as high, 10–15% as average, 5–10% as low and 0–5% as very low.

Average house value

Property value (in Dutch: waarde onroerende zaken (WOZ)) is defined as the value, evaluated periodically by municipalities, in the legal framework of the law on property values. The variable is obtained from the register from the Tax Office, which on its turn receives these values from the municipalities. The municipalities are obliged to value all real estate. The reference date for the house values in 2005 is January 1, 2003. The values for the property tax are available on a postal code area level. The values are categorized into 14 classes.

Part of country and province and large cities

Part of country is the regional grouping of provinces. The classification corresponds to level 1 in the European NUTS classification for regional statistics. NUTS level 2 concerns provinces. The Netherlands is divided in 12 provinces. Larger cities are put in a different category because they are known to have a different relationship with response behaviour and some socio-economic indicators.

3.3.3 Unit of analysis

For the analysis of nonresponse, it is important to clearly define the response indicator. The response indicator has a different meaning, depending on the unit of analysis. We distinguish between personal- and household surveys.

3.3.3.1 Personal survey

In a personal survey, a response implies that the sampled person participated. The survey answers concern the sample element and non-contact is defined at the level of the sample person. Nonresponse due to a longtime illness is an individual characteristic and thus also defined at the sample element level.

For the other types of nonresponse, i.e. not able due to language problems or a refusal, establishing the identity of the contacted person can cause difficulties. Sometimes it is difficult to assess whether the contacted person is indeed the sample element, for instance when language problems prevent the interviewer from obtaining this information, or in case of an eminent refusal where the refusing person does not want to answer any questions at all.

Usually the survey organisation provides a classification of different survey outcomes, so that it can be reconstructed what happened at each address. When analysing a personal survey, it is fairly safe to define the response indicator on the personal level; thereby assuming that also the refusal and the not able due to language problems are characteristics of the sample element.

3.3.3.2 Household survey

In a household survey, the response indicator is defined on the household level. A response implies that all eligible members of the selected household provided answers to the survey questions. Possibly, some eligible members do not participate. In that case, the response is partial. Usually, the minimum requirement for a partial response is that the members of the household core participate in

the survey. Otherwise the response is regarded as nonresponse too.

In case of a non-contact, no person in the household could be contacted. In that case the nonresponse can still be attributed to the household. For a nonresponse due to language problems, longtime illness or refusal it is required to make some assumptions. One person in the household can cause this type of nonresponse. However, this does not necessarily mean that the same reason holds for all eligible household members. If the person causing the nonresponse would have been asked for participation later in the interview process, when some other members of the household already answered the survey, maybe the same cause would have resulted in a partial response instead of a nonresponse.

Therefore, the definition of the response indicator is slightly different in a household survey, compared to a personal survey. A nonresponse due to a language problem, longtime illness or a refusal implies that this nonresponse cause holds for at least some member of the household. The household characteristic then becomes that at least one eligible person in the household could not participate, or refused participation. For this reason, partial response is usually regarded as nonresponse too.

3.4 Nonresponse analysis

3.4.1 Model selection and interpretation

In this section, we present results from the analysis of different types of response in the ELFS sample. The different processes are described separately, i.e. we do not allow for correlation between response types. The results are presented and discussed in section 3.4.2 to 3.4.5.

The sequential character of the response process is accounted for by analysing each process conditional on the preceding process(es). That implies that participation is analysed on the subsample of observations that were processed, contacted and able to participate; being able is conditional on being contacted and processed; and contact is conditional on being processed. Every analysis thus is based on a different number of observations. See table 3.3. The aim of the analysis is to construct logistic regression models with the auxiliary variables to estimate the probabilities ξ, θ, γ and ϱ. We refer to the estimates of the probabilities $\hat{\xi}, \hat{\theta}, \hat{\gamma}$ and $\hat{\varrho}$ as propensities, see Chapter 1. The logistic models are estimated separately, when interpreted together they are referred to as a nested logit model. The description of the nested logit model is beyond the scope of this chapter. We present the nested logit model and other models to analyse

Table 3.3: *Summary of different processes for ELFS sample July - October 2005*

Process	*Probability*	*Available observations*
Being processed	ξ	17,628
Contact	γ	17,274
Being able	θ	16,190
Participation	ϱ	15,835

nonresponse in detail in Chapter 8.

We construct models for the response types on the household-level, see section 3.3.3. The total number of households in the ELFS sample is $n = 17,628$. All analyses are run in Stata. The analyses use the available linked variables, see table 3.2. First, these variables are evaluated bivariately. Cramér's V statistic is used for that purpose, see e.g. Agresti (2002). Cramér's V is a χ^2-based measure of association that accounts for the different degrees of freedom of the variables (different number of categories) and the dimension of the dataset, i.e.

$$V = \sqrt{\frac{\chi^2}{N \times min(r-1, c-1)}} \tag{3.1}$$

where r is the number of categories from the linked variable (rows), and c the number of categories of the response type (columns). Since the response type is a binary indicator, Cramér's V reduces to $\sqrt{\chi^2/N}$. $V = 0$ in case the selected variable and the propensity indicator are completely independent. $V = 1$ indicates the selected variable and the propensity indicator are completely dependent. Based on Cramér's V, the strongest candidate for the model is selected. Stepwise, the logistic model with one additional variable is evaluated. The final model is found if none of the additional variables has a significant relation with the dependent variable. The relative significance, i.e. adjusted for the effect of the other variables in the model, is evaluated by the p-value of a Wald test for joint significance. The null hypothesis of the Wald test is that all coefficients corresponding to the categories of a variable are equal to zero, against the alternative hypothesis that at least one of the coefficients is not equal to zero. If $p > 0.05$, the variable is not added to the model.

We report the estimated coefficients β with the corresponding standard errors (se) of the final model, for all categories except the reference category. Also, the corresponding χ^2-value for the Wald-test of joint significance of all categories of a variable is presented. An asterisk reflects that a category is significant on

a 1%-level. The sign of the coefficient β determines whether the propensity is increasing or decreasing with respect to the reference category. By exponentiating β, i.e. e^{β}, the odds ratio is obtained. The odds ratio is used to compare whether the probability of an event is the same for two groups. For a variable with two categories, for instance gender with woman as reference category, and the indicator C for being contacted, e^{β} is the odds ratio of being contacted for men compared to women. This ratio varies with the value of the other auxiliary variables in the model. This implies that its value can only be interpreted by assuming that the household belongs to the reference group of the other variables, see Agresti (2002).

For the general model fit the Nagelkerke pseudo R^2, the total χ^2 and the degrees of freedom (df) of the model are reported. The Nagelkerke pseudo R^2 (Nagelkerke, 1991) is expressed as

$$\text{pseudo } R^2 = \frac{1 - \{L(0)/L(\hat{\boldsymbol{\beta}})\}^{2/n}}{1 - L(0)^{2/n}} \tag{3.2}$$

where $L(0)$ denotes the likelihood of the null model, and $L(\hat{\boldsymbol{\beta}})$ the likelihood of the fitted model. The measure can be interpreted as the proportion of explained variation by the model, like the R^2 in classical regression analysis. Furthermore, the pseudo R^2 takes on values between 0 and 1.

3.4.2 The process propensity

Unprocessed cases are selected sample elements that have not been approached in the field, see Chapter 1. U denotes whether or not a household is processed. For the experimental LFS July - October 2005, the percentage of unprocessed cases was only 2%, i.e. $U = 0$ for 354 households, compared to $U = 1$ for $17,274$ households.

The results for the analysis of $\hat{\xi}$ are given in table 3.4 to 3.6. From the bivariate analysis (table 3.4) it becomes clear that the relation between the auxiliary variables and being processed is not very strong. None of the unprocessed cases has a listed telephone, therefore Cramér's V for listed telephone could not be calculated. Only for the regional variables province and largest cities (V = 0.2344) and part of country (V = 0.1078) Cramér's V shows a dependency with U. Including both variables in the multivariate analysis introduced collinearity in the model. Since part of country has the lowest value for Cramér's V it is not included in the model.

This result is confirmed in the multivariate analysis (table 3.5). There are two significant variables: province and largest cities and degree of urbanization.

The largest contribution comes from the province and largest cities. None of the other variables shows a significant relation to being processed, in addition to province and degree of urbanization. The model fit is reasonably good; the pseudo R^2 equals 13.9% which can be considered as a reasonable fit for cross-sectional data.

In two of the provinces, being Drenthe and Limburg, there were no unprocessed cases. These provinces are therefore not included in the model which reduces the total number of observations to $15,866$. The results for the final model are given in table 3.6. Especially in Groningen and Amsterdam, the propensity for households to be processed is low. In general, the less urbanized areas show a higher propensity of being processed. The reference category for degree of urbanization is a high degree. The coefficients increase with a lower degree of urbanization. The only category that is not significant, is low degree of urbanization. The reference category for province and largest cities is the province of Groningen, where the percentage of unprocessed cases is highest (22.2%). Therefore, the coefficients of the other categories compared to Groningen are relatively high. Especially Utrecht (excluded the city of Utrecht), Noord-Brabant and The Hague have a high propensity for being processed.

Conclusively, the variables that have a significant relation to $\hat{\xi}$ are province and largest cities and degree of urbanization. Based on this analysis we can assume that the process propensity is unrelated to any of the household characteristics. Whether or not a household is being processed depends mostly on its location, and to a lesser extent on the degree of urbanization of the area.

Table 3.4: *Bivariate analysis of processed versus unprocessed cases*

Variable	*Cramér's V*	*Variable*	*Cramér's V*
Province and largest cities	0.2344	Average age	0.0142
Part of country	0.1078	CWI registration	0.0124
Degree of urbanization	0.0450	Gender	0.0111
Average house value	0.0447	Self employed	0.0098
Ethnic group	0.0271	Paid job	0.0086
Type of household	0.0243	Unemployment allowance	0.0072
Percentage non-natives	0.0227	Disability allowance	0.0034
Number of persons	0.0217	Listed telephone	-
Social allowance	0.0195		

Table 3.5: *Multivariate analysis of processed versus unprocessed cases*

Variable	*Wald* χ^2	
Province and largest cities	462.06	489.37
Degree of urbanization		60.09
pseudo R^2	0.1213	0.1394
χ^2	411.56	472.86
df	11	15

Table 3.6: *Logistic model for the propensity of being processed* $\hat{\xi}$

Variable	*Category*	β
Province and largest cities	Friesland	3.570*
(Reference Groningen)	Overijssel	3.203*
	Gelderland	2.839*
	Utrecht (except Utrecht city)	4.380*
	Noord-Holland (except Amsterdam)	3.093*
	Zuid-Holland (except Den Haag and Rotterdam)	2.744*
	Noord-Brabant	4.259*
	Amsterdam	1.251*
	Rotterdam	3.299*
	Den Haag	4.949*
	Utrecht	2.250*
Degree of urbanization	high	-1.086*
(Reference very high)	average	-0.985*
	low	-0.510*
	very low	-0.097
Constant		1.668*
pseudo R^2	0.1394	
χ^2	472.86	
df	15	

3.4.3 The contact propensity

Contact is defined as: either a face-to-face meeting, intercom talk (CAPI) or a telephone conversation (CATI) with the sample element, see also Chapter 1. The total number of observations to analyse contact $C = 1$ versus no contact $C = 0$ equals $n = 17,274$. Of these households, $16,190$ were contacted.

From the bivariate analysis in table 3.7 it is clear that there are relations between a number of variables and the contact propensity $\hat{\gamma}$. The highest value for Cramér's V is obtained for gender of the household core ($V = 0.1965$), closely followed by type of household ($V = 0.1872$) and number of persons in the household ($V = 0.1785$). These results are confirmed in the multivariate analysis, given in table 3.8. Besides the variables mentioned above, a number of other variables also show a significant relationship to the contact propensity. These are province and largest cities, listed telephone, degree of urbanization and age. In the final model, the largest contribution to the model is obtained from the variable province and largest cities, and listed telephone. It is remarkable how much the high value for the Wald statistic from the variable gender decreases when the next variable, type of household, is added to the model. It turns out that the negative effect of only men in the household is confounded by the fact that most single households are a man alone. The final model's fit is reasonably good, with a pseudo R^2 of 13.8%.

When we look at the results from the logistic regression in table 3.9, we see that single households are very difficult to contact. Households with children present (either married, unmarried or single parent) show a significantly higher propensity of contact. In the same line of reasoning it follows that the higher the number of persons in the household, the higher the contact propensity. Only for the households that consist of 5 or more persons the results are not significant. With respect to the province and largest cities, the only significant relationship is found for Amsterdam, where the contact propensity is lower than in the reference category Groningen. As expected (Schouten, 2004), having a listed land-line telephone is a good indication of a high contact propensity. Even in the multivariate analysis, when adjusted for the effects of other variables, having a listed telephone implies a significantly higher contact propensity. Amongst the unlisted telephone households, the percentage of contact is 89.6%, whereas for listed telephone households this percentage is 96.7%.

The reference category for degree of urbanization is highly urbanized. The contact rate in areas with a high degree of urbanization is 88.6%, compared to 96.4% for areas with a low degree. The coefficients for the contact model are higher for the categories with a lower degree of urbanization. Thus, the lower the degree of urbanization, the higher the contact propensity $\hat{\gamma}$. The only category that is not significant, is very low degree of urbanization. And finally, the age variable is included in the multivariate model. The reference category is the age group 15 to 35. The contact propensity increases with increasing age. For the reference group, the contact rate is 89.2%. For the age group 35 to 55, this is 94.4% and for the group aged 55+ the contact rate is even higher, 96.0%.

Our analysis to a large extent confirms the presented correlates of non-contact in Chapter 2. With respect to the household composition, we find a significant positive effect of the presence of children. The household types with children present, either married, unmarried or single-parent households have a higher contact propensity. This effect was also reported in the studies in Chapter 2. The larger the number of persons in the household, the higher the contact propensity. As noted in Chapter 2, the probability that at least one person is at home when the interviewer calls increases with the number of persons in the household. Furthermore, we also find that the contact propensity increases with age. The negative effect or urbanicity is also replicated in our analysis. The strongest relationship is the negative effect for households that are located in Amsterdam. This effect remains even when adjusting for degree of urbanization and other variables. In Amsterdam the propensity to be processed is also very low, but the analysis of the contact propensity is performed on the households that have been processed, so there should be no confounding between these two types of nonresponse. A similar strong effect for large cities is reported in Chapter 2.

Households with a listed land-line telephone have a higher contact propensity. This correlate was not investigated in any of the studies that we reported on in Chapter 2. In our analysis, we find a positive effect of a listed telephone on the contact propensity. Notice that the data collection for the LFS is face-to-face, and not by telephone. Households that have a listed telephone are easier to reach, which implies that having a listed telephone is strongly related to the at-home behaviour. Households with a listed telephone are more often at home than households without a land-line telephone, an unlisted land-line telephone or mobile only households. A possible hypothesis is that mobile only households display behaviour that places them out of their homes more often.

Table 3.7: *Bivariate analysis of contacted versus not contacted cases*

Variable	*Cramér's V*	*Variable*	*Cramér's V*
Gender	0.1965	Ethnic group	0.0873
Type of household	0.1872	Part of country	0.0659
Number of persons	0.1785	Social allowance	0.0365
Province and largest cities	0.1560	Self employed	0.0310
Listed telephone	0.1451	Paid job	0.0174
Degree of urbanization	0.1288	CWI registration	0.0169
Average house value	0.1036	Disability allowance	0.0158
Average age	0.1026	Unemployment allowance	0.0046
Percentage non-natives	0.0924		

Table 3.8: *Multivariate analysis of contacted versus not contacted cases*

Variable	*Wald* χ^2						
Gender	562.90	86.36	71.26	72.17	63.47	62.45	55.86
Type of household		81.94	56.96	44.13	36.09	31.41	26.91
Number of persons			14.47	15.16	13.70	13.96	12.46
Province and largest cities				232.67	189.60	132.76	136.19
Listed telephone					148.76	141.66	111.68
Degree of urbanization						25.20	21.73
Average age							47.18
pseudo R^2	0.0705	0.0805	0.0823	0.1096	0.1290	0.1322	0.1381
χ^2	570.71	651.95	666.47	887.63	1045.39	1070.52	1118.51
df	2	9	14	27	28	32	34

Table 3.9: *Logistic model for the contact propensity* $\hat{\gamma}$

Variable	*Category*	β
Gender	All women	0.624*
(Reference all men)	Mixed	0.629*
Type of household	Unmarried no/c	0.191
(Reference single)	Married no/c	0.333
	Unmarried with/c	0.692*
	Married with/c	0.758*
	Single parent	0.361*
	Other	0.125
	Mixed	0.357
Number of persons	2	0.441*
(Reference 1)	3	0.414*
	4	0.633*
	5	0.430
	6+	-0.156
Province and largest cities	Friesland	0.232
(Reference Groningen)	Drenthe	-0.357
	Overijssel	-0.289
	Gelderland	0.124
	Utrecht (except Utrecht city)	0.151
	Noord-Holland (except Amsterdam)	-0.379
	Zuid-Holland (except Den Haag and Rotterdam)	0.258
	Noord-Brabant	0.343
	Amsterdam	-0.820*

Continued on next page

Table 3.9: *Logistic model for the contact propensity $\hat{\gamma}$ - continued*

Variable	*Category*	*β*
	Rotterdam	-0.210
	Den Haag	0.441
	Utrecht	0.412
Listed telephone	yes	0.781*
(Reference no)		
Degree of urbanization	high	0.318*
(Reference very high)	average	0.494*
	low	0.441*
	very low	0.215
Average age	>= 35 and < 55	0.325*
(Reference >= 15 and < 35)	>= 55	0.686*
Constant		0.840*
pseudo R^2	0.1381	
χ^2	1118.51	
df	34	

3.4.4 The propensity for being able to participate

Conditional on being processed and contacted, we can now proceed to analyse what households are able to participate ($L = 1$ for able versus $L = 0$ for not able). In our dataset, the only cause that we consider for not being able to participate is a language problem. Another cause of not being able could be having a longtime illness, but this did not occur in our dataset. The total number of households available for the analysis equals $n = 16,190$. Of these households, 355 households could not participate due to language problems ($L = 0$). Hence, $15,835$ of the households that were contacted, were also able to participate.

In the bivariate analysis, given in table 3.10, we find a very large Cramér's V for ethnic group, $V = 0.3459$. It is not very surprising that language problems are much related to a household's ethnic group. Other variables that have a high value for Cramér's V are: Province and largest cities, percentage non-natives, degree of urbanization, average house value and listed telephone.

The final multivariate model, given in table 3.11, consists of ethnic group, province and largest cities and percentage of non-natives. The model fit is good, with a high pseudo R^2 of 30.3% for the final model. As the bivariate analysis already pointed out, the largest contribution to the model is obtained by the ethnic group. The coefficients for the logistic model are displayed in table 3.12. The province of Drenthe is excluded from the analysis because no language problems occurred there, reducing the total number of observations to $n = 16,108$. The

strongest variable in the model is ethnic group. With respect to the reference category native, we see large negative coefficients for the other categories, which reflects a larger probability of language problems in the household. The second variable in the model is province and largest cities. None of the categories is significant in the model. For the percentage of non-natives, only the category with a very high percentage is not significant. The other categories show a low propensity for being able to participate compared to areas with a low percentage of non-natives. We thus observe an increasing propensity to experience language problems in areas with an increasing percentage of non-natives.

It can be concluded that we are well capable of understanding nonresponse due to language problems. The model fit is remarkably high with a pseudo R^2 of 30.3%. Furthermore, the strongest explanatory variable is the ethnic group which is based on the country of origin of the household core and therefore highly correlated to the language spoken by the household.

Table 3.10: *Bivariate analysis of able versus not able cases*

Variable	*Cramér's V*	*Variable*	*Cramér's V*
Ethnic group	0.3459	Type of household	0.0604
Province and largest cities	0.1311	Number of persons	0.0571
Percentage non-natives	0.1248	Average age	0.0538
Degree of urbanization	0.1118	Paid job	0.0400
Social allowance	0.1058	Gender	0.0395
Average house value	0.1024	Self employed	0.0157
Listed telephone	0.1000	Unemployment allowance	0.0058
CWI registration	0.0641	Disability allowance	0.0056
Part of country	0.0605		

Table 3.11: *Multivariate analysis of able versus not able cases*

Variable		*Wald* χ^2	
Ethnic group	717.98	635.23	529.50
Province and largest cities		58.10	51.12
Percentage non-natives			13.30
pseudo R^2	0.2829	0.2994	0.3033
χ^2	965.86	1021.09	1034.44
df	6	18	22

Table 3.12: *Logistic model for the propensity to be able* $\hat{\theta}$

Variable	*Category*	β
Ethnic group	Morrocan	-3.915*
(Reference native)	Turkish	-3.881*
	Suriname / Netherlands Antilles	-2.134*
	other non-Western	-3.175*
	other Western	-4.347*
	mixed	-1.757*
Province and largest cities	Friesland	0.761
(Reference Groningen)	Overijssel	0.391
	Gelderland	1.070
	Utrecht (except Utrecht city)	-0.174
	Noord-Holland (except Amsterdam)	0.337
	Zuid-Holland (except Den Haag and Rotterdam)	0.299
	Noord-Brabant	0.248
	Limburg	0.556
	Amsterdam	-0.588
	Rotterdam	0.467
	Den Haag	-0.347
	Utrecht	0.402
Percentage of non-natives	low	-0.439*
(Reference very low)	average	-0.383*
	high	-0.675*
	very high	0.140
Constant		5.576*
pseudo R^2	0.3033	
χ^2	1034.44	
df	22	

3.4.5 The participation propensity

The last step in the response process concerns the decision to refuse or participate, i.e. the participation propensity $\hat{\varrho}$. For participating households, $P = 1$ and $P = 0$ for households that refuse participation, conditional on being processed, contacted and able to participate. The total number of households is $n = 15,835$. Of those households 69.5% participated, and 30.5% refused participation.

In table 3.13 Cramér's V values are displayed. In general these values are very low, which implies no or negligible relationships between the participation variable P and the auxiliary variables. The multivariate analysis confirms this result, the best model fit is obtained by including the variables province and

largest cities, paid job and average house value, but this is still not a good fit with a pseudo R^2 of 0.86%. Details of the final logistic model are given in table 3.15.

The largest contribution to the model is obtained from the variable province and largest cities, followed by paid job and average house value. Again, the only category of province and largest cities that shows a significant relationship in the multivariate model is Amsterdam. In the reference category Groningen, 68.6% of the households participated whereas in Amsterdam this is only 58.0%. The other categories of the variable show no significant difference with the participation propensity in Groningen. The reference category for the average house value is the group of houses for which no value has been reported. This is usually the case if the houses are recently build. With respect to this category, the only significant findings based on the participation model are for the average house value of 75,000 - 100,000 and 100,000 - 125,000, each of which shows a lower participation rate (66.3% resp. 67.4%) than the reference group (71.9%). The average participation rate is 69.5%. The most striking and perhaps disturbing finding for the LFS is the relationship between paid job and participation. The reference category for paid job is not having a paid job. Households that have at least one paid job amongst the members of the household core have a higher participation propensity. The participation rate for having a paid job is 71.1%, compared to 65.2% for not having a paid job.

In the final model there are relationships between participation and province and largest cities, having a paid job and average house value. The significant effect of the variable province and largest cities is caused by the lower participation propensity in Amsterdam. We also reported a lower process- and contact propensity for Amsterdam. Above that, it seems that households from Amsterdam are less prone to participate in the LFS. The same result is found by Stoop (2005) for large cities, i.e. also including The Hague, Rotterdam and Utrecht, and by Durrant and Steele (2007) for London. We would be inclined to adhere the explanation of crowding, as suggested by Groves and Couper (1998), that leads to a social overload and consequently a lower participation propensity. It appears that this is especially the case in Amsterdam, rather than strongly urbanized areas in general since the degree of urbanization is not included in the final model.

If one of the members of the household core has a paid job, the participation propensity is higher than for households with no paid jobs in the household core. This finding can be motivated by the social isolation hypothesis, i.e. persons that are not very involved in society will be less inclined to participate with surveys that come from an institution of that society. They will also tend to experience

less feelings of civic duty, see Chapter 2. This result was also reported by Durrant and Steele (2007). Another explanation is that the topic of the survey caused a refusal. It may be that persons that do not have a paid job, think that the survey does not apply to them. Although the interviewers are trained to identify this behaviour and persuade persons to participate in spite of little topic interest, the result from this analysis indicates that this situations still occurs.

The final relation is found for the average house value. Two groups have a significant lower participation propensity, namely houses with an average value between 75,000 and 100,000 euro and between 100,000 and 125,000 euro. For Dutch standards, these average house values are fairly low. They have a lower participation propensity than the group of households for which the average house value is not reported. This relationship can be explained by the social exchange theory. If we regard the average house value as an indicator for the social economic status (SES), it appears that a lower SES leads to a lower participation propensity. Groves and Couper (1998) discuss one possible hypothesis based on the social exchange theory, that persons with a lower SES may feel disadvantaged in relationships with persons that are better off than themselves. If they view the interviewer as someone that is better off, making a request and offering little in return, this could motivate the lower participation rate amongst lower SES households. In Chapter 2 we discussed another hypothesis for participation and SES, leading to the opposite effect. Groves and Couper (1998) find support for this hypothesis. Durrant and Steele (2007) report findings that support the first hypothesis, leading to a lower participation propensity in lower SES groups. We find supporting evidence for the first hypothesis here as well.

We should note that the relationships between the participation propensity and the linked variables are not very profound. The model fit is poor, with a pseudo R^2 of only 0.86%. It is questionable how much attention should be paid to the relationships in the final model. It seems that the final participation decision does not depend strongly on any of the variables that we used in our analysis.

Table 3.13: *Bivariate analysis of participation versus refusal*

Variable	*Cramér's V*	*Variable*	*Cramér's V*
Province and largest cities	0.0737	Percentage non-natives	0.0247
Paid job	0.0557	Gender	0.0224
Average house value	0.0506	Number of persons	0.0201
Degree of urbanization	0.0493	Self employed	0.0000
Listed telephone	0.0489	Social allowance	0.0000
Part of country	0.0472	Unemployment allowance	0.0000

Continued on next page

Table 3.13: *Bivariate analysis of participation versus refusal - continued*

Variable	*Cramér's V*	*Variable*	*Cramér's V*
Type of household	0.0468	CWI registration	0.0000
Average age	0.0432	Disability allowance	0.0000
Ethnic group	0.0391		

Table 3.14: *Multivariate analysis of participation versus refusal*

Variable	*Wald* χ^2		
Province and largest cities	84.96	83.45	84.02
Paid job		47.61	42.10
Average house value			35.42
pseudo R^2	0.0043	0.0067	0.0086
χ^2	84.22	131.26	166.79
df	13	14	24

Table 3.15: *Logistic model for the participation propensity* $\hat{\varrho}$

Variable	*Category*	β
Province and largest cities	Friesland	0.140
(Reference Groningen)	Drenthe	-0.014
	Overijssel	0.146
	Gelderland	0.104
	Utrecht (except Utrecht city)	-0.010
	Noord-Holland (except Amsterdam)	-0.116
	Zuid-Holland (except Den Haag and Rotterdam)	-0.114
	Noord-Brabant	0.043
	Limburg	0.215
	Amsterdam	-0.470*
	Rotterdam	-0.095
	Den Haag	0.071
	Utrecht	0.285
Paid job	Yes	0.254*
(Reference No)		
Average house value[5]	0 - 50	0.027
(Reference missing)	50 - 75	-0.149

Continued on next page

[5] × 1,000 euro

Table 3.15: *Logistic model for the participation propensity $\hat{\varrho}$ - continued*

Variable	*Category*	β
	75 - 100	-0.238*
	100 - 125	-0.221*
	125 - 150	-0.118
	150 - 200	0.008
	200 - 250	0.030
	250 - 300	0.022
	300 - 400	-0.095
	400 and more	0.070
Constant		0.752*
pseudo R^2	0.0086	
χ^2	166.79	
df	24	

3.4.6 Contrast with more simplified response processes

In the previous sections we analysed four different types of response; being processed, contact, being able and participation. This response process with four types of response is referred to as the elaborate response process. It is interesting to contrast the analysis of the elaborate response process with that of more simplified response processes. In this section, we therefore analyse two simplified response processes. The most common distinction between response types made in the survey literature (see also Chapter 2), is that between contact and participation. We refer to this model as the restricted response process. The most simplified response process, however, is that where the only distinction made is between response and nonresponse.

3.4.6.1 The restricted response process

Figure 3.3 displays the restricted response process. Compared to the analysis of the elaborate response process, we now merge the response types being processed and contact, as well as being able and participation. Hence all the observations are being contacted and the observations that are not processed are regarded as non-contacts. The total number of observations is $17,628$. The number of non-contacts is $1,438$. These observations are censored when regarding the participation process. For participation, the total number of observations reduces to $16,190$. All sample elements that are not able to participate due to language problems are now regarded as refusals.

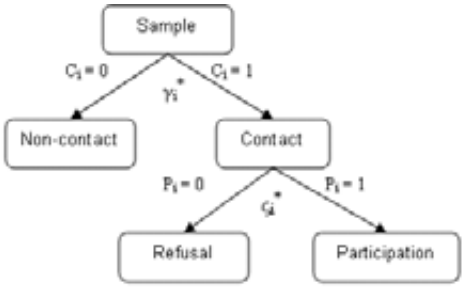

Figure 3.3: *The restricted response process distinguished by contact and participation*

Tables 3.16 and 3.17 display the final logistic regression models for contact and participation in the restricted response process.

Table 3.16: *Logistic model for the contact propensity* $\hat{\gamma}$

Variable	*Category*	β
Gender	All women	0.545*
(Reference all men)	Mixed	0.633*
Type of household	Unmarried no/c	0.339
(Reference single)	Married no/c	0.427*
	Unmarried with/c	0.734*
	Married with/c	0.679*
	Single parent	0.506*
	Other	0.321
	Mixed	0.480*
Province + large cities	Friesland	1.682*
(Reference Groningen)	Drenthe	1.365*
	Overijssel	1.178*
	Gelderland	1.402*
	Utrecht	1.730*
	(except Utrecht city)	
	Noord-Holland	1.066*
	(except Amsterdam)	
	Zuid-Holland	1.403*
	(except Den Haag, Rotterdam)	
	Noord-Brabant	1.879*
	Limburg	2.035*
	Amsterdam	0.293
	Rotterdam	1.168*
	Den Haag	1.825*

Continued on next page

Table 3.16: *Logistic model for the contact propensity $\hat{\gamma}$ - continued*

Variable	*Category*	β
	Utrecht	1.434*
Listed telephone	yes	1.281*
(Reference no)		
Average age	>= 35 and < 55	0.183*
(Reference >= 15; < 35)	>= 55	0.398*
Constant		-0.358*
pseudo R^2	0.1466	
χ^2	1460.57	
df	25	

Table 3.17: *Logistic model for the participation propensity $\hat{\varrho}$*

Variable	*Category*	β
Province + large cities	Friesland	0.169
(Reference Groningen)	Drenthe	0.030
	Overijssel	0.151
	Gelderland	0.125
	Utrecht (except Utrecht city)	-0.045
	Noord-Holland (except Amsterdam)	-0.117
	Zuid-Holland (except Den Haag, Rotterdam)	-0.116
	Noord-Brabant	0.036
	Limburg	0.225
	Amsterdam	-0.574*
	Rotterdam	-0.094
	Den Haag	-0.062
	Utrecht	0.273
Paid job	Yes	0.249*
(Reference No)		
Average house value[6]	0 - 50	-0.036
(Reference missing)	50 - 75	-0.187
	75 - 100	-0.219*
	100 - 125	-0.213*
	125 - 150	-0.101
	150 - 200	0.045
	200 - 250	0.058
	250 - 300	0.079
	300 - 400	-0.066
	400 and more	0.074
Ethnic group	Morrocan	-0.678*

Continued on next page

[6] × 1,000 euro

Table 3.17: *Logistic model for the participation propensity $\hat{\varrho}$ - continued*

Variable	*Category*	β
(Reference native)	Turkish	-0.404*
	Suriname, NL Antilles	0.242*
	other non-Western	-0.275*
	other Western	-0.311*
	mixed	0.035
Constant		0.716*
pseudo R^2	0.0148	
χ^2	300.77	
df	30	

In the restricted model, the sample elements that were not processed are regarded as non-contacts. We see that the variable province and large cities is remarkably more significant in the restricted model for contact than it was in the elaborate model in section 3.4.3. For instance, the parameter estimate for the province of Limburg is highest, i.e. Limburg has a very high contact propensity. In the analysis of being processed in the elaborate response process in section 3.4.2, Limburg was excluded from the analysis because there were no unprocessed cases in the province of Limburg. Because of the merging of being processed and contact, this effect is now ascribed to a high contact propensity whereas we know from the analysis of the elaborate response process that this effect is more likely to be caused by a zero proportion of unprocessed cases. Or put differently, the contact propensity for the other provinces is lower than it should be because of non-contact due to unprocessed cases. Besides that, some variables were not significant anymore in the restricted model, namely number of persons and degree of urbanization. This may be indicative of some sort of confounding between the process propensity and the contact propensity.

Likewise, the sample elements that could not participate because they were not able (due to language problems) are regarded as refusing participation in the restricted model. As a consequence, the participation process in the restricted model receives an additional explanatory variable, namely the ethnic group. There is a strong negative effect of non-native groups on participation. However, as we showed in Chapter 3, this effect is caused by language problems. It is now mistakenly assigned to a lower participation propensity of non-native groups.

3.4.6.2 The simplified response process

Compared to the analyses of the elaborate- and the restricted response process, we now merge all the response types. Hence all the observations are either response or nonresponse. The corresponding indicator is R; $R = 1$ for respondents and $R = 0$ for nonrespondents. The total number of observations is $17,628$. The number of respondents is $11,012$ and the number of nonrespondents is $6,616$.

The final logistic model for the simplified response process is displayed in table 3.18.

Table 3.18: *Logistic model for the response propensity* $\hat{\rho}$

Variable	*Category*	β
Province + large cities	Friesland	0.708*
(Reference Groningen)	Drenthe	0.534
	Overijssel	0.590*
	Gelderland	0.631*
	Utrecht	0.531*
	(except Utrecht city)	
	Noord-Holland	0.365*
	(except Amsterdam)	
	Zuid-Holland	0.425*
	(except Den Haag, Rotterdam)	
	Noord-Brabant	0.626*
	Limburg	0.798*
	Amsterdam	-0.264
	Rotterdam	0.419*
	Den Haag	0.575
	Utrecht	0.783*
Listed telephone	Yes	0.416*
(Reference No)		
Type of household	Unmarried no/c	0.038
(Reference single)	Married no/c	0.098
	Unmarried with/c	0.052
	Married with/c	0.283*
	Single parent	0.261*
	Other	-0.038
	Mixed	0.271*
Ethnic group	Morrocan	-0.605*
(Reference native)	Turkish	-0.321*
	Suriname, NL Antilles	0.094
	other non-Western	-0.241*
	other Western	-0.233
	mixed	0.094
Average house value[7]	0 - 50	-0.090
(Reference missing)	50 - 75	-0.126

Continued on next page

[7] $\times$ 1,000 euro

Table 3.18: *Logistic model for the response propensity $\hat{\rho}$ - continued*

Variable	*Category*	*β*
	75 - 100	-0.161
	100 - 125	-0.158
	125 - 150	-0.071
	150 - 200	0.064
	200 - 250	0.054
	250 - 300	0.054
	300 - 400	-0.136
	400 and more	0.111
Constant		-0.253
pseudo R^2	0.0322	
χ^2	750.52	
df	37	

Compared to the restricted model for the participation propensity in the previous section, the first observation is that the pseudo R^2 for this model is slightly higher (3.2% versus 1.5%). Also, the variable paid job has disappeared from the model. In addition, the variables listed telephone and type of household entered the model. It seems that the contribution of these variables can be attributed to the contact propensity. Compared to the restricted response process, the variables that do not appear in the simplified response model are gender and average age.

It seems that the less specified the response process, the less we are able to understand the causes of nonresponse.

3.4.7 Summary of the models

In this chapter, we demonstrate how to deal with nonresponse with the use of linked data from external registers. For that purpose, we use data from the experimental LFS from the period of July - October 2005. We linked this data to a number of registers. We have analysed the four distinctive probabilities in the response process: being processed ξ, being contacted γ, being able θ and participating ϱ. In the analyses, we have not allowed for correlation between the different types of nonresponse. Instead, we have analysed each of the types conditional on the preceding process. For instance, the analysis of the participation propensity ϱ has been performed on the subsample of households that are processed, contacted and found able for participation.

In Chapter 2 we presented an overview of the determinants and correlates of contact and participation based on survey literature. These correlates are

reported for general household surveys. Our analysis concerns one specific survey; the Dutch Labour Force Survey. Additional to the two most common types of response, contact and participation, we analysed being processed and being able to participate. The analysis of the distinctive response types to a large extent replicates the findings in the survey literature. Some of the external registers that we linked to the sample are related to the survey topic, for instance having a paid job, an unemployment allowance, or being registered as seeking employment.

In table 3.19 the final models for the different response types are summarised. The variables are listed according to the Wald χ^2 value, the strongest variable first. It is remarkable that the variable province and largest cities enters into every model. The strongest effect for this regional variable is found for the propensity to be processed. Only 2% of the households were not processed in the field. The main correlate of the process propensity ξ is the variable province and largest cities. Especially in Amsterdam there is a large number of unprocessed cases. Thus, none of the household characteristics in the sample is disturbed due to unprocessed households. The propensity to be able θ in this analysis only comprises nonresponse due to language problems. The percentage of households that could not participate due to language problems constitutes another 2% of the total sample size. The model for the propensity θ has a very high model fit with a pseudo R^2 of 30.3%. The variable that has the largest contribution to the model, is the ethnic group. Furthermore, small other contributions come from the variables province and largest cities and percentage of non-natives. In areas with a high percentage of non-natives, the propensity θ is smaller compared to areas with a lower percentage of non-natives. The results for these two propensities ξ and θ are comforting in the sense that none of the variables that are related to the survey topic (and that are included in the analysis) seem to be disturbed by these nonresponse types. And besides that, the process

Table 3.19: *Summary of the different models for the response types*

Response type	*Variables*
Being processed	Province and largest cities, degree of urbanization
Contact	Gender, type of household, number of persons, Province and largest cities, listed telephone, degree of urbanization, average age
Being able	Ethnic group, province and largest cities, percentage of non-natives
Participation	Province and largest cities, paid job, average house value

propensity ξ is to a large extent correlated to the distribution of the fieldwork in the different provinces and large cities. Furthermore, the language problems are almost exclusively related to the ethnic group of the household core. None of the other household characteristics emerged in either of the two analyses.

The correlates of contact, conditional on being processed, confirm the general results reported in Chapter 2. The most important correlates are province and largest cities, average age and household type. Typically, the contact propensity is very low in Amsterdam. In general, the contact propensity in urbanized areas is lower than in rural areas. Older aged household cores are more easy to contact. Single households (especially when they consist of man alone) are difficult to contact whereas households with children present have a high contact propensity. The larger the number of persons in the household, the larger the contact propensity.

Besides these generally supported results, we found that having a listed telephone is a very good indication of the contact propensity. Households with a listed land-line telephone have a high contact propensity. Even when the effect of the other variables in the model is accounted for, this variable still contributes greatly to γ. This result has also been found by Schouten (2004). A possible hypothesis is that households that list their telephone number, compared to households that do not, are less socially isolated and therefore more open to communication. This should be further investigated. Based on the results of our analyses, we strongly recommend the inclusion of this variable in nonresponse analysis studies.

The final type of nonresponse, refusal, constitutes the largest part of the nonresponse. In our ELFS dataset, 37.5% of the households refused to participate. We analysed the participation propensity ϱ, conditional on the preceding types of nonresponse i.e. on being processed, contacted and able to participate. We found that we were not able to really grasp the participation propensity with the available variables in our analysis. The pseudo R^2 equals 0.86% and the total χ^2 is only 166.79. Caution is in place when interpreting the results of a model with such a poor model fit. There are three variables in the final model; province and largest cities, having a paid job and average house value. These results also confirm the analyses described in Chapter 2. Again, the inclusion of province and largest cities is caused by Amsterdam. Households in Amsterdam have a lower process propensity, a lower contact propensity and also a lower participation propensity. Having a paid job is an indication of a higher participation propensity ϱ. The participation propensity model is the only model in which one of the linked variables that can be associated with the topic of the survey plays a role. It remains questionable how much attention should be paid

to this effect because of the poor model fit. The last variable in the model is the average house value. If we interpret this variable as an indication of the SES, lower SES groups display a lower participation propensity.

In section 3.4.6 we compare the results from the elaborate analysis with two more simplified response processes. The first comparison is made with the restricted response process consisting of contact and participation. This restricted model finds less relationships between the variables. Furthermore, compared with the elaborate analyses on contact and participation, the variable ethnic group enters into the participation model but this effect can be clearly attributed to language problems. For the contact model, the variable province and largest cities is more significant than before. This effect is most probably caused by the unprocessed cases. The merging of the different response types seems to cause a confounding of the effects of the variables. The effects become even more obvious when the response process is simplified to only response and nonresponse. We can conclude from this comparison that, as expected, analysing the more elaborate response process provides a better understanding of the causes and correlates of (non)response.

3.5 Concluding remarks

The assumption of no correlation between the different types in the response process might not be realistic, see e.g. Nicoletti and Peracchi (2005). However, for the purpose of analysing the different steps we made the assumption for reasons of simplicity. In Chapter 8 we present a model that does allow for correlation between the different types of nonresponse and we come back to this assumption in more detail.

The analyses are all made on a household level. Due to the aggregation of variables on a personal level, these analyses have a lower explanatory power than analyses where the unit of analysis is the person. Despite of this, the models for ξ and γ fit the data reasonably well. The model for θ has a good model fit. Only the model for ϱ is weak.

Chapter 4

Re-Approaching Nonrespondents

In this chapter we present two approaches to obtain extra information from nonrespondents on the survey items: the call-back approach and the basic-question approach. The approaches have been applied to the 2005 Dutch LFS in a pilot study. We discuss the results from this study. Furthermore, the extra response in the two approaches is evaluated with the use of linked data described in Chapter 3.

4.1 Introduction

In Chapter 3 we analysed the different types of response with the use of linked data. Sometimes, however, there is no data available for linkage. Another way of obtaining information on nonrespondents is by simply asking them. In this chapter, we describe two approaches to obtain additional information from the nonrespondents on the survey items: the call-back approach and the basic-question approach. Hansen and Hurwitz (1946) proposed to investigate nonresponse in mail surveys by taking a sample of nonrespondents and trying to obtain the required information by means of a face-to-face interview. This so-called call-back approach (CBA) consists of taking a sample of nonrespondents and re-approach selected initial nonrespondents. In the re-approach one may choose to use a different data collection mode, specially trained interviewers, an extended fieldwork period, or incentives for the prospective respondents.

The call-back approach is a rather expensive method to obtain information from nonrespondents. Furthermore, the CBA will considerably lengthen the fieldwork period. An alternative to the call-back approach is the basic-question approach (BQA). This method can be applied when the call-back approach is no option due to time- or budget constraints. The basic-question approach assumes that many survey questionnaires are composed around a few basic questions. With the answers to these questions the most important conclusion of the survey can be formulated. The procedure was first proposed by Kersten and Bethlehem (1984). They observed that persons who refused to participate in a survey often could be persuaded to answer just a few basic questions. The main objective of the basic-question approach is to gain insight in possible differences between respondents and nonrespondents with respect to the most important variables of the survey. If such differences are detected, the approach also provides information for adjusting estimates for other variables. For applications of the BQA see for instance Van den Brakel and Renssen (1998) and Voogt (2004).

Both the call-back approach and the basic-question approach re-approach a sample of nonrespondents. If there is no nonresponse to the re-approach or the nonresponse is MCAR, an indication of the difference between respondents and nonrespondents is obtained.

Example 4.1.1 An application of the BQA and the CBA
Voogt (2004) has applied both the call-back- and the basic-question approach in his research on nonresponse bias in election research. He selected a simple random sample of 995 voters from the election register of the town of Zaanstad in the Netherlands. There were two basic questions in his survey:

- Did you vote in the parliamentary election on Wednesday May 6, 1998?
- Are you interested in politics, fairly interested or not interested?

In the first wave of the survey, people were contacted by telephone if a telephone number was available. If not they were send a paper questionnaire by mail. The BQA was applied in a separate follow-up. All refusers were offered the possibility to answer just the two basic questions (by telephone or by paper questionnaire). The CBA was applied to those persons who refused to answer the basic questions. This time the refusers were visited at home by interviewers. The results of the fieldwork are summarised in table 4.1. One conclusion from this table is that the situation need not be hopeless if the response is low in the first wave. With additional measures response rates can be increased substantially. This conclusion is in line with Stoop (2005).

Table 4.1: Response in the election research by Voogt (2004)

Result	*Cases*	*Percentage*
Response in first wave	508	51.1%
Response in basic-question approach	196	19.7%
Response in call-back approach	224	22.5%
Final nonresponse	67	6.7%
Total	995	100.0%

Because the researcher had access to the voting register of the town, he could establish with certainty whether all 995 people in the survey had voted or not. In this group 72.9% had voted. The voting behaviour for the various groups is presented in table 4.2. The groups are ordered in growing reluctance to participate. There seems to be a relationship between this reluctance and voting behaviour: the more reluctant the group the lower the percentage of voters. If we assume that the response in the basic-question approach is representative for all the nonresponse after the first wave, an estimate for the percentage of voters would be $(508/995) \times 85.4 + (487/995) \times 66.3 = 76.4\%$. This value is much better than the 85.4% for the initial response, but still too high. This implies that the response in the basic-question approach is not representative for all the nonresponse. ■

In this chapter, we describe the results from a study on the Dutch LFS with the basic-question approach and the call-back approach. We use the linked data to validate the results of the two re-approaches. The design for the LFS and the available linked data for the analysis are described in Chapter 3. We discuss in detail how the re-approaches were implemented, and we provide recommendations for further fine-tuning of the approaches. In addition, we analyse the

Table 4.2: Voting behaviour by follow-up approach

Group	*% voters*
Response in first wave	85.4%
Response in basic-question approach	66.3%
Response in call-back approach	55.8%
Final nonresponse	53.7%

response that is obtained in the two re-approaches with the use of linked data. We also compare the estimates of the main survey items based on the regular response with the estimates based on the additional response obtained in the two approaches. And finally, we discuss whether the nonresponse after the re-approaches has become more selective than the regular nonresponse.

This chapter is organised as follows. In section 4.2, we discuss the design of the pilot in general, and the implementation of the two re-approaches in particular. The data are analysed in section 4.3, where we also answer the main research questions. Finally, in section 4.4 we summarise the main findings and make some recommendations for further research.

4.2 Design of the re-approach strategies

4.2.1 General design of the LFS pilot

The pilot was motivated by the studies by Stoop (2004) and Voogt (2004), where good results were obtained in re-approaching nonrespondents. The study by Stoop (2004) re-approached a sample of 350 refusers that were classified by the interviewers as hard to persuade. It was not possible to link the data to external data sources so that little information was available on the nonrespondents, see Schouten and Bethlehem (2002). The study by Voogt (2004) surveyed the small town of Zaanstad in the Netherlands about voting behaviour, see also example 4.1.1. He had access to the voting register of Zaanstad, which enabled him to confront the answers to the survey questions with the real voting behaviour. He investigated both nonresponse bias and response bias in election research. The experiences from the studies from Stoop (2004) and Voogt (2004) were useful inputs for our own pilot study, for example regarding the sample size and types of respondents in the re-approach.

From July to December 2005, we re-approached a group of nonrespondents to the LFS. We decided that, besides refusals, also former non-contacts and unprocessed cases were eligible for the re-approach and we used a larger sample size than Stoop (2004). We linked the dataset to the SSB for a large number of auxiliary variables. Some of these variables are related to employment status, e.g. having a paid job or receiving an unemployment allowance (see Chapter 3 for a detailed description of the linked data).

In 2005, the Netherlands was divided in thirteen interviewer districts. Eleven districts participated in the pilot. Two districts could not participate because of a lack of available interviewers. In the participating districts two or three

of the best interviewers were selected, depending on the geographical size of the district. Because the respondents were to be re-approached by a different interviewer than the interviewer that obtained the original nonresponse, we needed at least two interviewers per district to assign each address to a new interviewer. In five districts a number of municipalities were excluded so that travelling distances were acceptable. For more details, see Schouten (2007). The exclusion of these municipalities resulted in the ELFS-sample that we also used for the nonresponse analysis in Chapter 3.

The basic-question approach was not conducted with face-to-face interviewers. However, we selected households from the same districts and municipalities as for the call-back approach to ensure that the target populations of the two approaches were similar. Due to the exclusion of a number of districts and municipalities, the estimates based on the re-approaches cannot be directly compared to the estimates in the regular LFS. However, the selected municipalities cover the full range from rural to strongly urbanized areas and make up about 75% of the population. Furthermore, if we compare estimates for the re-approach and the regular LFS, we select the same municipalities from the regular LFS. See table 4.3 for the sizes of the different groups in the ELFS-sample.

4.2.2 The call-back approach

In the call-back approach we selected samples of eligible nonresponding households for the months July to October 2005. The total sample size for the CBA equals 775 households. The features of the CBA-design are summarised in table 4.4.

The questionnaire and interview mode of the CBA are the same as in the regular LFS to avoid mode- and questionnaire effects. No advance letter was sent to nonrespondents to prevent alarming them or giving them additional reasons

Table 4.3: *Groups in the re-approach*

Group	*Size*
Eligible households	18,074
Regular response	11,275
Regular nonresponse	6,799
Eligible nonresponse	6,170
Call-back approach	775
Basic-question approach	942

Table 4.4: *Design features of the call-back approach*

Feature	*Call-back approach*
Sample unit	All eligible household members aged 15+
Questionnaire	Original LFS questionnaire
Mode	CAPI
Fieldwork period	2 months
Control group	No
Advance letter	No
Timing	One week after being processed as nonresponse
Incentive	Yes
Special interviewers	Yes
Eligible response types	Refusal, non-contact, unprocessed cases

to refuse cooperation. The addresses in the CBA were allocated to a different interviewer than the interviewer that received the original LFS non-response.

A total of 28 interviewers was selected from the best CAPI interviewers. The interviewers received an additional training in doorstep interaction. Furthermore, a special newsletter was made to inform them and halfway through the study a meeting was organised to exchange experiences and to boost motivation. The interviewers could offer incentives in the form of gift vouchers as they deemed fit, and in addition the interviewers themselves could receive a bonus based on their response rate.

Paper summaries were sent to the interviewers, containing background information on the composition, gender and age of the household, the type of non-response and the timing and number of previous contact attempts. The interviewers could use this information to fine-tune their contact attempts and prepare the doorstep interaction. Interviewers were, however, not obliged to adapt their strategy to the background information. It was solely provided to support them.

We used an extended fieldwork period of two months to enhance participation, which causes a delay in time for the call-back response with respect to the LFS response. It is known that employment has seasonal and cyclical components. Hence, a delay in time may imply that we find a difference in employment rate that is to be attributed to the month of observation and not to non-response bias. Unless a time series model is posed for employment, it is not

possible to disentangle the effects of non-response and time. We will not do that but accept the time lag. Furthermore, we assume that the time effect is small.

In the analysis we assigned weights to all households selected for the CBA in the LFS nonresponse to let these households represent all eligible households, see table 4.5. The months July - August and September - October have different selection probabilities due to a different workload for the interviewers.

4.2.3 The basic-question approach

In the basic-question approach we also selected samples of eligible nonresponding households for the months July to October 2005. The design features of the BQA are summarised in table 4.6. The design of the BQA is more different from the original LFS than the design of the CBA. The questionnaire for the BQA contained a small number of questions to determine the key variable of the Labour Force Survey, i.e. the labour force status (see Chapter 3). The time it took to answer the questions varied between 1 and 3 minutes, depending on the labour force status of the respondent. In the regular LFS, these questions are asked in the beginning of the questionnaire. The questions are thus not taken out of their context, and we assume that there are no questionnaire effects.

The design for the BQA uses multiple modes. Addresses for which a listed land-line telephone number could be linked to the survey sample were approached by CATI. The households without a listed land-line telephone actually consist of three groups: households that have an unlisted land-line telephone, households with only (a) mobile telephone(s), and households without a telephone. We cannot make a distinction between these groups, although they are different on socio-economic variables, see e.g. Vehovar et al. (2004) and Callegaro and Poggio (2004). Addresses without a listed land-line telephone were approached by a combination of a paper questionnaire and a web questionnaire. An advance letter was sent to the selected addresses. The letter contained a personal login to a secured website where the web survey could be found, as

Table 4.5: *Weights and sizes of eligible households for the CBA*

Group	*Total size*	*Size selection*	*Weight*
Eligible CBA July-Aug	3,039	391	7.77
Eligible CBA Sep-Oct	3,133	384	8.16
Not eligible July-Oct	629	0	1

Table 4.6: *Design features of the basic-question approach*

Feature	*Basic-question approach*
Sample unit	One randomly selected person per household
Questionnaire	Strongly condensed questionnaire containing only the key questions of the LFS to determine the employment situation
Mode	Mixed-mode: Combination of CATI, web- and paper surveys
Fieldwork period	1 month
Control group	Yes
Advance letter	Yes
Timing	One week after being processed as nonresponse
Incentive	No
Special interviewers	No
Eligible response types	Refusal, non-contact, unprocessed cases

well as a paper questionnaire with a return envelope. In the letter the selected person in the household was presented the choice between the web questionnaire and the paper questionnaire.

As mentioned before, the idea behind the BQA is that respondents who refuse to participate can often be persuaded to answer a few questions. However, asking too many questions risks getting no information at all. To reduce the response burden and thus stimulate participation we decided to only interview one randomly selected person per household employing the next birthday method, see section 4.2.3.1.

In the BQA one may allow the interviewers to ask the basic questions immediately after they received a refusal for the main questionnaire. This may lead, however, to higher nonresponse rates for the main questionnaire, as Van den Brakel and Renssen (1998) and Van Goor and van Goor (2003) found. Interviewers may switch to the BQA too soon, thereby shortening the doorstep interaction and putting less effort in refusal conversion. Therefore we re-approached the nonrespondents one week after the household was processed as a nonresponse, with a different interviewer.

We also approached a fresh control sample of $1,000$ households with the basic-questionnaires. Ideally, the basic-questionnaires should also have been pre-

sented to a sample of LFS-respondents. That would enable us to directly compare the answers obtained by the nonrespondents with that of the respondents. However, the experiment was designed not to intervene with the regular LFS. Therefore, we chose to approach a fresh control sample instead of LFS respondents to analyse possible mode effects and the effect of the random respondent selection procedure, see sections 4.2.3.1 and 4.2.3.2. Table 4.7 summarises the different groups in the basic-question approach. In the analysis we assigned weights to all households selected for the BQA in the LFS nonresponse to let these households represent all eligible households. See table 4.8. The response rate in CATI is much higher than in the paper and web variant. This result is obtained in both the BQA for the LFS nonrespondents and the control group. Interviewer-assisted data collection modes in general obtain higher response rates due to the persuasive power of the interviewers. Moreover, we know from previous research (Schouten 2004) that persons without a listed telephone less often participate in surveys. Furthermore, Griffin et al. (2001) found that offering multiple modes of response in a mailing led to a lower response rate. They suspect that offering a mode of response other than mail, in combination with a paper questionnaire, contributes to a break in the response process and thus results in a lower overall response rate.

The design of the BQA hence causes a strong relationship between telephone ownership and response as a result of the low response rates to the paper and web questionnaires. This finding confounds the analysis of response patterns and mode effects. Therefore, we chose to restrict these analyses to households with a listed land-line telephone. In table 4.9 we give the sizes of the various groups.

Table 4.7: *Groups in the basic-question approach*

	LFS nonresponse		control group	
	sample	*response rate*	*sample*	*response rate*
Eligible households	6,170			
Selected households	942	40%	1,000	62%
CATI	573	55%	667	80%
Paper/Web	280	23%	333	25%

Table 4.8: *Weights and sizes of eligible households for the BQA*

Group	*Total size*	*Size selection*	*Weight*
Eligible BQA	6,170	942	6.55
Not eligible BQA	629	0	1
LFS response	11,275	0	1

Table 4.9: *The different groups with a listed telephone in the BQA and the regular LFS*

Group	*Telephone*	*Response*	*Response rate*
Regular LFS	10,137	6,893	68%
Nonresponse BQA	573	315	55%
Control group	667	534	80%

4.2.3.1 Within-household respondent selection techniques

One important aspect of the basic-question approach is the within-household respondent selection. The literature identifies several methods to randomly select one person within a household, see for example Kish (1949), Salmon and Nichols (1983), Oldendick et al. (1988), Binson et al. (2000) or Gaziano (2005). The main objective of these procedures is to obtain a sample that is representative of the target population of the survey. However, recently Clark and Steel (2007) discussed the selection of respondents within households to derive a maximal accuracy of survey items as persons within households are often very similar with respect to survey topics. Binson et al. (2000) compare the three methods that are most often used, the Kish-grid, the last birthday method and the next birthday method. Based on their findings, we decided to use the next-birthday method in the BQA.

The pre-notification letter of the paper/web survey instructs that the questionnaire has to be completed by the person of 15 years or older who is the next to have its birthday. In the telephone version, the interviewer explains that the questions have to be answered by (or for, because proxy is allowed) the person in the household that is at least 15 years old and that is the next to have its birthday.

We analysed the control sample to validate the respondent selection with the next birthday method. For the analysis, we compared the birth date of the

respondent to the birth date of the intended respondent. Unfortunately, only the data from the CATI survey facilitated this analysis (667). Furthermore, we had to leave out a number of addresses: addresses of the nonrespondents (170), addresses where the reported birth date was not found in the register (69), addresses where more than one household was registered (22), addresses where the birthday was in the month of the interview (53) and addresses where more than one person had its birthday in the month of the next birthday (31). We had to leave out the last two groups as we could not link exact birth dates, but only the year and the month.

The total observations for the analysis thus reduced to 322. On these addresses, the respondent was the intended respondent in 270 cases. That is, in 83.9% of the cases we obtained answers for the person that was the next to have its birthday. Of those addresses, 91 were a single person household in which the person is by definition the intended respondent. On the remaining 16.1% addresses the person who responded was not the intended respondent. These results are slightly better than the results found by Lavrakas et al. (2000).

The percentage of correct selection is high, but there is still a large number of addresses where the selection failed. However, if the selection went wrong it is still possible to obtain a random sample of within-household respondents. To verify this, we compared the composition of the intended BQA respondents to the composition of the other respondents. For this analysis, we did not include the single households where the selection is by definition successful. The total number of observations than reduces to 231. We used Pearson's χ^2 with $\alpha = 0.05$ as a test for independence. Table 4.10 displays the results.

Table 4.10: *The distribution of auxiliary variables for a correct selection and a false selection (in percentage)*

		The respondent is the intended respondent	
Variable	*p-value*	*Correct*	*False*
Gender	0.080		
Male		46	60
Female		54	40
Position in household	0.003		
Child		5	19
Core member		95	77
Other		1	4
Ethnicity	0.479		
Native		87	83

Continued on next page

Table 4.10: *The distribution of auxiliary variables for a correct selection and a false selection (in percentage) - continued*

		The respondent is the intended respondent	
Variable	*p-value*	*Correct*	*False*
Non-native		13	17
Age	0.186		
15 to 34 y.		26	35
35 to 54 y.		45	31
55+		30	35
Household size	0.689		
2		48	52
3		22	19
4		25	19
5		5	8
6		1	2
Household composition	0.295		
Couple		44	48
Couple with children		48	44
One parent		7	6
Other		0	2
Number of call attempts	0.445		
1		60	54
2+		40	46
Total group size		179	52

The results indicate that there is a difference with respect to position in the household and gender. It appears that the composition of the respondents is distorted when it comes to children. We have to note that there are only 18 households where the next birthday is a child. In those 18 households, only 8 times the child responded. In the other 10 cases, the basic questions have been answered by a parent. When the respondent is a child (only 16 households), this is the intended respondent in 50% of the cases. The slight distortion in gender might be a result of the at-home effect. As Lavrakas et al. (2000) note, also in our sample it seems that women are slightly over represented. However, this result is not significant, possibly due to the small sample. We do not find a relationship between the number of persons on the address, household composition, ethnicity or the age of the next birthday-person. We did not spend more efforts (in terms of call attempts) on addresses where the selection succeeded. This implies that the small over representation of women in our sample is not caused by the fact that they are more cooperative and available, since the same number of call attempts was made to contact them.

The analysis in table 4.10 is bivariate. We also implemented a multivariate analysis to account for the influences of other variables. Therefore we performed a logistic regression of the indicator TR_i for the intended next birthday respondent ($TR_i = 1$) or other respondent ($TR_i = 0$), $i = 1, \ldots, 231$. We included as explanatory variables all variables in table 4.10. It turns out that only the household composition is significant in this analysis. The p-value for a test of significance equals 0.016. When adjusting for all other variables, the true distortion is caused by the household composition. When there are children present in the household, the selection more often fails than not. With respect to gender, ethnicity, age, household size and number of contact attempts we obtain a random sample of within-household respondents.

Our main interest lies in the labour force status, the basic-question variable. Therefore, we tested bivariately for a difference in labour force status between the intended respondents and other respondents. The p-value for the χ^2 test of independent equals 0.407. Moreover, a multinomial regression of the labour force status on the variables in table 4.10 and the indicator TR_i showed no significant influence of TR_i on the labour force situation, when adjusted for the influences of the other, auxiliary variables. Hence we conclude that the next birthday selection method did not result in a bias in the basic-question variable.

4.2.3.2 Mode effects

The objective of the re-approach of nonrespondents is the detection of differences between respondents and nonrespondents. The implementation of the basic-question approach involves a change in both the questionnaire and the interview modes. Therefore, we face the risk of errors other than nonresponse error. For an introduction to survey errors see for example Groves (1989) and Biemer and Lyberg (2003). A change of wording and context as well as a change in mode may seriously affect the answers, and, hence, statistics that are based on the survey. When such mode effects are imminent, we cannot draw any conclusions about nonresponse error, because we cannot decompose the total survey error. We refer to mode effects as the sum of all errors that occur because of a different interview mode and the use of the condensed questionnaire with basic questions.

The control group consists of households that have not been contacted before; therefore this group can be used to assess mode effects of the BQA. These mode effects can be made visible by comparing the composition of the BQA response in the control group to the LFS response. Because the control group and the regular LFS response are different samples, it is likely that the nonresponse error is also different. The nonresponse error could confound the mode effect,

and we thus have to take the nonresponse error into account when analysing the mode effect.

By linking auxiliary data to both samples we can directly assess nonresponse error with respect to the auxiliary variables (see Chapter 3 for a detailed description of the linked data). From the analysis of nonresponse error we can define strata in which households have a homogeneous response behaviour with respect to the auxiliary variables. Within those strata we can investigate the mode effects by considering the answers to the survey questions. We can still not fully disentangle nonresponse error from mode effects, however. Within strata households may have a BQA response behaviour that is different from the LFS response behaviour when it comes to the survey topics. Therefore, we settle for conclusions about the sum of mode effects and remaining nonresponse error. In practice we will always have to deal with them simultaneously.

For the analysis, we had to account for the difference in design between the two samples. In the BQA only one person in the household was selected. We therefore assigned weights to the BQA respondents equal to 1 / (number of persons in the household). Based on the combined respondents data, the indicator LFS_i is defined to be equal to 1 if person i responded to the regular LFS, and $LFS_i = 0$ if person i responded to the CATI BQA, where $i = 1, 2, \ldots, N$ and $N = 6,893 + 534 = 7,427$. A bivariate comparison of BQA respondents and regular LFS respondents revealed large differences between the two groups for telephone ownership ($p = 0.000$) and whether the household received an unemployment allowance ($p = 0.006$).

We performed a weighted logistic regression to analyse the response patterns in the two groups, see for example Johnston and Dinardo (1997) or Greene (2003). The logistic regression models the probability that a person is a respondent in the LFS, given that he or she is a respondent in either the LFS or the BQA. By using weights we adjust for the unequal inclusion probabilities in the two samples. If response probabilities are the same for the LFS and BQA, then the model explains the difference in the number of sampled addresses in the BQA and LFS. Clearly, this difference is independent of any background characteristic. Hence, in that case it suffices to incorporate an intercept in the logistic regression. Of course, we know that the response probabilities are different for the LFS and BQA. However, if the response probabilities are the same within the LFS and also within the BQA, then again a logistic regression model with an intercept is sufficient. As a consequence, if any of the background characteristics is incorporated in the logistic regression it means that the BQA and LFS response mechanisms are different. This difference is not just in overall level of response but also relative within different groups in the population. The

logistic regression model thus points to those groups in the population that have a different response behaviour in the LFS and the BQA.

We used the adjusted Wald F statistic (Skinner et al. 1989) to test the joint significance of the categorical variables. The final logistic model contains only two significant variables: disability allowance ($F = 4.58$; $p = 0.032$) and social allowance ($F = 7.34$; $p = 0.007$). With respect to these two variables, the response amongst the telephone-owners in the BQA control sample is selective compared to the regular LFS.

If we assume that the response mechanism is MAR, then within the strata defined by disability allowance and social allowance the households are homogeneous. In other words in those strata respondents and nonrespondents cannot be distinguished. This assumption cannot be tested, however. The response mechanism may be NMAR with respect to the main survey items. We, however, used auxiliary variables that are closely related to employment like having an unemployment allowance, having a paid job and being subscribed to the CWI database. As a consequence the impact of the detailed employment status on nonresponse error is reduced.

We can now proceed to compare the survey item for the LFS: the employment status. To account for the different response patterns, we again perform a weighted logistic regression. The dependent variable of the regression is the indicator LFS_i, like in the nonresponse analysis. As explanatory variables we include the variables disability allowance and social allowance to adjust for the difference in response pattern (i.e. nonresponse bias), and the survey item employment situation to see whether this variable is different in the two groups. We fitted this model to the respondents in both the LFS and the control sample that have a listed land-line telephone, and then we used the adjusted Wald statistic to test for the joint significance of the categories of the variables. The results are given in table 4.11. The null hypothesis of independence on employment status conditional on receiving a disability and/or social allowance is not accepted, i.e. $p > 0.050$. We, hence, find no evidence that employment status

Table 4.11: *Analysis of mode effects in the BQA*

Variable	*F-value*	*p-value*
Employment situation	2.22	0.109
Disability allowance	4.00	0.046
Social allowance	8.83	0.003

is affected by mode effects. However, we have to note that the p-value for the employment status is rather low.

4.3 Analysis of the response in the re-approach strategies

We analyse the additional response that is obtained by the call-back and basic-question approaches by answering two questions: What households participate in the re-approach strategy, and do we get new households? By new households we mean households that have different characteristics than the ones in the regular LFS response. We investigate these questions separately for both re-approaches with the use of the linked data described in Chapter 3.

4.3.1 The call-back approach

4.3.1.1 What households respond in the call-back approach?

We compare the call-back response with the full call-back sample of initial nonrespondents. First, we perform bivariate tests for independence between response and the available auxiliary variables. Second, we fit a logistic regression model to estimate the probability of response in the call-back approach. Note that both the tests and the response probabilities are conditional on an initial nonresponse in the LFS.

Table 4.12 gives p-values corresponding to Pearson's χ^2 test for independence. The only variable that shows a significant association with response at the 5% level is the average age of the household core. However, if we account for multiple testing, then we would not reject independence of call-back response behaviour and average age. p-values for the geographic variables region and province and largest cities are also small but not significant. Table 4.13 gives the fit and the corresponding parameter estimates of the logistic regression model for call-back response. The resulting model has only two variables: average age and part of country. The proportion of variance that is explained by the model is small, 2.5% using Nagelkerke's pseudo R^2, see (3.2). The lack of dependence with the available covariates is a remarkable result; even with the relatively small sample size of 775 addresses in mind. It implies that the response in the call-back approach is not very selective. From table 4.13 we can conclude that the households with an average age between 35 and 54 years had a higher response rate in the call-back approach. Furthermore, in the eastern

Table 4.12: *Bivariate test of independence between response and nonresponse in the CBA*

Variable	*p-value*	*Variable*	*p-value*
Job	0.913	CWI registration	0.488
Social all.	0.893	Self-employed	0.471
Disability all.	0.842	Urbanization	0.199
Gender	0.793	Ethnicity	0.135
Unemployment all.	0.656	Household type	0.110
Average house value	0.644	Part of country	0.080
% non-native	0.625	Province and largest cities	0.072
Listed telephone	0.502	Average age	0.048

parts of the Netherlands fewer households participated than in the other parts of the country.

4.3.1.2 Do we get new households in the call-back approach?

We investigate the question whether we get new households in two ways. First, we compare p-values of univariate tests for independence between response and several auxiliary variables before and after the call-back response is added to the regular LFS response. Second, we construct logistic regression models for the LFS response alone, for the combined response of LFS and call-back, and for the type of response given a household responded in either the LFS or call-back.

Table 4.14 gives p-values for bivariate tests for independence between response and linked data for the LFS response and for the combined response of LFS and call-back approach. For many variables independence of response behaviour is rejected at 5% or 1% levels before and after the combination of the

Table 4.13: *Logistic regression model for response and nonresponse in the CBA*

Variable	*Category*	β
Constant		-0.207
Average age	35 - 54	0.490*
(Reference < 35)	> 54	0.096
Part of country	East	-0.868*
(Reference North)	West	0.051
	South	-0.332

response groups. The exceptions are average age, CWI registration, unemployment allowance, self-employed and disability allowance. Surprisingly, the p-value for age is high. Age is often one of the strongest explanatory characteristics for response in household surveys, see Chapter 3. In the ELFS-sample we do not find a strong relation to response at the household level.

From table 4.14 we can conclude that the p-values have increased for all variables after the call-back response is added to the LFS response. After inclusion of the call-back response independence is not rejected at the 5% level for variables average age and disability allowance. These conclusions are the same for the LFS response alone. However, at the 1% level independence of gender, average house value and paid job is not rejected anymore, contrary to the LFS response. These results imply that we find smaller deviations from independence between response behaviour and household characteristics. The distributions of these characteristics in the combined response are closer to the distributions in the original sample. In general, we conclude that the response has become more representative because of the additional response that is obtained in the call-back approach. In the last column we give p-values corresponding to weighted

Table 4.14: *Bivariate test of independence between response behaviour and different groups*

	Pearson's χ^2 p-value		
Variable	*LFS response*	*Combined*	*LFS vs CBA*
Urbanization	< 0.009	< 0.009	0.00
Listed telephone	< 0.009	< 0.009	0.00
Part of country	< 0.009	< 0.009	0.00
Province and largest cities	< 0.009	< 0.009	< 0.009
Gender	< 0.009	0.02	0.00
Ethnicity	< 0.009	< 0.009	0.00
Average house value	< 0.009	0.03	0.02
% non-native	< 0.009	0.00	0.00
Household type	< 0.009	0.00	0.01
Paid job	< 0.009	0.04	0.10
Disability allowance	0.00	0.16	0.54
Social allowance	< 0.009	0.00	0.10
CWI registration	< 0.009	0.38	0.26
Unemployment allowance	0.73	0.86	0.62
Self-employed	0.52	0.98	0.17
Average age	0.32	0.66	0.05
Job status	-	-	0.56
Employment status	-	-	0.32

bivariate tests for independence of response group, LFS or call-back, and the various linked auxiliary variables. Here, we have first isolated the respondents from the LFS and the call-back approach and tested whether households with different characteristics have different distributions in the two groups. As the analysis is now restricted to respondents we can also give the bivariate tests for the survey items job status of the household and employment status of the household. The independence for the two survey items is not rejected. Between the two response groups there is no significant difference in job and employment status on the household level. Also, none of the job related auxiliary variables lead to a rejection of independence of response group at the 5% level. For all other auxiliary variables independence is rejected at the 5% level. The strongest differences are in the geographical variables and the availability of a listed land-line telephone.

Next, we fitted three logistic regression models. One for the probability of response in the regular LFS; one for the probability of response in the combined response of LFS and call-back; and one for the probability of being a LFS respondent given that a household responded in either the LFS or call-back.

In table 4.15 the three resulting models are given. In all cases we applied a forward selection algorithm with a 5% significance level. Hence, only significant variables are shown. '*' denotes significance at a 5% level, '**' denotes significance at a 1% level. For the last two regressions we performed a weighted logistic regression to account for the sampling of nonrespondents in the call-back approach. For the third regression we added the survey items job and employment status.

From table 4.15 we conclude that the call-back approach does bring in new households; LFS and call-back respondents are different with respect to part of country, having a listed land-line telephone and ethnicity. As a result, the logistic regression model for the combined response finds less or weaker relations than the model for the LFS response alone. The only exception is the variable telephone, which becomes slightly more unbalanced. However, we do not find a significant relation with the two survey items. In other words, if we keep demographic characteristics the same then job and employment status do not have significantly different distributions.

Table 4.15: *Logistic regression models for the response behaviour in different groups in the CBA*

		Analysis		
Variable	*Category*	*LFS*	*LFS + CBA*	*LFS vs CBA*
Constant		0.414*	0.723**	2.15**
Urbanization	strong	0.196*	0.309**	
(Reference very strong)	moderate	0.288*	0.160	
	little	0.312*	0.405**	
	not	0.258*	0.450**	
Listed telephone	yes	0.378**	0.447**	0.38**
(Reference no)				
Part of country	East	-0.437*	0.090	-1.02**
(Reference North)	West	-0.558**	-0.224	-1.03**
	South	-0.091	0.276	-0.61
Household type	not married	0.049	0.091	
(Reference single)	married	0.087	-0.130	
	unmarried w/children	0.134	0.200	
	married w/children	0.372**	0.249*	
	single parent	0.314*	0.264	
	other	0.181	0.022	
	>1 household	0.483*	0.745**	
Ethnicity	Moroccan	-0.898**		-1.00**
(Reference native)	Turkish	-0.372		-0.18
	Suriname/Antilles	-0.040		0.22
	other non-Western	-0.415**		-0.52*
	other Western	-0.345*		0.18
	mix	0.023		0.31

We have predicted the corresponding distributions conditional on the known auxiliary variables, i.e. the call-back response is missing-at-random. Furthermore, the results of the logistic regression suggest that the response became more representative. After all, the number of significant variables reduced after the call-back approach. This conforms to the results of section 4.3.1.1 where we found that the call-back approach was not strongly selective.

4.3.2 The basic-question approach

4.3.2.1 What households respond in the basic-question approach?

For the analysis of the BQA we follow the strategy of section 4.3.1.1. First, we perform bivariate tests. Next we construct logistic regression models for the basic-question response. p-values for the tests of independence between response in the basic-question approach and the linked data are given in table 4.16. These values are computed for all households and for the households with a listed land-line telephone. In the full population, the variables that show a significant association with response at the 1% level are telephone, gender, job and social allowance. At a 5% level independency is rejected additionally for the variables ethnicity and household type. When we restrict the analysis to households with a listed land-line telephone we find three auxiliary variables that give a significant explanation for response: average age of the household, having a job and self-employment of a household core member. Next, we construct a logistic regression model for the probability of response when the basic question strategy is applied to LFS nonrespondents. Again we perform the analysis for all households and for the listed households. The model and regression parameters are given in table 4.17. '*' denotes significance at a 5% level, '**' denotes significance at a 1% level.

The model shows significant differences in basic question response for households with different scores on telephone, age, job and gender characteristics. In table 4.17 we see that the households that consist only of women or a mixture of men and women have a higher response rate, as do households where at least one person had a job at the 1st of January 2005. Having a listed land-line telephone has a large influence on the probability of response in the basic question re-approach, see section 4.2.3. This can to a large extent be explained by the data collection design. Based on telephone ownership households were assigned either to CATI or to a paper/web questionnaire. The response in the CATI group was much higher, and this is reflected in the high regression parameter for telephone. The relation between response behaviour and having a

Table 4.16: *Bivariate test of independence between response and nonresponse in the BQA*

	p-value for Pearson's χ^2 test	
Variable	*All households*	*Telephone households*
Urbanization	0.42	0.58
Telephone	< 0.009	-
Part of country	0.97	0.76
Province and largest cities	0.52	0.33
Gender	0.00	0.31
Ethnicity	0.04	0.93
Average house value	0.19	0.94
Household type	0.03	0.14
Paid job	0.00	0.02
Disability allowance	0.35	0.75
Social allowance	0.01	0.42
CWI registration	0.08	0.97
Unemployment allowance	0.40	0.60
Self-employed	0.12	0.01
Average age	0.06	< 0.009

job is alarming, since having a job at the 1st of January 2005 can be expected to relate strongly to unemployment in the summer of 2005. We conclude that the BQA is more selective than the CBA; there is a stronger relation between the selected auxiliary variables and response.

The logistic model changes when we restrict the analysis to the households with a listed land-line telephone. We do not find a significant relation with having a job at January 1st and with the gender composition of the household core. We do, however, find differences in response for age and self-employment within the households with a listed telephone. Household cores with an average age below 35 years respond better than the other households. The same is true for households of which none of the household members is self-employed, which relates to the key items of the LFS. The relation between gender and BQA response behaviour must, therefore, arise from the unlisted households that received a paper and web questionnaire. The bivariate relation between having a job and BQA response is still present for the listed households, see table 4.16, but disappears when conditioning on age and self-employment.

The results indicate that the BQA tends to be more selective than the CBA. We find strong relations between response and auxiliary variables of which some are job-related. The results do not necessarily imply, however, that it is useless

Table 4.17: *Logistic regression model for response and nonresponse in the BQA*

Variable	*Category*	β	
		All households	*Telephone households*
Constant		-1.532**	1.302**
Telephone	yes	0.765**	
(Reference no)			
Average age	35 - 54		-1.379**
(Reference < 35)	> 55		-1.571**
Gender	All female	0.458*	
(Reference All male)	Mix	0.470*	
Job	yes	0.456**	
(Reference no)			
Self-employed	yes		-0.730*
(Reference no)			

to conduct a BQA; we may get respondents that were lacking in the regular LFS. If the BQA is selective in the opposite direction of the LFS, then the combined response may still have a composition that comes closer to that of the population.

4.3.2.2 Do we get new households in the basic-question approach?

We follow the strategy of section 4.3.1.2. First, we perform bivariate tests. Next, we construct logistic regression models. Table 4.18 contains p-values of bivariate weighted tests for independence. The first column contains p-values for the weighted test for independence between LFS response and the selected auxiliary variables. For most variables, the p-value is similar to the p-value for the LFS response in table 4.14. However, there are differences in some of the work related variables and in average age. The independence tests now reject that a CWI registration and the average age are independent of response in the BQA. Also the p-value for an unemployment allowance dropped considerably from 0.79 to 0.17. We conclude that the additional regions in the BQA lead to a stronger dependence between response behaviour and these variables. Nonetheless, for the majority of the auxiliary variables the p-values remain very low. The second column of table 4.18 contains p-values after the addition of the basic question respondents. If we compare these two sets then it follows that many of the p-values again remain small for the combined response. For the auxiliary variable

average age the p-value has increased and is now above the usual significance levels; independence between response and the average age of the household core is not rejected anymore. Furthermore, the p-value for self-employment has increased. Other than for the CBA we cannot conclude that the BQA response has become more representative when adding the re-approach response. We would still reject bivariate independence for many variables including some of the (un)employment related variables. We also computed p-values for weighted tests of independence of response group, LFS or basic-question, given that the household responded. Here, we make a distinction between all households and the listed households. p-values are given in the third and fourth column of table 4.18. We performed the test also for the LFS key variable employment status, which could be derived from the basic questions in the condensed questionnaires.

The p-values indicate that for the majority of the auxiliary variables no significant differences are found in the bivariate composition of the response of LFS and basic-question re-approach. This conclusion remains true after we isolated the listed households. This indicates that the BQA respondents resemble the LFS respondents more than did the CBA respondents. The auxiliary variables urbanization, province, household type and average age lead to a rejection of independence at the 5% level. The observed differences with respect to urbanization and province are not surprising. The LFS is a face-to-face survey and contact rates in the LFS differ considerably between regions of the country due to interviewer workload, accessibility of addresses, and physical impediments. A re-approach using a different mode can be expected to be especially beneficial for response from households that were not contacted in the LFS. Remarkably, the independence of the group of response and having a listed telephone is not rejected. The BQA strengthens the selectiveness of response for having a listed telephone. We get more households with a listed telephone while we would like to have more households without a listed telephone.

Next, and analogous to table 4.15, we constructed three logistic regression models. These are given in table 4.19. Again, '*' denotes significance at a 5% level, '**' denotes significance at a 1% level. The first model for the LFS response probability is the same as the first model in 4.15. The second model explains the response probability to the combined response of LFS and basic question re-approach. The third model describes the probability that a household responded in the LFS given that the household is a respondent in either the LFS or basic question re-approach. Additionally we derived a model for the same probability but restricted to listed households.

If we compare the model for the combined response of LFS and BQA with the model for the LFS response, then we see that ethnicity, region and house-

Table 4.18: *Bivariate test of independence between response behaviour and different groups*

Variable	Pearson's χ^2 p-value			
	LFS response	*Combined*	*LFS vs BQA*	
			All	Listed
Urbanization	< 0.009	< 0.009	0.01	0.02
Telephone	< 0.009	< 0.009	0.80	-
Region	< 0.009	< 0.009	0.10	0.25
Province	< 0.009	< 0.009	< 0.009	< 0.009
Gender	< 0.009	< 0.009	0.49	0.63
Ethnicity	< 0.009	< 0.009	0.65	0.52
House value	0.00	< 0.009	0.79	0.28
Household type	< 0.009	< 0.009	0.05	0.06
Job	0.00	< 0.009	0.66	0.72
Disability allowance	0.49	0.20	0.47	0.96
Social allowance	< 0.009	< 0.009	0.46	0.12
CWI registration	0.00	0.00	0.42	0.53
Unemployment all.	0.17	0.16	0.26	0.23
Self employed	0.37	0.60	0.12	0.11
Average age	0.03	0.17	0.01	0.01
Employment status			0.42	0.61

hold type are no longer significant while household cores with a job more often are respondents. The variables degree of urbanization, telephone, gender and receiving a social allowance all remain in the model but give stronger and more significant contributions. We still find more households where one of the core members receives social allowance. Especially, the selectiveness of response with respect to having a listed telephone has become more pronounced, as expected. In table 4.19, we can see that the strongest differences between LFS and BQA respondents are with respect to urbanization and region. Here, we do get new respondents in the BQA. As a consequence, the dependence between response and region has disappeared. The picture is different for the listed households. If we restrict the BQA to this group then apart from the more urban areas we attract more households with a social allowance as well as younger households.

4.3.3 Summary of the results

We investigated two questions; what households participate in the re-approaches and do we get new households. Furthermore we investigated whether the new households, if any, are also different with respect to their employment status. In other words, we investigated whether the LFS nonresponse is not-missing-at-random.

Table 4.19: *Logistic regression models for the response behaviour in different groups in the BQA*

Variable	*Category*	Pearson's χ^2 p-value			
		LFS response	*Combined*	*LFS vs BQA*	
				All	Listed
Constant		0.173	-0.008	1.211**	0.883**
Urbanization	strong	0.205*	0.310**	0.139	0.066
(Reference very strong)	moderate	0.151	0.135	0.265	0.396
	little	0.420**	0.487**	0.459*	0.557*
	not	0.346**	0.457**	0.425*	0.493*
Telephone	yes	0.371**	0.710**		
(Reference no)					
Ethnicity	Moroccan	-0.563*			
(Reference native)	Turkish	-0.331			
	Suriname/Antilles	0.064			
	other non-Western	-0.349**			
	other Western	-0.205*			
	mix	0.134			
Gender	all female	0.319**	0.516**		
(Reference all male)	mix	0.123	0.494**		
Social allowance	yes	-0.407**	-0.433*		-0.630*
(Reference no)					
Average age	35 - 54				0.505**
(Reference < 35)	55 +				0.556**
Having a job	yes		0.264**		
(Reference no)					
Part of country	East	0.081		0.159	
(Reference North)	West	-0.059		0.029	
	South	0.260		0.367	
Household type	not married	0.012			
(Reference single)	married	0.210			
	unmarried w/children	0.109			
	married w/children	0.366*			
	single parent	0.217			
	other	-0.459			
	> 1 household	0.447*			

A comparison of the call-back response and nonresponse revealed only small differences between those groups. We only found differences for average age of the household and part of the country where the household lives. The small difference between call-back response and nonresponse suggests that with respect to the available demographic, geographic and socio-economic characteristics, the response has become more similar to the sample. The analyses indicate that this is indeed true. A comparison of the combined response to the original LFS sample leads to smaller p-values and, hence, less significant differences.

Next, we found that the LFS and call-back respondents are indeed different. When the two groups are compared directly, they are especially different for geographic variables, having a listed land-line telephone and ethnic background. The follow-up response has an over representation of the most urbanized, western parts of the Netherlands. Also, households without a listed telephone are over represented in the follow-up. Finally, Moroccan and non-western households other than Moroccan and Turkish households are over represented in the study.

When it comes to key LFS questions, we do not find a significant difference between the job and employment status of the households in the LFS and the call-back response. This important finding holds both in a bivariate comparison as well as in a multivariate setting. It implies that there is no indication that households are not-missing-at-random in the LFS when it comes to these items.

The analysis of the call-back response indicates that call-back respondents are more similar to the remaining nonrespondents than to the LFS respondents. The selected refusals are the largest nonresponse group; they make up 77% of the re-approach sample. Not surprisingly, the follow-up response rate was highest for the former non-contacts, 58%. This finding conforms to that of Lynn et al (2002), i.e. an intensive follow-up is especially successful for the difficult to contact. However, contrary to Lynn et al. (2002), we do not find that the survey estimates of employment and unemployment are affected by the addition of the follow-up response.

The picture for the basic-question respondents is different. We believe that this difference is caused by the design of the basic-question approach. When we compare the basic-question response with the basic-question nonresponse, we find differences with respect to telephone ownership, age, having a paid job and gender. The results indicate that the basic-question respondents are more similar to the LFS respondents than to the remaining nonrespondents. There is still no difference in employment status between the two groups. However, we seem to get only new respondents with respect to region and age. The total combined response of the LFS and the basic-question approach is slightly

more representative with respect to these characteristics, but not with respect to other characteristics. We believe that the improved composition of region and age again is mostly due to the conversion of former non-contacts and unprocessed cases. Region and age relate strongly to the ability to contact households in a face-to-face survey.

4.4 Concluding remarks

Both re-approaches were successful in obtaining additional response. The response rates in the call-back and basic-question approaches respectively, were 43% and 40%. The rather passive way we re-approached unlisted households in the BQA is reflected in the response percentage: for CATI, 50% of the households participated in the survey. In the web- and paper combination, only 23% of the households responded. Possibly the response to the web survey can be increased by an improved design of the web questionnaire, for instance by using the Total Design Method (TDM) suggested by Dillman (2000).

The low response rate in the paper and web group of the basic-question approach is attenuated by the types of LFS nonrespondents assigned to this strategy. Non-contacts more often do not have a listed telephone than refusals. It is reasonable to expect lower conversion rates for refusals, especially when we put relatively little effort into converting them. Another possible explanation for the low response rates in the paper and web group is raised by Griffin et al. (2001), who found that offering multiple modes of response in a mailing led to a lower response rate. One may consider to use only the paper questionnaires to avoid a possible break in the response process.

The results for the basic-question approach as a whole are not satisfactory. The analyses indicate that the implemented basic-question approach was, however, more successful when restricted to households with a listed telephone. In future research on the basic-question approach, therefore, the design should be changed to get also the unlisted households into the response. One may either consider to use the combination of paper and web modes for the entire sample, as they are cheap, and to send a number of reminders or to include a pre-paid incentive. One may also consider employing face-to-face interviewers carrying paper questionnaires to make a quality cost trade-off.

In both re-approaches we have to be careful when drawing conclusions about the employment status of the non-converted nonrespondents. While there is an indication that they resemble converted nonrespondents in socio-economic characteristics in both cases, the low conversion rate of refusals leaves ample room for

differences in employment status. However, this study gives no indication that the nonresponse of the LFS may follow a not-missing-at-random data pattern for this key item.

Chapter 5

The R-Indicator As a Supplement to the Response Rate

In Chapter 1 we showed that nonresponse has two main consequences for survey estimates. First, it reduces the sample size, i.e. it decreases the precision of the estimates. Second, it deteriorates the data since the inclusion probabilities that were chosen in the design no longer hold. As a consequence, nonresponse may also introduce bias to the estimates. The decreased precision can be dealt with by an increased sample size. Without any auxiliary information about the sample units, not much can be done, however, about the nonresponse bias. One needs to assume that on average, nonrespondents are the same as respondents with respect to the key survey topics, but this is not always a realistic assumption. In case auxiliary information from administrative sources can be linked to the sample or in case good auxiliary population statistics are available, the corresponding auxiliary variables can be used to calibrate the response to sample or population totals as we describe in Chapter 6. In this chapter[1] we show how the auxiliary information can be employed to measure the impact of the nonresponse.

[1]This chapter is based on Schouten and Cobben (2007) and Cobben and Schouten (2008). A revised version of this chapter has been accepted for publication in *Survey Methodology* as "Indicators for the representativeness of survey response" authored by Schouten, Cobben, and Bethlehem.

5.1 Introduction

It is a well-developed finding in the survey methodological literature that response rates by themselves are poor indicators of nonresponse bias, see e.g. Curtin, Presser and Singer (2000), Groves and Heeringa (2006), Groves, Presser and Dipko (2004), Groves and Peytcheva (2006), Keeter et al. (2000), Merkle and Edelman (2002), Heerwegh et al. (2007) and Schouten (2004). However, the field has yet to propose alternative indicators of nonresponse bias that are less ambiguous indicators of survey quality.

The survey literature contains an ever growing body of analyses into nonresponse bias. For general overviews see Groves et al. (2002), Stoop (2005), Groves (2006) and Groves and Peytcheva (2006). These analyses make use of auxiliary population or sample totals of demographic and socio-economic characteristics of households. Although response behaviour depends on the topic of the survey, a number of characteristics has been identified that relate to lower response rates, see Chapter 3. Age, type of household and degree of urbanization usually have a different composition in the response than in the original sample.

With the analysis of nonresponse came the concept of a continuum of resistance, see Chapter 2. Two of these dimensions are ease of contact and ease of participation. Attached to those dimensions are individual contact and participation probabilities, and when combined overall individual response probabilities. Clearly, these probabilities are unknown but can be modelled using the available auxiliary information. Associated with the continuum of resistance is the level of effort of the survey researcher. The more effort the survey researcher invests in contacting households and converting reluctant respondents, the higher the response rate. It seems that the level of effort has increased during the past decades in order to maintain acceptable response rates. The level of effort represents costs (and time) and can be balanced to response rates, see Kalsbeek et al. (1994). One may also attempt to differentiate the level of effort between households to get a balanced composition of the response, see Groves and Heeringa (2006), Biemer and Link (2006) and Van der Grijn, Schouten and Cobben (2006).

The question arises whether increased efforts, apart from a higher response rate, lead to a survey that is more 'representative' of the population under study. This has been investigated by e.g. Lynn et al. (2002), Stoop (2005) and Schouten and Bethlehem (2007). But exactly what do we mean by representative and how can we measure this concept? In this chapter, we focus attention on these questions. We propose an indicator, which we call an R-indicator ('R' for representativity), for the similarity between the response to a survey and the

sample or the population under investigation.

In the literature there are many different interpretations of the concept of representativeness, see Kruskal and Mosteller (1979 a, b and c) for a thorough investigation of the statistical and non-statistical literature. Some authors explicitly define representativeness. Hajèk (1981) links 'representative' to the estimation of population parameters; the pair formed by an estimator and a missing-data-mechanism are representative in case with probability one the estimator is equal to the population parameter. Following Hajèk's definition, calibration estimators (e.g. Särndal et al., 1992) are representative for the auxiliary variables that are calibrated. Bertino (2006) defines a so-called univariate representativeness index for continuous random variables. This index is a distribution free measure based on the Cramér-Von Mises statistic. Kohler (2007) defines an internal criterion for representativeness. This univariate criterion resembles the Z-statistic for population means.

We disconnect the concept representativeness from the estimation of a specific population parameter but relate the concept to the overall composition of response. By disconnecting indicators from a specific parameter they can be used as tools for comparing different surveys and surveys over time, and for a comparison of different data collection strategies and modes. Also, the measure gives a multivariate perspective of the dissimilarity between sample and response.

The R-indicator that we propose (Schouten and Cobben, 2007) employs estimated response probabilities. The estimation of response probabilities implies that the R-indicator itself is a random variable, and, consequently, has a precision and possibly a bias. The sample size of a survey, therefore, plays an important role in the assessment of the R-indicator. However, this dependence exists for any measure; small surveys simply do not allow for strong conclusions about the missing-data-mechanism. We show that the proposed R-indicator relates to Cramér's V measure for the association between response and auxiliary variables. In fact, we view the R-indicator as a lack of association measure. The weaker the association the better, as this implies there is no evidence that nonresponse has affected the composition of the observed data. To use R-indicators as tools for monitoring and comparing survey quality in the future, they need to have the features of a measure. That is, we want an R-indicator to be interpretable, measurable and normalizable. Furthermore, the R-indicator has to satisfy the mathematical properties of a measure. Especially the interpretation and normalization of the R-indicator are not straightforward.

We apply the R-indicator to the re-approach study of the nonrespondents to the Dutch LFS, described in Chapter 4. We compare the values of the R-

indicator with the conclusions from that study as an empirical evaluation of the R-indicator. Additionally, we apply the R-indicator to two other studies that were conducted at Statistics Netherlands in 2006. These studies involved different data collection modes. A detailed analysis was done and documented, which make them suitable for an empirical evaluation of the R-indicator as well. We refer to Schouten and Cobben (2007) and Cobben and Schouten (2008) for more illustrations and empirical investigations.

In section 5.2, we start with a discussion of the concept of representative response. Next, in section 5.3, we define the R-indicator. Section 5.4 is devoted to the features of the R-indicator. Sections 5.5.2 and 5.5.3 describe the application of the R-indicator to field studies. Finally, section 5.6 ends this chapter with concluding remarks.

5.2 The concept of representative response

In this section we first discuss what it means that a subsample of respondents is representative of the whole sample. Next, we make the concept of representative response mathematically rigorous by giving a definition.

5.2.1 The meaning of representative

Why we should not single-mindedly focus on response rates as an indicator of survey quality can be illustrated by an example from the 1998 Dutch survey POLS (short for Permanent Onderzoek Leefsituatie or Integrated Survey on Household Living Conditions in English).

Example 5.2.1 'More' is not necessarily 'better'
Table 5.1 contains the one and two month POLS survey estimates for the percentage of the Dutch population that receives some form of social allowance and the percentage non-natives. Both variables are taken from administrative data and are artificially treated as survey questions. The sample proportions are also given in table 5.1. After one month the response rate was 47.2%, while after the full period of interview of two months the rate was 59.7%. In the 1998 POLS the first month was CAPI. Nonrespondents after the first month were allocated to CATI if they had a listed land-line telephone. Otherwise, they were allocated once more to CAPI. The second month of interview gave another 12.5% of response. However, from table 5.1 we can see that after the second month the survey estimates have a larger bias than after the first month. The increased ef-

Table 5.1: *Response means in POLS 1998 after one month of interview and for the full fieldwork period*

Variable	*After 1 month*	*After 2 months*	*Sample*
Social allowance	10.5%	10.4%	12.1%
Non-native	12.9%	12.5%	15.0%
Response rate	47.2%	59.7%	100%

fort led to a less representative response with respect to both auxiliary variables. ■

In this specific example, it is clear that the response became less representative with respect to the two auxiliary variables. But what do we mean by representative in general? The term representative is often used with hesitation in the statistical literature. Kruskal and Mosteller (1979 a, b and c) make an extensive inventory of the use of the word 'representative' in the literature and identify nine interpretations. The statistical interpretations that Kruskal and Mosteller named 'absence of selective forces', 'miniature of the population', and 'typical or ideal cases' relate to probability sampling, quota sampling and purposive sampling. In section 5.2.2 we propose a definition that corresponds to the interpretation of the 'absence of selective forces'. First, we motivate why we make this choice.

The concept of representative response is also closely related to the missing data mechanisms missing-completely-at-random (MCAR), missing-at-random (MAR) and not-missing-at-random (NMAR), see Chapter 1. These mechanisms find their origin in model-based statistical theory. Somewhat loosely translated, with respect to a survey item, MCAR means that respondents are on average the same as nonrespondents, MAR means that within known sub populations respondents are on average the same as nonrespondents, while NMAR means that the respondents are different from the nonrespondents. The addition of the survey item is essential. Within one questionnaire some survey items can be MCAR, while other survey items are MAR or NMAR. Furthermore, the MAR assumption for one survey item holds for a particular stratification of the population. A different survey item may need a different stratification.

We wish to monitor and compare the response to different surveys in topic or time. As a consequence, it is not interesting to define a representative response dependent on the survey items itself or dependent on the estimator used. Instead, we focus on the quality of data collection. We, therefore, compare the

response composition with that of the sample. The survey topic influences the probability that households participate in the survey, but this influence cannot be measured or tested. Hence, from our perspective, this influence cannot be used to assess the quality of the response. We propose to judge the composition of the response by pre-defined sets of variables that are observed external to the survey and can be employed for each survey under investigation. We want the respondent selection to be as close as possible to a 'simple random sample of the survey sample', i.e. with as little relation as possible between response and characteristics that distinguish sample elements from eachother. This can be interpreted as the absence of selective forces in the respondent selection, or as MCAR with respect to all possible survey items.

5.2.2 Definition of a representative subsample of respondents

Let the population U consist of N elements i, $i = 1, \ldots, N$. By δ_i we denote the sample indicator which is 1 in case element i is sampled and 0 otherwise. By R_i we denote the response indicator for i. If element i is sampled and did respond then $R_i = 1$, otherwise $R_i = 0$. The sample size is $n = \sum_{i=1}^{N} \delta_i R_i + \sum_{i=1}^{N} \delta_i (1 - R_i) = n_r + n_{nr}$. Finally, π_i denotes the first-order inclusion probability of sample element i.

The key to our definitions lies in the individual response propensities. Let ρ_i be the probability that element i responds in case it is sampled, i.e. $\rho_i = P(R_i = 1 | \delta_i = 1)$. We follow a model-assisted approach, i.e. the only randomness is in the selection of the sample and the response to the survey. A response probability is a feature of a labelled and identifiable element, so to say a biased coin that the element carries in a pocket, and is therefore inseparable from that element. First, we give a strong definition.

Definition 1 (Strong)**.** *A subsample of respondents is strongly representative with respect to the sample if the response probabilities ρ_i are the same for all elements in the population*

$$\rho_i = P(R_i = 1 | \delta_i = 1) = \rho \tag{5.1}$$

for $i = 1, 2, \ldots, N$, and if the response of an element is independent of the response of all other elements.

The strong definition implies that the response is MCAR with respect to all possible survey questions. Although this definition is appealing, its validity

cannot be tested in practice since we have no replicates of the response of one single element. Therefore, we also construct a weak definition that can be tested in practice.

Definition 2 (Weak)**.** *A subsample of respondents is weakly representative for a categorical variable X with H categories if the average response probability is constant over categories*

$$\overline{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho \tag{5.2}$$

for $h = 1, \ldots, H$, where N_h is the population size of category h, ρ_{hk} is the response probability of element k in class h and summation is over all elements in this category.

The weak definition corresponds to a missing-data-mechanism that is MAR, or MCAR with respect to X as MCAR states that we cannot distinguish respondents from nonrespondents based on knowledge of X.

5.3 R-indicators

In the previous section we defined strong and weak representative response. Both definitions make use of individual response probabilities that are unknown in practice. In this section, we propose a representativity indicator, or R-indicator. First, we start with a population R-indicator. From there on we base the same R-indicator on a sample and on estimated response probabilities, or response propensities.

5.3.1 Population based R-indicator

Let us consider the hypothetical situation in which the individual response probabilities are known. In this situation, we can test the strong definition. We would like to measure the amount of variation in the response probabilities; the more variation, the less representative in the strong sense. Let $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_N)'$ be a vector of response probabilities, let $\mathbf{1} = (1, 1, \ldots, 1)'$ be the N-vector of ones, and let $\boldsymbol{\rho}_0 = \mathbf{1} \times \overline{\rho}$ be the vector consisting of the average population response probability $\overline{\rho} = \frac{1}{N} \sum_{i=1}^{N} \rho_i$.

Any distance function d in $[0,1]^N$ would suffice to measure the deviation from a strong representative response by calculation of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_0)$. Note that the

height of the overall response does not play a role. When we apply the Euclidean distance function to a distance between $\boldsymbol{\rho}$ and $\boldsymbol{\rho}_0$, this measure is proportional to the standard deviation of the response probabilities

$$S(\boldsymbol{\rho}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\rho_i - \overline{\rho})^2} \tag{5.3}$$

When fixing the average response probability $\overline{\rho}$, the maximum variation is obtained by letting $\overline{\rho}N$ of the response probabilities be equal to 1, and $(1-\overline{\rho})N$ of the response probabilities be equal to 0. Then, $S(\boldsymbol{\rho}) \leq \sqrt{\overline{\rho}(1-\overline{\rho})}$. In addition, by taking $\overline{\rho} = \frac{1}{2}$ it follows that

$$S(\boldsymbol{\rho}) \leq \sqrt{\overline{\rho}(1-\overline{\rho})} \leq \frac{1}{2} \tag{5.4}$$

We want the R-indicator to take values on the interval $[0, 1]$ with the value 1 being strong representativeness and the value 0 being the maximum deviation from strong representativeness. Therefore we propose the R-indicator $R(\boldsymbol{\rho})$ defined by

$$R(\boldsymbol{\rho}) = 1 - 2S(\boldsymbol{\rho}) \tag{5.5}$$

Note that the minimum value of $R(\boldsymbol{\rho})$ depends on the response rate, see figure 5.1. For $\overline{\rho} = \frac{1}{2}$ it has a minimum value of zero. For $\overline{\rho} = 0$ and $\overline{\rho} = 1$ no variation is possible and the minimum value is one. Paradoxically, the lower bound increases when the response rate decreases from 0.5 to 0. For a low response rate there is less room for individual response probabilities to vary.

The R-indicator $R(\boldsymbol{\rho})$ can be seen as a *lack of* association measure. When $R(\boldsymbol{\rho}) = 1$ there is no relation between any auxiliary variable and the missing-data-mechanism. In fact, $R(\boldsymbol{\rho})$ has a close relation to the well-known χ^2-statistic that is often used to test independence and goodness-of-fit. Suppose that the response probabilities are only different for classes h defined by a categorical variable X. Let $\overline{\rho}_h$ be the average response probability for class h, and let f_h be the population fraction of class h, i.e. $f_h = \frac{N_h}{N}$ for $h = 1, 2, \ldots, H$. In other words, for all i with $X_i = h$ the response probability is $\rho_i = \overline{\rho}_h$. Since the variance of the response probabilities is the sum of the 'between' and 'within' variances over classes h, and the within variances are assumed to be zero, it holds that

$$S^2(\boldsymbol{\rho}) = \frac{1}{N-1}\sum_{h=1}^{H} N_h(\overline{\rho}_h - \overline{\rho})^2 = \frac{N}{N-1}\sum_{h=1}^{H} f_h(\overline{\rho}_h - \overline{\rho})^2 \approx \sum_{h=1}^{H} f_h(\overline{\rho}_h - \overline{\rho})^2 \tag{5.6}$$

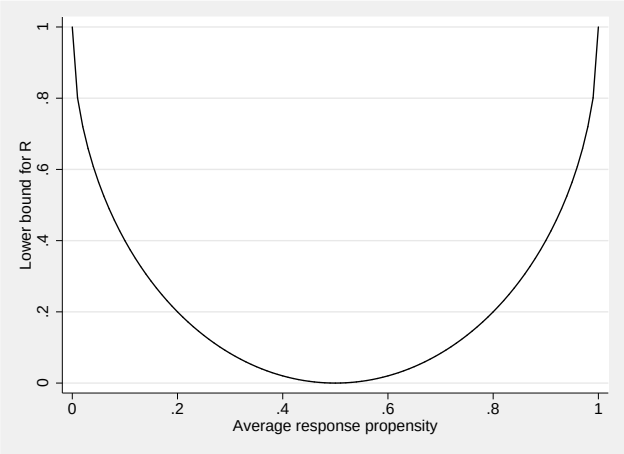

Figure 5.1: *Minimum value of $R(\boldsymbol{\rho})$ as a function of the average response probability*

The χ^2-statistic measures the distance between observed and expected proportions. However, it is only a true distance function in the mathematical sense for fixed marginal distributions f_h and $\overline{\rho}$. We can apply the χ^2-statistic to X in order to 'measure' the distance between the true response behaviour and the response behaviour that is expected when response is independent of X, i.e. to measure the deviation from weak representativeness with respect to X. Rewriting the χ^2-statistic leads to

$$\begin{aligned}
\chi^2 &= \sum_{h=1}^{H} \frac{(N_h\overline{\rho}_h - N_h\overline{\rho})^2}{N_h\overline{\rho}} + \sum_{h=1}^{H} \frac{(N_h(1-\overline{\rho}_h) - N_h(1-\overline{\rho}))^2}{N_h(1-\overline{\rho})} && (5.7)\\
&= \sum_{h=1}^{H} \frac{N \times f_h(\overline{\rho}_h - \overline{\rho})^2}{\overline{\rho}} + \sum_{h=1}^{H} \frac{N \times f_h(\overline{\rho}_h - \overline{\rho})^2}{(1-\overline{\rho})} && (5.8)\\
&= \frac{N}{\overline{\rho}(1-\overline{\rho})} \sum_{h=1}^{H} f_h(\overline{\rho}_h - \overline{\rho})^2 && (5.9)\\
&\approx \frac{N}{\overline{\rho}(1-\overline{\rho})} S^2(\boldsymbol{\rho}) && (5.10)
\end{aligned}$$

The χ^2-statistic is thus related to the R-indicator $R(\boldsymbol{\rho})$ by

$$\chi^2 \approx \frac{N}{4\overline{\rho}(1-\overline{\rho})}(1 - R(\boldsymbol{\rho})^2) \tag{5.11}$$

An association measure that transforms the χ^2-statistic to the $[0, 1]$ interval is Cramér's V, see equation (3.1). Cramér's V attains a value zero if observed proportions exactly match expected proportions and its maximum is one. In our case the denominator equals N, since the response indicator only has two categories, response and nonresponse. As a consequence, (3.1) becomes

$$V = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{1}{\overline{\rho}(1-\overline{\rho})}} S(\boldsymbol{\rho}) \tag{5.12}$$

From (5.12) we can see that for large N Cramér's V is approximately equal to the standard deviation of the response probabilities, standardized by the maximum standard deviation $\sqrt{\overline{\rho}(1-\overline{\rho})}$ for a fixed average response probability $\overline{\rho}$. However, unless we fix $\overline{\rho}$, (5.12) is not a distance function. Because we do not want the R-indicator to depend on the average response probability, we continue with the R-indicator defined by (5.5).

5.3.2 Sample based R-indicator

In section 5.3.1 we assume that we know the individual response probabilities. In practice, however, these probabilities are unknown. Furthermore, we only have information about the response behaviour of sampled elements. Therefore, we are interested in alternatives to the indicator $R(\boldsymbol{\rho})$. An obvious way to do this, is to use response-based estimators for the individual response probabilities and the average response probability. Therefore, we swith to the estimated response probabilities, or response propensities $\hat{\rho}_i$, introduced in Chapter 1. Methods to estimate response probabilities, for instance logistic or probit regression models, are discussed in appendix 5.A. By $\hat{\overline{\rho}}$ we denote the weighted sample average of the response propensities, i.e.

$$\hat{\overline{\rho}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\delta_i}{\pi_i}\hat{\rho}_i \tag{5.13}$$

We replace $R(\boldsymbol{\rho})$ by the estimator $\hat{R}(\hat{\boldsymbol{\rho}})$

$$\hat{R}(\hat{\boldsymbol{\rho}}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \frac{\delta_i}{\pi_i}(\hat{\rho}_i - \hat{\overline{\rho}})^2} \tag{5.14}$$

Note that in (5.14) there are in fact two estimation steps based on different probability mechanisms. The response probabilities $\boldsymbol{\rho}$ are estimated and the variation in the probabilities $S^2(\boldsymbol{\rho})$ is estimated. Hence, we denote the sample-based R-indicator by $\hat{R}(\hat{\boldsymbol{\rho}})$. We return to the consequences of these two estimation steps in section 5.4.

Example 5.2.1 'More' is not necessarily 'better' - continued
The proposed R-indicator is applied to the survey data from the 1998 POLS. This survey was a combination of CAPI and CATI, in which the first month was CAPI only. The sample size was close to $40,000$ and the response rate was approximately 60%. By linking the fieldwork administration to the sample it could be deduced for each contact attempt whether it resulted in a response. This enables monitoring the traces of the R-indicator during the fieldwork period. For the estimation of response probabilities we used a logistic regression model with region, ethnic background and age as independent variables. Region was a classification with 16 categories, the 12 provinces and the 4 largest cities Amsterdam, Rotterdam, The Hague and Utrecht as separate categories. Ethnic background has seven categories: native, Moroccan, Turkish, Suriname, Netherlands Antilles, other non-western non-native and other western non-native. The variable age has three categories: 0–34 years, 35-54 years, and 55 years and older, see Chapter 3 for a detailed description of these variables. In figure 5.2 $\hat{R}(\hat{\boldsymbol{\rho}})$ is plotted against the response rate for the first six contact attempts in POLS. The leftmost value corresponds to the subsample of respondents after one attempt was made. For each additional attempt the response rate increases, but the indicator shows a drop in representativeness. Hence, the subsample of respondents becomes less representative with respect to region, ethnic background and age. This result confirms the findings in Schouten (2004). ■

5.4 Features of an R-indicator

In section 5.3 we proposed a candidate indicator for representativeness. However, other indicators can also be constructed. There are many association measures or fit indexes, e.g. Goodman and Kruskal (1979), Bentler (1990) and Marsh et al. (1988). Association measures have a strong relation to R-indicators. Essentially, an R-indicator attempts to measure the lack of association in a multivariate setting. In this section we discuss the desired features of an R-indicator. We show that the proposed R-indicator $\hat{R}(\hat{\boldsymbol{\rho}})$ in (5.14) allows for a straightforward upper bound on the nonresponse bias.

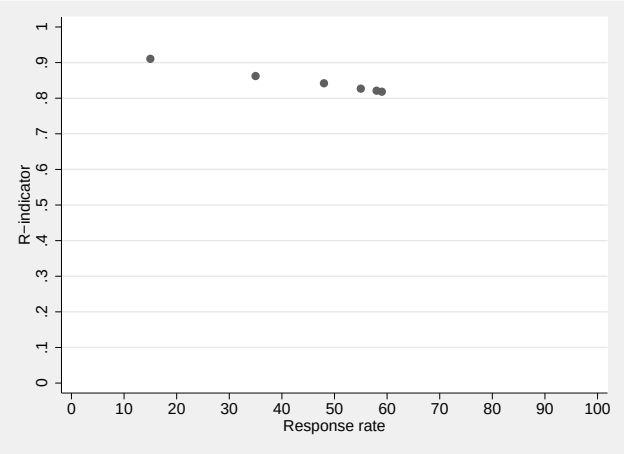

Figure 5.2: *Sample based R-indicator for the first six contact attempts in POLS 1998*

5.4.1 Features in general

The triangle inequality property of a distance function allows for a partial ordering of the variation in response propensities. A distance function can easily be derived from any mathematical norm. In section 5.2 we chose to use the Euclidean norm as this norm is commonly used. The Euclidean norm led to an R-indicator that uses the standard deviation of response propensities. In section 5.4.3 we show that the Euclidean norm based R-indicator has interesting normalization features.

We must make a subtle distinction between R-indicators and distance functions. Distance functions are symmetric while an R-indicator measures a deviation with respect to a specific point, namely the situation where all response propensities are equal. If we change the vector of individual propensities, then in most cases this point is shifted. However, if we fix the average response propensity then the distance function facilitates interpretation.

The desirable features of an R-indicator are: measurement, interpretation and normalization. In section 5.3.1 we derived the 'population' R-indicator (5.5) based on response probabilities. However, such an indicator is not measurable when response probabilities are unknown and all we have is the response to one survey. Hence, in section 5.3.2 we made the R-indicator measurable by switching

to response propensities, which led to the sample based R-indicator (5.14). In the next two sections we discuss the interpretation and normalization of the sample based R-indicator $\hat{R}(\hat{\boldsymbol{\rho}})$.

5.4.2 Interpretation

The second desirable feature of an R-indicator is the ease with which we can interpret its value and the concept it is measuring. The R-indicator $\hat{R}(\hat{\boldsymbol{\rho}})$ is based on a sample and on estimated individual response probabilities. Both have far-reaching consequences for the interpretation and comparison of the R-indicator. Since the R-indicator is an estimator itself, it is also a random variable. This means that it depends on the sample, i.e. it is potentially biased and has a certain accuracy. But what is it estimating?

Let us first assume that the sample size is arbitrarily large so that precision does not play a role. We also assume here that the selection of a model for response propensities is no issue. In other words, we are able to fit any model for any fixed set of auxiliary variables. There is a strong relation between the R-indicator and the availability and use of auxiliary variables. In section 5.2 we defined strong and weak representativeness. Even when we are able to fit any model, we are not able to estimate response propensities beyond the 'resolution' of the available auxiliary variables. Therefore, we can only draw conclusions about weak representativeness with respect to the set of auxiliary variables. This implies that whenever an R-indicator is used, it is necessary to complement its value by the set of covariates that served as a grid to estimate individual response propensities. If the R-indicator is used for comparative purposes, then those sets must be the same. We must add that it is not necessary that all auxiliary variables are indeed used for the estimation of propensities, since they may not add any explanatory power to the model. However, the same sets should be available. The R-indicator $\hat{R}(\hat{\boldsymbol{\rho}})$ then measures a deviation from weak representativeness.

The R-indicator does not capture differences in response probabilities within subgroups of the population other than the subgroups defined by the classes of X. If we let again $h = 1, 2, \ldots, H$ denote strata defined by X, N_h be the size of stratum h, and $\overline{\rho}_h$ be the population average of the response probabilities in stratum h, then it is not difficult to show that $\hat{R}(\hat{\boldsymbol{\rho}})$ is a consistent estimator of

$$R_X(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{h=1}^{H} N_h(\overline{\rho}_h - \overline{\rho})^2} \tag{5.15}$$

in case standard models like logistic regression or linear regression are used to estimate the response probabilities. Of course, (5.15) and (5.5) may be different.

In practice, the assumption of an arbitrarily large sample size is not realistic. The sample size affects both estimation steps; the estimation of response propensities and the estimation of the R-indicator using a sample. If we would know the individual response probabilities, the sample based estimation of the R-indicator would only lead to variance and not to bias. Hence, for a small sample size the estimated sample based R-indicator would have a small precision but this could be accounted for by using a confidence interval instead of a point estimator. Because of model selection and model fit, the implications for the estimation of response probabilities are different. There are two strategies. Either a model is imposed to estimate probabilities, thereby fixing the auxiliary variables beforehand, or the model is constructed using only the significant contribution of auxiliary variables with respect to some predefined level.

In the first case, again no bias is introduced but the standard error may be affected by overfitting. In the second case, the model for the estimation of response probabilities depends on the size of the sample: the larger the sample, the more interactions are accepted as significant. Although it is standard statistical practice to fit models based on a significance level, model selection may introduce bias and variance when estimating an R-indicator. This can be easily understood by regarding the extreme case of a sample of, say, size 10. For such a small sample no interaction between response behaviour and auxiliary variables will be accepted, leaving an empty model and an estimated R-indicator of 1. Small samples simply do not allow for the estimation of response probabilities. In general, a smaller sample size will thus lead to a more optimistic view on representativeness.

We should make a further subtle distinction. It is possible that for one survey a lot of interactions contribute to the prediction of response probabilities but only very little each, while in another survey there is only one but strong contribution of a single interaction. None of the small contributions may be significant, but together they are as strong as the one large contribution that is significant. Hence, we would be more optimistic in the first example even if sample sizes would be comparable. These observations show that one should always use an R-indicator with care. It cannot be viewed separately from the auxiliary variables that were used to compute it. Furthermore, the sample size has an impact on both bias and precision.

5.4.3 Normalization

The third important feature of the R-indicator is normalizability. We would like to determine bounds for the R-indicator so that the scale of the R-indicator, and as a consequence also changes in the R-indicator, get a meaning. Clearly, the interpretation issues that we raised in the previous section also affect the normalization of the R-indicator. Therefore, in this section we assume the ideal situation that we can estimate response probabilities without bias. This assumption holds for large surveys. We discuss the normalization of the R-indicator $\hat{R}(\hat{\boldsymbol{\rho}})$.

5.4.3.1 Upper bounds for the absolute bias and the mean square error

We show that for any survey item Y the R-indicator can be used to determine upper bounds for the nonresponse bias and for the root mean square error (RMSE) of adjusted response means. Let Y be a survey item and let $\overline{y}_{ht}^r$ be the Horvitz-Thompson estimator for the population mean based on the survey response, i.e.

$$\overline{y}_{ht}^r = \frac{\sum_{i=1}^N \frac{R_i}{\pi_i} Y_i}{\sum_{i=1}^N \frac{R_i}{\pi_i}} \tag{5.16}$$

Bethlehem (1988) refers to (5.16) as the modified Horvitz-Thompson estimator. It can be shown (e.g. Bethlehem, 1988, Särndal and Lundström, 2005) that its bias $B(\hat{\overline{y}}_{ht}^r)$ is approximately equal to

$$B(\overline{y}_{ht}^r) = \frac{Cov(\mathbf{Y}, \boldsymbol{\rho})}{\overline{\rho}} \tag{5.17}$$

with $Cov(\mathbf{Y}, \boldsymbol{\rho}) = \frac{1}{N}\sum_{i=1}^N (Y_i - \overline{y})(\rho_i - \overline{\rho})$ the population covariance between the survey item and the response probabilities. Bethlehem (1988) gives an approximation of the variance $S^2(\overline{y}_{ht}^r)$ of $\overline{y}_{ht}^r$

$$S^2(\overline{y}_{ht}^r) = \frac{1}{(N\overline{\rho})^2} \sum_{k=1}^N \sum_{l=1}^N (\rho_{kl}\pi_{kl} - \rho_k\rho_l\pi_k\pi_l) \frac{(Y_k - E(\overline{y}_{ht}^r))(Y_l - E(\overline{y}_{ht}^r))}{\pi_k\pi_l} \tag{5.18}$$

where $E(\overline{y}_{ht}^r) = \frac{1}{N}\sum_{i=1}^N \frac{\rho_i Y_i}{\overline{\rho}}$, $\rho_{kl} = \rho_k\rho_l$ for $k \neq l$ and $\rho_{kk} = \rho_k$.

A normalization of $\hat{R}(\hat{\boldsymbol{\rho}})$ is obtained by using the Cauchy-Schwarz inequality. From this inequality it follows that the covariance between any two random

variables is bounded in absolute sense by the product of the standard deviations of the two random variables. We can translate this to bounds for the bias (5.17) of $\overline{y}_{ht}^r$ as

$$|B(\overline{y}_{ht}^r)| \leq \frac{S(\boldsymbol{\rho})S(\mathbf{Y})}{\overline{\rho}} = \frac{(1-R(\boldsymbol{\rho}))S(\mathbf{Y})}{2\overline{\rho}} = B_m(\boldsymbol{\rho}, \mathbf{Y}) \tag{5.19}$$

Clearly, we do not know the upper bound $B_m(\boldsymbol{\rho}, \mathbf{Y})$ in (5.19) but we can estimate it using the sample and the response propensities

$$\hat{B}_m(\hat{\boldsymbol{\rho}}, \mathbf{Y}) = \frac{(1-\hat{R}(\hat{\boldsymbol{\rho}}))\hat{S}(\mathbf{Y})}{2\hat{\overline{\rho}}} \tag{5.20}$$

where $\hat{S}(\mathbf{Y})$ is the response-based estimator of $S(\mathbf{Y})$ adjusted for the sampling design.

In a similar way we can set an upper bound to the RMSE of $\overline{y}_{ht}^r$. It holds approximately that

$$\begin{aligned} RMSE(\overline{y}_{ht}^r) &= \sqrt{B^2(\overline{y}_{ht}^r) + S^2(\overline{y}_{ht}^r)} && (5.21) \\ &\leq \sqrt{B_m^2(\boldsymbol{\rho}, \mathbf{Y}) + S^2(\overline{y}_{ht}^r)} = E_m(\boldsymbol{\rho}, \mathbf{Y}) && (5.22) \end{aligned}$$

Again, we do not know $E_m(\boldsymbol{\rho}, \mathbf{Y})$. Instead, we use a sample-based estimator that employs the response propensities

$$\hat{E}_m(\hat{\boldsymbol{\rho}}, \mathbf{Y}) = \sqrt{\hat{B}_m^2(\hat{\boldsymbol{\rho}}, \mathbf{Y}) + \hat{S}^2(\overline{y}_{ht}^r)} \tag{5.23}$$

with $\hat{S}^2(\overline{y}_{ht}^r)$ an estimator for $S^2(\overline{y}_{ht}^r)$.

The bounds $\hat{B}_m(\hat{\boldsymbol{\rho}}, \mathbf{Y})$ and $\hat{E}_m(\hat{\boldsymbol{\rho}}, \mathbf{Y})$ in (5.20) and (5.23) are different for each survey item Y. For reasons of comparison it is convenient to define a hypothetical survey item. The maximum variance of a $0-1$ variable is 0.5. Therefore, we assume that $\hat{S}(\mathbf{Y}) = 0.5$. The corresponding bounds are denoted by $\hat{B}_m(\hat{\boldsymbol{\rho}})$ and $\hat{E}_m(\hat{\boldsymbol{\rho}})$. They are equal to

$$\hat{B}_m(\hat{\boldsymbol{\rho}}) = \frac{1-\hat{R}(\hat{\boldsymbol{\rho}})}{4\hat{\overline{\rho}}} \tag{5.24}$$

$$\hat{E}_m(\hat{\boldsymbol{\rho}}) = \sqrt{\hat{B}_m^2(\hat{\boldsymbol{\rho}}) + \hat{S}^2(\overline{y}_{ht}^r)} \tag{5.25}$$

We compute (5.24) and (5.25) for all studies described in section 5.5. Note that (5.20), (5.23), (5.24) and (5.25) are again random variables that have a certain precision and that are potentially biased.

5.4.4 Response-representativeness functions

In the previous section we used the R-indicator to set upper bounds to the nonresponse bias and to the root mean square error of the (adjusted) response mean. In the opposite case we may set a lower bound to the R-indicator by demanding that either the absolute nonresponse bias or the root mean square error is smaller than some prescribed value ψ. Such a lower bound may be chosen as one of the ingredients of quality restrictions put upon the survey data by the user of the survey. If the user does not want the nonresponse bias or root mean square error to exceed a certain value, then the R-indicator must have a value above the corresponding bound.

Clearly, lower bounds to the R-indicator depend on the survey item. Therefore, again we restrict ourselves to a hypothetical survey item for which $\hat{S}(\mathbf{Y}) = 0.5$. It is not difficult to show from (5.24) that if we assume that

$$\hat{B}_m(\hat{\boldsymbol{\rho}}) \leq \psi \tag{5.26}$$

than it must hold that

$$\hat{R}(\hat{\boldsymbol{\rho}}) \geq 1 - 4\hat{\bar{\rho}}\psi = r_1(\psi, \hat{\boldsymbol{\rho}}) \tag{5.27}$$

Analogously, using (5.25) and assuming that

$$\hat{E}_m(\hat{\boldsymbol{\rho}}) \leq \psi \tag{5.28}$$

we arrive at

$$\hat{R}(\hat{\boldsymbol{\rho}}) \geq 1 - 4\hat{\bar{\rho}}\sqrt{\psi^2 - \hat{S}^2(\overline{y}^r_{ht})} = r_2(\psi, \hat{\boldsymbol{\rho}}) \tag{5.29}$$

In (5.27) and (5.29) $r_1(\psi, \hat{\boldsymbol{\rho}})$ and $r_2(\psi, \hat{\boldsymbol{\rho}})$ denote lower bounds to the R-indicator. In the following, we refer to $r_1(\psi, \hat{\boldsymbol{\rho}})$ and $r_2(\psi, \hat{\boldsymbol{\rho}})$ as response-representativeness functions. We compute them for the studies in section 5.5.

Example 5.2.1 'More' is not necessarily 'better' - continued
We again illustrate the normalization by the same example used in sections 5.2 and 5.3. Figure 5.3 contains response-representativeness function $r_1(\psi, \hat{\boldsymbol{\rho}})$ and the R-indicators $\hat{R}(\hat{\boldsymbol{\rho}})$ for the six contact attempts in POLS 1998. Three values of ψ are chosen, $\psi = 0.1$, $\psi = 0.075$ and $\psi = 0.05$. Figure 5.3 indicates that after the second contact attempt, the values of the R-indicator exceed the lower bound corresponding to the 10%-level ($\psi = 0.1$). Hence, after two contact attempts the subsample of respondents satisfies the restriction of an upper bound of 10% for

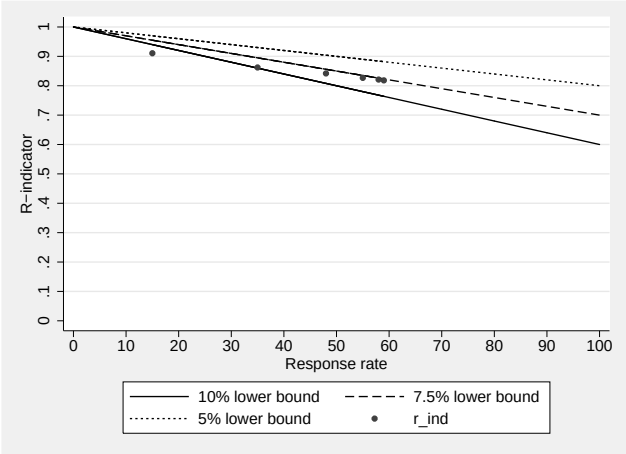

Figure 5.3: *Response-representativeness functions* $r_1(\psi, \hat{\boldsymbol{\rho}})$ *for POLS 1998*

the nonresponse bias. After four attempts the R-indicator is close to the 7.5%-level ($\psi = 0.075$). However, the values never exceed the lower bound that is based on the 5%-level ($\psi = 0.05$).

In figure 5.4 the upper bound to the absolute bias $\hat{B}_m(\hat{\boldsymbol{\rho}})$ is plotted against the response rate of the six contact attempts. After the third contact attempt the upper bound to the absolute bias has converged to a value around 8%. ■

5.5 An empirical validation of the sample-based R-indicator

In this section we apply the R-indicator $\hat{R}(\hat{\boldsymbol{\rho}})$ to three studies that investigate different refusal conversion techniques, combinations of data collection modes and contact strategies. The first study involves the re-approach of nonrespondents to the Dutch Labour Force Survey with the call-back approach and the basic-question approach, described in Chapter 4. The second and third study both deal with mixed-mode data collection designs applied to the Dutch Safety Monitor survey and the Dutch Informal Economy survey, see also Chapter 9.

In sections 5.5.2 and 5.5.3 we pay a closer look to the representativeness of the different fieldwork strategies of the studies. First, in section 5.5.1 we describe

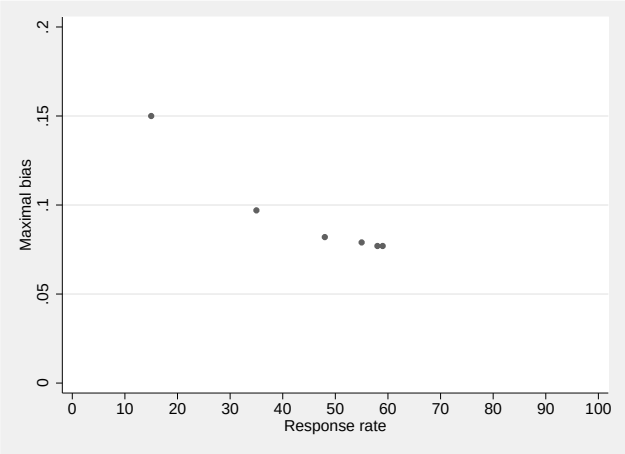

Figure 5.4: *Upper bound to the absolute bias for six contact attempts of POLS 1998*

how we approximate the standard error of $\hat{R}(\hat{\boldsymbol{\rho}})$.

5.5.1 Standard error and confidence interval

To compare the values of the R-indicator for different surveys or data collection strategies, we need to estimate the standard error. The R-indicator $\hat{R}(\hat{\boldsymbol{\rho}})$ is based on the sample standard deviation of the estimated response probabilities. This means that there are two random processes involved. The first process is the sampling of the population. The second process is the response mechanism of the sampled elements. If the true response probabilities were known, then drawing a sample would still introduce uncertainty about the population R-indicator and, hence, lead to a certain loss of precision. However, since we do not know the true response probabilities, these probabilities are estimated based on the sample.

An analytical derivation of the standard error of $\hat{R}(\hat{\boldsymbol{\rho}})$ is not straightforward due to the estimation of the response probabilities. Therefore, we resort to a numerical approximation of the standard error. We estimate the standard error of the R-indicator by non-parametric bootstrapping (Efron and Tibshirani 1993). The non-parametric bootstrap estimates the standard error of the R-indicator by drawing a number $b = 1, 2, \ldots, B$ of so-called bootstrap samples. These are

samples drawn independently and with replacement from the original dataset, of the same size n as the original dataset. The R-indicator is calculated for every bootstrap sample b. We thus obtain B replications of the R-indicator; $\hat{R}_b$, for $b = 1, 2, \ldots, B$. The standard error for the empirical distribution of these B replications is an estimate of the standard error of the R-indicator, that is

$$S_R = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{R}_b - \hat{\overline{R}})^2} \tag{5.30}$$

where $\hat{\overline{R}} = \frac{1}{B}\sum_{b=1}^{B}\hat{R}_b$ is the average estimated R-indicator. In the approximations we take $B = 200$ for all studies. We experimented with the number of bootstrap samples B and found that in all cases the estimate of the standard error had converged at much smaller values than $B = 200$. We determine the $100(1-\alpha)\%$ confidence interval by assuming a normal approximation of the distribution of $\hat{R}(\hat{\boldsymbol{\rho}})$ employing the estimated standard error using (5.30)

$$CI_\alpha = \hat{R}(\hat{\boldsymbol{\rho}}) \pm \xi_{1-\alpha} \times S_R \tag{5.31}$$

with $\xi_{1-\alpha}$ the $1-\alpha$ quantile of the standard normal distribution.

5.5.2 Re-approaching nonrespondents to the Dutch LFS

In previous chapters we presented and discussed the Dutch Labour Force Survey (Chapter 3) and the re-approach of LFS nonrespondents with the call-back approach and the basic-question approach (Chapter 4). In Chapter 4, we analysed the additional response that was obtained with these two re-approaches using the available linked data described in Chapter 3.

For the call-back approach we found that the converted nonrespondents are different from the LFS respondents with respect to the selected auxiliary variables. Furthermore, we found no evidence that the converted nonrespondents were different from persistent nonrespondents with respect to the same characteristics. These findings lead to the conclusion that the combined response of the LFS and call-back approach is more representative with respect to the linked data. The additional response in the basic-question approach was also analysed using the same set of linked data. For the basic-question approach the findings were different for households with and without a listed telephone. When restricted to listed households, the results are the same as for the call-back approach: the response becomes more representative after the addition of the listed basic-question respondents. However, for the overall population, i.e. including

the unlisted households, the opposite was found. The basic-question approach gives 'more of the same' and sharpens the contrast between respondents and nonrespondents. Combining LFS response with basic-question response leads to a less representative composition of the response. In the logistic regression models described in Chapter 4, the indicators for having a listed telephone and having a paid job gave a significant contribution.

Recall that for both re-approaches a simple random sample from the nonrespondents to the LFS is selected. Therefore, the sample is weighted so that the selected nonrespondents represent all the nonrespondents to the LFS. Table 5.2 shows the weighted sample size, response rate, $\hat{R}(\hat{\boldsymbol{\rho}})$, $CI_{0.05}$, $\hat{B}_m$ and $\hat{E}_m$ for the response to the LFS, the response to the LFS combined with the call-back response and the response to the LFS combined with the basic-question response. The standard errors are computed with $B = 200$ bootstraps. The standard errors are relatively large with respect to the studies in subsequent sections due to the weighting. There is an increase in $\hat{R}(\hat{\boldsymbol{\rho}})$ when the call-back respondents are added to the LFS respondents. As both the response rate and the R-indicator increase, the upper bound to the absolute bias $\hat{B}_m$ decreases. The upper bound to the RMSE $\hat{E}_m$ is only slightly higher than the upper bound to the absolute bias $\hat{B}_m$ for the LFS + CBA response, and equal to the upper bound to the absolute bias for the other response groups. The large sample size leads to RMSEs that are only slightly larger than the upper bound to the absolute bias as the variance term vanishes. The confidence intervals $CI_{0.05}$ for the LFS response and the combined LFS and call-back response overlap. However, the one-sided null hypothesis $H_0 : R_{LFS} - R_{LFS+CBA} \geq 0$ is rejected at the 5%-level. Hence, we can conclude that the call-back approach has led to a significant increase of the R-indicator.

In table 5.2 there is a decrease in $\hat{R}(\hat{\boldsymbol{\rho}})$ when we compare the LFS response to the combined response with the basic-question approach. This decrease is not significant. The upper bound to the absolute bias $\hat{B}_m$ slightly decreases.

Table 5.2: *Composition of the LFS response and the combined LFS response with the additional response from the two re-approaches*

Response	*n*	*Response rate*	$\hat{R}(\hat{\boldsymbol{\rho}})$	$CI_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
LFS	18,074	62.2%	80.1%	(77.5; 82.7)	8.0%	8.0%
LFS + CBA	18,074	76.9%	85.1%	(82.4; 87.8)	4.8%	4.9%
LFS + BQA	18,074	75.6%	78.0%	(75.6; 80.4)	7.3%	7.3%

In table 5.3 we restrict the comparison to households with a listed telephone. Again, the standard errors are computed with $B = 200$ bootstraps. In general, the R-indicator is much higher than for all the households. Because the sample size is now smaller, the estimated standard errors are larger as is reflected in the width of the confidence interval. The upper bound to the absolute bias $\hat{B}_m$ is lower than for all households. Due to the large sample size, it is again approximately equal to the upper bound to the RMSE $\hat{E}_m$. The R-indicator $\hat{R}(\hat{\boldsymbol{\rho}})$ for the combined response of the LFS and the basic-question approach now is higher than LFS response alone, but the increase is not significant. The upper bound to the absolute bias $\hat{B}_m$ decreases when the response to the BQA is added to the LFS response. Hence, the R-indicators for the LFS re-approach confirm the conclusions for the call-back approach and the basic-question approach that were drawn in Chapter 4.

5.5.3 Mixed mode pilots

In 2006, Statistics Netherlands conducted two pilots to investigate mixed-mode data collection strategies. The first pilot concerned the Safety Monitor. See Cobben et al. (2007) for details. The second pilot was based on a new survey, the Informal Economy survey; see Gouweleeuw and Eding (2006).

Safety Monitor

The regular Safety Monitor surveys persons of 15 years and older in the Netherlands about issues that relate to safety and police performance. The regular Safety Monitor is a concurrent mixed-mode survey, see also Chapter 9. Persons with a listed telephone are approached by CATI while persons that cannot be reached by telephone are approached by CAPI.

In the pilot in 2006, the possibility to use the internet as one of the modes in a mixed-mode strategy was evaluated. The pilot Safety Monitor is a combination

Table 5.3: *Composition of the LFS response and the combined LFS response with the additional response from the two re-approaches, restricted to households with a listed telephone*

Response	*Size*	*Response rate*	$\hat{R}(\hat{\boldsymbol{\rho}})$	$CI_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
LFS	10,135	68.5%	86.3%	(83.1; 89.5)	5.0%	5.1%
LFS + BQA	10,135	83.0%	87.5%	(84.3; 90.7)	3.8%	3.8%

of a sequential and a concurrent mixed mode design, see figure 5.5. Persons in the pilot were first approached with a web survey. Nonrespondents to the web survey were re-approached by CATI if they had a listed telephone and by CAPI otherwise. In table 5.4 we give the response rates for the regular survey, the pilot response to the web only, and the response to the pilot as a whole. The response to the web survey alone is low. Only 30% of the persons returned the web questionnaire. This implied that close to 70% of the sampled units was re-allocated to either CAPI or CATI. This resulted in an additional response of approximately 35%. The overall response rate is slightly lower than that of the regular Safety Monitor.

Fouwels et al. (2006) performed a univariate analysis of the response compositions in the regular Safety Monitor and the pilot Safety Monitor. They concluded that the response rate is lower for the pilot but that this decrease is quite stable over various demographic subgroups. Furthermore, they observed a univariately declining response rate for the auxiliary variables age, ethnic group, degree of urbanization and type of household. They did, however, find indications that the response becomes less representative when restricting the comparison to the web respondents only. This holds, not surprisingly, especially for the age of the sampled persons.

Table 5.4 gives the sample size, response rate, $\hat{R}(\hat{\boldsymbol{\rho}})$, $CI_{0.05}$, $\hat{B}_m$ and $\hat{E}_m$ for three groups: the regular survey, the pilot survey restricted to web and the pilot survey as a whole. The auxiliary variables age, ethnic group, degree of urbanization and type of household were linked from registers and were selected in logistic model for the response probabilities. The standard errors are computed with $B = 200$ bootstraps. Table 5.4 shows that the R-indicator for the web response is lower than that of the regular survey. The corresponding p-value is close to 5%. As a consequence of both a low response rate and a low R-indicator, the upper bound to the absolute bias $\hat{B}_m$ is more than twice as high as for the regular survey. However, for the pilot as a whole both the R-indicator and $\hat{B}_m$ are close to the values of the regular survey. Due to the smaller sample size of the pilot the estimated standard errors are larger than in the regular survey. The upper bound to the RMSE $\hat{E}_m$ is only slightly higher than the upper bound to the absolute bias $\hat{B}_m$. The findings in table 5.4 do not contradict the conclusions of Fouwels et al. (2006). We also find that the web response in the pilot has a less balanced composition whereas the composition of the full pilot response is not significantly worse than that of the regular Safety Monitor.

Informal Economy

The second mixed-mode pilot in 2006 concerned the Informal Economy. This

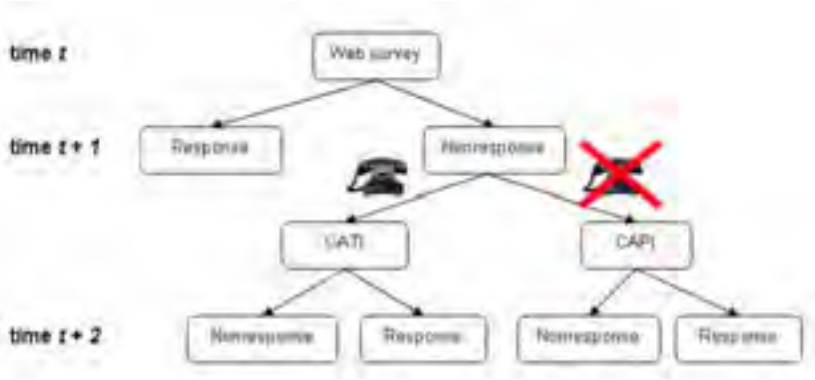

Figure 5.5: *The mixed mode design of the pilot Safety Monitor*

Table 5.4: *Composition of the response to the regular Safety Monitor, the response to the pilot web survey only, and the total response to the pilot Safety Monitor*

Response	*Size*	*Response rate*	$\hat{R}(\hat{\rho})$	$CI_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
Regular	30,139	68.9%	81.4%	(80.3; 82.4)	6.8%	6.8%
Pilot web only	3,615	30.2%	77.8%	(75.1; 80.5)	18.3%	18.4%
Pilot total	3,615	64.7%	81.2%	(78.3; 84.0)	7.3%	7.4%

is a new survey that is set up at Statistics Netherlands. The 2006 survey served as a pilot for subsequent years. The target population consists of persons of 16 years and older. Questions about unpaid labour, or moonlighting, are the main interest of this survey. In the pilot, two samples are selected. One sample is approached by CAPI. The second sample is first approached by a combination of a web- and a mail survey. Nonrespondents to the web/mail survey that have a listed telephone are re-approached by CATI whereas nonrespondents without a listed telephone are not re-approached. The second sample is thus approached by a combination of a concurrent and a sequential mixed mode design. See figure 5.6. We consider three groups: the CAPI respondents, the web/mail respondents, and the web/mail respondents supplemented by the CATI response in the re-approach. See table 5.5 for the results of this pilot.

Gouweleeuw and Eding (2006) compared the three groups univariately with respect to age, gender, level of education and ethnic group. They found small

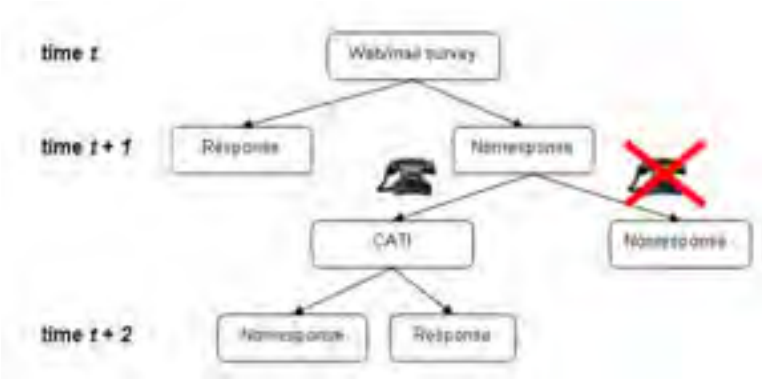

Figure 5.6: *The mixed mode design of the second sample in the pilot Informal Economy*

differences with respect to age and gender. However, with respect to ethnic group they concluded that the composition of the CAPI response is more comparable to the population than those of the other two groups. With respect to level of education they found that all groups deviate from the population but in different ways. More high educations are found in the web/mail and web/mail/CATI group while low educations are over represented in the CAPI group. From the univariate comparisons it is, therefore, not immediately clear to which degree the overall compositions of the response are different. Table 5.5 gives the R-indicators for the three groups with their corresponding confidence intervals and maximal absolute bias and RMSE. The standard errors are computed with $B = 200$ bootstraps.RMS The response rate in the web/mail

Table 5.5: *Composition of the response to the CAPI sample, the response to the web/mail survey, and the total response to second sample in the pilot Informal Economy*

Response	*Size*	*Response rate*	$\hat{R}(\hat{\boldsymbol{\rho}})$	$CI_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
CAPI	2,000	56.7%	77.2%	(73.0; 81.4)	10.1%	10.2%
Web/mail	2,001	33.8%	85.1%	(81.5; 88.7)	11.0%	11.2%
Web/mail + CATI	2,001	49.0%	78.0%	(74.4; 81.6)	11.2%	11.3%

group is considerably lower than in the other samples. However, surprisingly the R-indicator of the web/mail group is significantly higher than in the other groups at the 5% level. Furthermore, despite of the lower response rate, the upper bound to the absolute bias $\hat{B}_m$ does not differ considerably from that of the other groups. Hence, from these results we may conclude that with respect to the four selected auxiliary variables, there is no reason to favour the other groups to the web/mail alternative. A possible explanation for this result, is that the web- and mail survey are not interviewer-assisted but self-administered and hence, sensitive questions are more easily answered than face-to-face or by telephone.

The re-allocation of nonrespondents to CATI resulted in an additional response of about 15%. Still the response rate is lower than in the CAPI group. The R-indicator of this group is, however, comparable to that of the CAPI group. The R-indicator is lower than the web/mail group. This result once more confirms that persons with a listed telephone are a special group that differ from non-listed persons in both composition and response behaviour.

5.6 Concluding remarks

The R-indicator presented in this chapter is promising because, under the restriction that the auxiliary information is available for all sample elements, it can easily be computed and allows for interpretation and normalization when response propensities can be estimated without error. The application to real survey data shows that the R-indicator confirms earlier analyses of the nonresponse composition. Other R-indicators can be constructed by choosing different distance functions between vectors of response propensities. The R-indicator and graphical displays that we showed in this chapter can be computed using most standard statistical software packages.

The computation of the R-indicator is sample-based and employs models for individual response propensities. Hence, the R-indicator itself is a random variable and there are two estimation steps that influence its bias and variance. However, it is mostly the modelling of response propensities that has important implications. The restriction to the sample for the estimation of the R-indicator implies that the indicator is less precise but this restriction does not introduce a bias. Model selection and model fit usually are performed by choosing a significance level and adding only those interactions to the model that give a significant contribution. The latter means that the size of the sample plays an important role in the estimation of response propensities. The model selection

strategy may introduce bias.

There are various approaches to deal with the dependency on the size of the sample. One may restrain from model selection and fix a stratification beforehand. That way bias is avoided, but standard errors are not controlled and may be considerable. Another approach is the development of 'best practices' based on an empirical validation.

It is important to constantly keep in mind that the R-indicator must be viewed relative to the set of auxiliary variables. When comparing R-indicator values one needs to fix the vector of auxiliary variables. The application of the R-indicator showed that there is no clear relation between response rate and representativeness of response. Larger response rates do not necessarily lead to a more balanced response. Not surprisingly, we do find that higher response rates reduce the risk of nonresponse bias. The higher the response rate, the smaller the maximal absolute bias of survey items.

Application to the selected studies learned that standard errors do decrease with increasing sample size as expected but they are still relatively large for modest sample sizes. For example for a sample size of $3,600$ we found a standard error of approximately 1.3%. Hence, if we assume a normal distribution, then the 95% confidence interval has an approximate width of 5.4%. The sample size of the LFS is about $30,000$ units. The standard error is approximately 0.5% and the corresponding 95% confidence interval is approximately 2% wide. The standard errors are larger than we expected, possibly due to the composition of the ELFS-sample, i.e. the elements that are selected for the re-approach strategies receive large weights compared to the regular respondents.

This chapter contains a first empirical study of an R-indicator and its standard error. Much more theoretical and empirical research is necessary to get a grip on R-indicators and their properties. First, we did not consider survey items at all. Clearly, it is imperative that we do this in the future. However, as we already argued, the R-indicator is dependent on the set of auxiliary variables. It can therefore be conjectured that, as for nonresponse adjustment methods, the extent to which the R-indicator predicts nonresponse bias of survey items is completely dependent on the missing-data mechanism. In a missing-data mechanism that is strongly non-ignorable, the R-indicator will not do a good job. However, without knowledge about the missing-data mechanism no other indicator may either. For this reason we constructed the notion of an upper bound to the absolute bias, as this gives an upper bound to the nonresponse bias under the worst-case-scenario.

A second topic of future research is a theoretical derivation of the standard error of the R-indicator used in this chapter. We believe that the non-parametric

bootstrap errors are good approximations. However, if we want the R-indicator to play a more active role in the comparison of different strategies, then we need (approximate) closed forms. Third, we will need to investigate the relation between the selection and number of auxiliary variables and the standard error of the R-indicator. These topics, as well as some other issues, will be investigated in the RISQ-project in the 7th Framework Programme of the European Union, see Bethlehem and Schouten (2008).

Appendix

5.A Estimating response probabilities

The response probability ρ_i is a latent variable. Furthermore, its value lies between 0 and 1. To estimate this type of variable we can use a probability model that restrains the predicted outcome to lie between 0 and 1. In other words, we want $P(R_i = 1) = F(\mathbf{X}_i'\boldsymbol{\beta})$ with some function F translating $\mathbf{X}_i'\boldsymbol{\beta}$ into a value between 0 and 1. $\mathbf{X}_i$ represents a $J \times 1$-vector of auxiliary variables with values for sample element i, and $\boldsymbol{\beta}$ is a $J \times 1$-vector of coefficients that correspond to the different auxiliary variables in $\mathbf{X}_i$.

Response probabilities are essential in nonresponse research. However, individual probabilities can only be estimated by looking at the aggregate response level. For each sample element there is simply only one replication. Therefore, we switch from theoretical response probabilities to the more practical response propensities defined by (1.10).

We consider three models to compute the response propensities $\rho(\mathbf{X}_i)$, each using a different transformation function F, or, in the framework of Generalized Linear Models (GLM, see for example Greene, 2003), a different known link function. The first choice of F is the identity function, which results in the linear probability model $P(R_i = 1) = \mathbf{X_i'}\boldsymbol{\beta}$. This model, however, does not restrict the propensities to lie between 0 and 1. Another transformation is the standard normal distribution, which does restrict the probability to lie between 0 and 1. This results in the probit model

$$\rho(\mathbf{X}_i) = \Phi(\mathbf{X_i'}\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{X_i'}\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} exp(\frac{-z^2}{2})dz \tag{5.32}$$

When choosing F to be the logistic distribution, this leads to the logit model

$$\rho(\mathbf{X}_i) = \Lambda(\mathbf{X_i'}\boldsymbol{\beta}) = \frac{exp(\mathbf{X_i'}\boldsymbol{\beta})}{1 + exp(\mathbf{X_i'}\boldsymbol{\beta})} \tag{5.33}$$

The probit and the logit function are the most common models but in fact any model with the right properties can be used. The logit transformation leads to the logistic regression model

$$log(\frac{R_i}{1 - R_i}) = logit(\rho(\mathbf{X}_i)) = \mathbf{X_i'}\boldsymbol{\beta} \tag{5.34}$$

whereas the probit transformation is defined as the solution of (5.32). Note that the logit regression model does not have an error term. The probit and the logit models can be described as latent variable regression models. To see that, we define a latent variable $\rho_i^*(\mathbf{X}_i)$ as

$$\rho^*(\mathbf{X}_i) = \mathbf{X}_i'\boldsymbol{\beta} + \boldsymbol{\epsilon}_\mathbf{i} \tag{5.35}$$

The latent variable $\rho^*(\mathbf{X}_i)$ can be regarded as a transformation of the response propensity into the $(-\infty, \infty)$-interval. The error terms are assumed to be identical and independently normally distributed, i.e. $\epsilon_i \sim NID(0, \sigma^2)$. We do not observe $\rho^*(\mathbf{X}_i)$, but rather we observe the response indicator R_i which takes on values of 0 or 1 according to

$$R_i = \begin{cases} 1, & \text{if } \rho^*(\mathbf{X}_i) > 0 \\ 0, & \text{else} \end{cases} \tag{5.36}$$

The latent variable $\rho^*(\mathbf{X}_i)$ is not equal to the response probability ρ_i because its value is not restricted to the interval $(0, 1)$. However, the probit or logit transformation ensures that the estimated values, the response propensities $\hat{\rho}_i$, can be interpreted as probabilities.

For more information on these models see for instance Johnston and Dinardo (1997) or Greene (2003). There are alternative ways to estimate response probabilities, for example by a Response Homogeneity Group model (RHG), see Särndal et al. (1992). The sample s is divided into different groups so that within each group all sample elements are assumed to have the same response probability. This model corresponds to a logit or probit model when the auxiliary variables $\mathbf{X}$ are all qualitative variables. Another possibility is the use of classification trees, see Breiman et al. (1998) and Schouten and De Nooij (2005). The sample is divided into classes using binary splits based on a certain criterion. The final leaves of the trees form strata of sample elements with homogeneous response behaviour.

Does it matter which model is used to compute the response propensities? According to Dehija and Wahba (1999) it does not. They use a logistic model and state that other models yield similar results. However, Laaksonen (2006) provides some examples in which it does matter which model is used. He evaluates logit, probit, complementary log-log and log-log models. Furthermore, there is no benchmark with respect to how well the model should fit the data in order to be used effectively in adjustment for nonresponse.

Chapter 6

A Review of Nonresponse Adjustment Methods

In Chapter 1 we discussed some basic sampling theory, and we showed how the information that is collected by means of a survey sample can be used to estimate population parameters, or population characteristics, such as the population total or the population mean. We introduced the Horvitz-Thompson estimator as an unbiased estimator for population characteristics, and we showed that by using auxiliary information in the estimation of population characteristics more accurate estimators are obtained. However, in Chapter 1 we restricted the situation to full response to the survey. This is not a realistic situation, sample elements do not participate in surveys for many reasons as we have discussed in Chapter 2.

One method to deal with nonresponse is re-approaching nonrespondents. We presented an application of two techniques to re-approach nonrespondents in Chapter 4. Another way of dealing with nonresponse is to take measures to reduce nonresponse. These measures can be taken before and during the data collection, see for example Groves and Couper (1998).

However, despite methods to enhance the response or re-approach nonrespondents, the response will never be 100%, nor will it be completely non-selective. A consequence of selective nonresponse is that the distribution of respondent characteristics differs from the distribution in the population. An example of this is provided in Chapter 3. The selectivity of the subsample of respondents can result in a bias in the estimates for population characteristics.

The aim of this chapter is to provide an overview of the traditional nonresponse adjustment methods that are used most frequently in survey organisations.

6.1 Introduction

The general idea behind nonresponse adjustment is to assign weights to the respondents so that they also represent the nonrespondents. This is usually done in such a way that the distribution of the respondents' characteristics is similar to the distribution of the characteristics in the sample or population, with respect to auxiliary variables.

In Chapter 1 we introduced the Horvitz-Thompson estimator and the generalized regression estimator as unbiased estimators of population characteristics in case of full response to a survey. Furthermore, we introduced nonresponse in the classic sampling theory by means of the random response model, in which every person i has a non-zero probability ρ_i of responding to the survey. Unlike the sampling phase, the nonresponse phase is out of control of the survey researcher. Based on the random response model, we showed what the effects of selective nonresponse may be on the estimators of population characteristics. In this chapter, we discuss methods to adjust for the bias due to selective nonresponse.

Nonresponse adjustment methods assign weights to the respondents, so that the weighted respondents represent the population as close as possible with respect to known auxiliary information. It is then assumed that the estimates for population characteristics become less biased, i.e. the nonresponse is assumed to be missing-at-random (see Chapter 1 for a detailed description of missing data mechanisms).

In Chapter 1 (equation (1.20)) we showed that the Horvitz-Thompson estimator can be written as

$$\overline{y}_{ht} = \frac{1}{N}\sum_{i\in s}\frac{Y_i}{\pi_i} = \frac{1}{N}\sum_{i\in s} d_i Y_i \tag{6.1}$$

where $d_i = \frac{1}{\pi_i}$ is the design weight. Adjustment weighting replaces this estimator by a new estimator

$$\overline{y}_w = \frac{1}{N}\sum_{i\in s} w_i Y_i \tag{6.2}$$

where $w_i = g_i d_i$ and g_i is the correction weight, often referred to as g-weight, produced by a weighting adjustment method. The Horvitz-Thompson estimator (6.1) is obtained by $g_i = 1$ for all i.

By introducing nonresponse into the sampling theory based on the random response model, another type of randomness enters into the framework. Besides the randomness of the sample selection, in addition every sample element has an unknown individual response probability ρ_i. So, sample elements are equipped with known inclusion probabilities and unknown response probabilities. The randomization theory, or design-based theory, now becomes a quasi-randomization theory (Oh and Scheuren, 1983). This leads to the two-phase approach for nonresponse adjustment (Särndal et al. 1992). First, a sample is selected from the target population. The selection is followed by nonresponse. This theory leads to a modified version of the Horvitz-Thompson estimator (1.20)

$$\overline{y}_{ht}^{r} = \frac{1}{N} \sum_{i \in r} \frac{d_i Y_i}{\rho_i} \tag{6.3}$$

where r is the subsample of respondents. The Horvitz-Thompson estimator is extended to two phases: sample selection, reflected in the design weight; and response, reflected in the response probability. However, the response probabilities are unknown in practice. Based on some appropriate model and by using the available information on the sample we can estimate the response probabilities (see Chapter 5, appendix 5.A). Särndal (1981) proposes to insert the estimated response probabilities $\hat{\rho}_i$ into (6.3) so that

$$\hat{y}_{ht}^{r} = \frac{1}{N} \sum_{i \in r} \frac{d_i Y_i}{\hat{\rho}_i} \tag{6.4}$$

Bethlehem (1988) suggests, as a first simple step, to replace the unknown response probabilities ρ_i by the Horvitz-Thompson estimator for the mean response probability, i.e.

$$\overline{\rho}_{ht} = \frac{1}{N} \sum_{i=1}^{N} d_i R_i \tag{6.5}$$

consequently, the modified Horvitz-Thompson estimator in case of nonresponse can be expressed as

$$\tilde{y}_{ht}^{r} = \frac{1}{N} \sum_{i \in r} \frac{d_i Y_i}{\overline{\rho}_{ht}} \tag{6.6}$$

Both modifications of the Horvitz-Thompson estimators, (6.4) and (6.6), use auxiliary information from the design stage, and in addition estimator (6.4)

uses auxiliary information for the estimation of ρ. The other approaches that we discuss in this section make the sample representative with respect to several auxiliary variables. The approaches thus depend on the quality of the auxiliary information. In Chapter 1 we described that an ideal auxiliary variable has three features: it explains well the response behaviour, it explains well the survey items, and it identifies the most important domains for publication of the survey statistics. Särndal and Lundström (2008) remark that in practice, even the best auxiliary variables will not be capable of completely removing the nonresponse bias. In their paper, they propose an indicator to compare and rank auxiliary variables in their effectiveness of adjusting for nonresponse bias. The indicator that Särndal and Lundström (2008) propose is mainly based on the first feature, explaining response behaviour. Schouten (2007) proposes an algorithm to select the most influential auxiliary variables for nonresponse adjustment. His approach is based on the first two features, taking into account both the relationship with the response behaviour and the survey items.

In this chapter, we provide a review of methods that use auxiliary information for the estimation of survey items. These methods differ in the way that they calculate the correction weights g_i for $i = 1, 2, \ldots, n$. We focus on the calibration framework. In this framework the design weights are adjusted so that the sample sum of weighted auxiliary variables equals its known population total, under the constraint that the final weights are as close as possible to the initial design weights.

The calibration framework embodies a large number of conventional nonresponse adjustment methods, for instance the generalized regression (GREG) estimator. The GREG is based on the relationship between the response behaviour, the survey items and the auxiliary variables. Another method that has become popular in adjusting for nonresponse bias, is the propensity score method. This method is based on estimated response probabilities, or response propensities and does not belong to the calibration framework.

This chapter is outlined as follows. Section 6.2 is devoted to the generalized regression estimator modified to nonresponse. Next, we outline the calibration framework in section 6.3. In section 6.4 we show how the propensity score method can be combined with the GREG-approach, thereby including the relationship with survey items. This chapter ends with some concluding remarks in section 6.5.

6.2 The generalized regression estimator

Bethlehem (1988) introduced the generalized regression (GREG) estimator to adjust for nonresponse bias. In Chapter 1 we already discussed that the GREG-estimator increases the precision of estimated population characteristics by using auxiliary information. In this section, we show how the GREG-estimator can be modified to deal with nonresponse. We thereby follow the design-based approach.

The GREG-estimator makes weighted sample estimates for quantitative variables conform to population parameters. It is derived from the standard regression estimator, and involves incorporating the adjustment for auxiliary variables as a modification of the weights. It can be used to cover several auxiliary variables, unequal weights, transformations of variables and interactions between variables.

In Chapter 1 we showed that the GREG-estimator can be derived from a linear regression of the values Y on $\mathbf{X}$, i.e.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6.7}$$

where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$ is an $N \times 1$-vector of population values for Y_i, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)'$ is an $J \times 1$-vector of regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)'$ is an $N \times 1$-vector of residuals. The general regression estimator can be estimated by the sample observations and computed as (see (1.26))

$$\overline{y}_{gr} = \overline{y}_{ht} + (\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht})'\hat{\boldsymbol{\beta}}^{(s)} \tag{6.8}$$

where $\overline{y}_{ht}$ is the Horvitz-Thompson estimator of the survey item mean, $\overline{\mathbf{x}}_{ht}$ is the analogue of $\overline{y}_{ht}$ and $\hat{\boldsymbol{\beta}}^{(s)}$ is the sample-based estimate of $\boldsymbol{\beta}$ (see (1.28))

$$\hat{\boldsymbol{\beta}}^{(s)} = (\mathbf{X}\boldsymbol{\Pi}^{-1}\mathbf{X}')^{-1}\mathbf{X}\boldsymbol{\Pi}^{-1}\mathbf{Y} = \Big(\sum_{i=1}^{n} d_i \mathbf{X}_i \mathbf{X}_i'\Big)^{-1}\Big(\sum_{i=1}^{n} d_i \mathbf{X}_i Y_i\Big) \tag{6.9}$$

and

$$\overline{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \tag{6.10}$$

In case of nonresponse Bethlehem (1988) introduces the following modified version of the generalized regression estimator

$$\overline{y}^r_{gr} = \overline{y}^r_{ht} + (\overline{\mathbf{X}} - \overline{\mathbf{x}}^r_{ht})'\hat{\boldsymbol{\beta}}^{(s)}_r \tag{6.11}$$

where $\overline{y}_{ht}^{r}$ is the modified Horvitz-Thompson estimator of the survey item mean, $\overline{\mathbf{x}}_{ht}^{r}$ is the analogue of $\overline{y}_{ht}^{r}$ and $\hat{\boldsymbol{\beta}}_{r}^{(s)}$ is defined by

$$\hat{\boldsymbol{\beta}}_{r}^{(s)} = \Big(\sum_{i=1}^{n} d_i R_i \mathbf{X}_i \mathbf{X}_i'\Big)^{-1} \Big(\sum_{i=1}^{n} d_i R_i \mathbf{X}_i Y_i\Big) \tag{6.12}$$

So $\hat{\boldsymbol{\beta}}_{r}^{(s)}$ is an analogue of the vector of coefficients $\hat{\boldsymbol{\beta}}^{(s)}$ in (6.9), but just based on the response data. Under the condition that there exists a vector $\boldsymbol{\lambda}'$ such that $1 = \boldsymbol{\lambda}'\mathbf{X}_i$ (see (1.31)), the general regression estimator in the case of nonresponse can be expressed as

$$\overline{y}_{gr}^{r} = \overline{\mathbf{X}}'\hat{\boldsymbol{\beta}}_{r}^{(s)} \tag{6.13}$$

This estimator can be expressed as a weighted sum of survey item observations based on the subsample of respondents, i.e.

$$\overline{y}_{w} = \frac{1}{N}\sum_{i\in r} w_i Y_i \tag{6.14}$$

where again $w_i = g_i d_i$. The g-weights can be expressed as

$$g_i = 1 + \Big(\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht}^{r}\Big)' \Big(\sum_{i=1}^{n_r} d_i \mathbf{X_i} \mathbf{X_i}'\Big)^{-1} \mathbf{X_i} \tag{6.15}$$

for $i \in r$. Again, $\overline{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i$ when the auxiliary information is available on the population level and $\overline{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{n} d_i \mathbf{X}_i$ when the auxiliary information is on the sample level. The GREG-estimator thus can also be expressed as

$$\overline{y}_{gr}^{r} = \frac{1}{N}\sum_{i\in r} d_i g_i Y_i \tag{6.16}$$

Bethlehem (1988) shows that an approximation of the bias of (6.11) is given by

$$B(\overline{y}_{gr}^{r}) = \mathbf{X}\boldsymbol{\beta}_r - \overline{Y} \tag{6.17}$$

where $\boldsymbol{\beta}_{\mathbf{r}}$ is defined by

$$\boldsymbol{\beta}_r = \Big(\sum_{i=1}^{N} \rho_i \mathbf{X}_i \mathbf{X}_i'\Big)^{-1} \Big(\sum_{i=1}^{N} \rho_i \mathbf{X}_i Y_i\Big) \tag{6.18}$$

The bias of this estimator vanishes if $\boldsymbol{\beta}_{\mathbf{r}} = \boldsymbol{\beta}$. Thus, the regression estimator is unbiased if nonresponse does not affect the regression coefficients. Practical experience (at least in the Netherlands) shows that nonresponse often seriously affects estimators like means and totals, but less often causes estimates of relationships to be biased. Particularly if relationships are strong, i.e. the regression line fits the data well, the risk of finding wrong relationships is small. Bethlehem (1988) shows that by writing

$$\boldsymbol{\beta}_r = \boldsymbol{\beta} + \Big(\frac{1}{N}\sum_{i=1}^{N}\frac{\rho_i \mathbf{X}_i \mathbf{X}_i'}{\overline{\rho}}\Big)\overline{\epsilon}_r \tag{6.19}$$

where

$$\overline{\epsilon}_r = \frac{1}{N}\sum_{i=1}^{N}\frac{\rho_i \epsilon_i}{\overline{\rho}} \tag{6.20}$$

two conclusions can be drawn. $\boldsymbol{\beta}_r$ and $\boldsymbol{\beta}$ will be approximately equal if (6.20) is close to or equal to zero. This is the case when

- there is a good fit of the regression model.
- there is little or no correlation between the residuals of the regression model and the response probabilities. This condition is satisfied if the data are missing-at-random (MAR), and the auxiliary variables concerned are included in the regression model.

6.3 Calibration estimators

Calibration estimators in the full response case are described in Deville and Särndal (1992) and Deville et al. (1993). The calibration framework for nonresponse adjustment is introduced by Lundström and Särndal (1999) and elaborately discussed in Särndal and Lundström (2005). It describes how we can estimate the population average $\overline{Y}$ for survey item Y in the presence of nonresponse. The distribution of the auxiliary information for the subsample of respondents r is calibrated to known auxiliary information, under the condition that the adjustment weights should not differ too much from the design weights.

When the distribution of the auxiliary variables is calibrated to the population, it is assumed that the weighted distribution of the survey item also resembles the distribution of the survey item in the population. This is the

missing-at-random (MAR) assumption. Within classes of the auxiliary variables $\mathbf{X}$, it is assumed that persons have the same value of the survey item Y. In those classes, the respondents closely resemble the nonrespondents and their values can be weighted to represent the total in their class.

Recall that we are interested in $\overline{Y} = \frac{1}{N}\sum_{i\in U} Y_i$. However, based on the observed persons only we would have the response mean $\overline{Y} = \frac{1}{n_r}\sum_{i\in s} R_i Y_i$. To adjust for the effects of nonresponse, we assign weights to the respondents so that they represent the sample or the population. The calibration equation assures that if we apply the calibration estimator to the auxiliary variables in the subsample of respondents, we will obtain the known values for the auxiliary variables in the population. In other words, calibration estimation assures that the distribution of the auxiliary variables in the subsample of respondents is equal to the distribution of the auxiliary variables in the population. This results in the following calibration estimator

$$\overline{y}_w = \frac{1}{N}\sum_{i\in r} w_i Y_i \tag{6.21}$$

where the weights w_i are determined by the calibration equation

$$\frac{1}{N}\sum_{i\in r} w_i \mathbf{X}_i = \overline{\mathbf{X}} \tag{6.22}$$

The calibration weights w_i are composed of the design weight d_i and an adjustment for the nonresponse effect, denoted by weight v_i, i.e. $w_i = d_i v_i$. These weights v_i have to be as close to 1 as possible. The weights w_i are not uniquely defined by (6.22). Therefore, Lundström and Särndal (1999) required that the difference between the weights w_i and the design weights d_i minimises some distance function. For example, the distance function

$$d(w_i, d_i) = \sum_{i\in r}(w_i - d_i)^2/d_i \tag{6.23}$$

leads to the GREG-estimator. If we assume that there exists a linear relationship between the weights and the auxiliary information, we can determine the weights by solving the Lagrange multiplier problem $v_i = 1+\boldsymbol{\lambda}_\mathbf{r}{}'\mathbf{X}_i$. Kott (2006) describes an approach that allows for a nonlinear relationship between the weights and the auxiliary information. For instance, the linear functional form described by $v_i = 1 + \boldsymbol{\lambda}_\mathbf{r}{}'\mathbf{X}_i$ can be expressed as $w_i = f(\boldsymbol{\lambda}_\mathbf{r}{}'\mathbf{X}_i)$. Nonlinear functions allow for a broader class of calibration estimators, see for example Kott (2006).

For the linear functional form, the solution for $\boldsymbol{\lambda_r}'$ becomes

$$\boldsymbol{\lambda_r}' = (\overline{\mathbf{X}} - \overline{x}_{ht}^r)'(\sum_r d_i \mathbf{X}_i \mathbf{X}_i')^{-1} \tag{6.24}$$

where $\overline{x}_{ht}^r$ is the modified Horvitz-Thompson estimator for the mean of the auxiliary variables. The specification of the calibration equation (6.22) is different depending on the level of the available auxiliary information. In case of information on the population level, the calibration equation (6.22) becomes

$$\frac{1}{N}\sum_{i\in r} w_i \mathbf{X}_i = \frac{1}{N}\sum_{i\in U} \mathbf{X}_i \tag{6.25}$$

Consequently, the population average $\overline{\mathbf{X}}$ in (6.24) is $\frac{1}{N}\sum_{i\in U}\mathbf{X}_i$. In case of information on the sample level, the calibration equation (6.22) becomes

$$\frac{1}{N}\sum_{i\in r} w_i \mathbf{X}_i = \frac{1}{N}\sum_{i\in s} d_i \mathbf{X}_i \tag{6.26}$$

and $\overline{\mathbf{X}}$ in (6.24) then refers to the estimated population average $\frac{1}{N}\sum_{i\in s} d_i \mathbf{X}_i$.

The weights w_i are referred to as the final calibration weights and can be described as

$$\begin{aligned} w_i &= d_i v_i \\ &= d_i(1 + \boldsymbol{\lambda_r}'\mathbf{X}_i) \\ &= d_i(1 + (\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht}^r)'(\textstyle\sum_{i\in r} d_i \mathbf{X}_i \mathbf{X}_i')^{-1}\mathbf{X}_i) \end{aligned} \tag{6.27}$$

Lundström and Särndal (1999) and Särndal and Lundström (2005) derive estimators for the variance and the mean square error (MSE) of the calibration estimators under the two-phase approach to nonresponse. That is to say, the first phase involves the selection of the sample from the population, the second phase involves the realisation of the subsample of respondents from the sample, given the sample. For the variance derivations we refer to these references. In Chapter 1 we discussed that the MSE consists of two components; a variance component and a bias component. When the sample size of the survey increases, the variance approaches zero if there is no nonresponse. However, the bias does not approach zero with increasing sample size. The mean squared error is therefore often dominated by the bias component. Särndal and Lundström (2005) show that the variance can be described as the sum of the sampling variance and the nonresponse variance.

Example 6.3.1 Special cases
The modified generalized regression estimator discussed in section 6.2 belongs to the family of calibration estimators for nonresponse adjustment. This can be seen by looking at the GREG-estimator expressed as a weighted sum of sample observations, described in (6.16). The g-weights in (6.15) are similar to the v-weights in (6.27).

The post-stratification estimator arises when the auxiliary information is qualitative. Suppose we can classify the population into H mutually exclusive and exhaustive groups, based on one auxiliary variable or a crossing of two or more auxiliary variables. The population is divided into H strata denoted by $U_1, U_2, \ldots, U_H$. Likewise, the sample can be divided into H strata $s_1, s_2, \ldots, s_H$. Let $X_1, X_2, \ldots, X_H$ be the corresponding dummy variables. Then, for an element in stratum h, the corresponding dummy variable X_h is assigned the value 1, and all other dummy variables are set to 0. The population size of stratum h is equal to N_h with corresponding sample size n_h. Furthermore, let r_h be the subsample of respondents in group h. Note that the size of the subsample r_h is a random variable and not a fixed number.

Post-stratification assigns the same weight to all responding elements in the same stratum. This implies that for all elements $i \in r_h$ the weight v_i equals

$$v_i = \frac{N_h}{n_h} \tag{6.28}$$

The calibration estimator (6.21) can be expressed as

$$\begin{array}{ll} \overline{y}_w & = \frac{1}{N} \sum_r d_i v_i Y_i = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in r_h} d_i Y_i \\ & = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{y}^r_{ht,h} \end{array} \tag{6.29}$$

where

$$\overline{y}^r_{ht,h} = \frac{1}{n_h} \sum_{i \in r_h} d_i Y_i \tag{6.30}$$

is the modified Horvitz-Thompson estimator for the stratum mean of survey item Y in stratum h. Särndal (1980) replaces the size n_h of stratum h by the Horvitz-Thompson estimator for n_h, i.e. $n_{ht,h} = \sum_{i \in r_h} d_i$. The post-stratification estimator is nearly unbiased under equal response probabilities within the groups defined by $\mathbf{X}$.

Other special cases are the ratio estimator and the regression estimator. The ratio estimator is obtained when there is one single quantitative auxiliary variable. This estimator assumes that the auxiliary variable X has values

that are more or less proportional to the values of the survey item. The regression estimator arises when there is one quantitative variable X, like the ratio estimator, but in addition N is known. The auxiliary vector can then be described as $\mathbf{X} = (1, X_i)'$. The ratio estimator is a nearly unbiased estimator for $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ if and only if the nonresponse occurs completely at random. Hence the response probabilities need to be equal for all the elements in the population. This is a very strong and rarely satisfied condition. The regression estimator provides better protection against nonresponse bias than the ratio estimator. ■

6.4 The propensity score method

6.4.1 Two-phase approach

As we have mentioned before, the quasi-randomization approach, or the two-phase approach refers to the idea of sample selection followed by the realisation of a response subsample. Särndal et al. (1992) describe the two-phase approach. Based on a known response distribution given the sample s, the response probabilities are known and can be used in the estimation of population characteristics. We already introduced in Chapter 1 the response indicator R_i and the corresponding conditional response probability $P(R_i = 1|\delta_i = 1) = \rho_i$. Now, we can use the modified Horvitz-Thompson estimator described in (6.3).

The assumption of a known response distribution is, unfortunately, not realistic. We do not know the individual response probabilities and we have to estimate them. In Chapter 1 we discussed a number of methods that can be used to estimate the individual response probabilities, denoted by $\hat{\rho}_i = \rho(\mathbf{X}_i)$. Särndal and Lundström (2005) show how the GREG-estimator can be extended to account for the two phases, i.e.

$$\overline{y}_{gr}^{r} = \frac{1}{N}\sum_{i=1}^{n_r} \frac{d_i}{\rho_i} g_{i\rho} Y_i \tag{6.31}$$

in which the g-weights are

$$g_{i\rho} = 1 + \left(\overline{\mathbf{X}} - \overline{x}_{ht}^{r}\right)' \Big(\sum_{i=1}^{n_r} \frac{d_i}{\rho_i} \mathbf{X}_i \mathbf{X}_i' \Big)^{-1} \mathbf{X}_i \tag{6.32}$$

Compared to the g-weights from the full response case, the distinction is that the modified version $\overline{x}_{ht}^{r}$ of the Horvitz-Thompson estimator for the mean of the

auxiliary variables is used and that the design weights d_i are updated for the second phase, the response phase, by dividing them by the response probabilities ρ_i. Based on a suitable model using auxiliary information, the unknown response probabilities ρ_i are replaced by their estimates $\hat{\rho}_i$ in (6.31) so that

$$\overline{y}_{gr}^{r} = \frac{1}{N}\sum_{i=1}^{n_r}\frac{d_i}{\hat{\rho}_i}g_{i\hat{\rho}}Y_i \tag{6.33}$$

and likewise

$$g_{i\hat{\rho}} = 1 + \Big(\overline{\mathbf{X}} - \overline{x}_{ht}^{r}\Big)'\Big(\sum_{i=1}^{n_r}\frac{d_i}{\hat{\rho}_i}\mathbf{X}_i\mathbf{X}_i'\Big)^{-1}\mathbf{X}_i \tag{6.34}$$

These weights can also be expressed as the final calibration weights w_i by

$$w_i = \frac{d_i}{\hat{\rho}_i}\Big[1 + (\overline{\mathbf{X}} - \overline{x}_{ht}^{r}\Big)'\Big(\sum_{i=1}^{n_r}\frac{d_i}{\hat{\rho}_i}\mathbf{X}_i\mathbf{X}_i'\Big)^{-1}\mathbf{X}_i\Big] \tag{6.35}$$

In Särndal et al. (1992), a variance estimator is derived for (6.33) in case the response probabilities are estimated by a Response Homogeneity Group model (see Chapter 1). The sample s is divided into different groups so that within each group all sample elements are assumed to have the same response probability. This model corresponds to a logit or probit model when the auxiliary variables $\mathbf{X}$ are all qualitative variables. If and only if the model used to estimate the response probabilities fits the data well, the estimator (6.33) is nearly unbiased.

Kalton and Flores-Cervantes (2003) describe the method of logistic regression weighting. The response probabilities are estimated with a logistic regression model and inserted in the modified Horvitz-Thompson estimator (6.4). Little (1986) has also applied this approach. Kalton and Flores-Cervantes (2003) note that, when the auxiliary information that is being used in the logistic regression model is a set of categorical variables, and there are no interactions included in the model, logistic regression weighting is similar to raking modelling. It is however, more flexible in that it can include continuous variables, unlike raking modelling.

Kalton and Flores-Cervantes (2003) also note that using logistic regression weighting, unlike raking modelling, does not ensure that the weighted sample marginal distributions conform to the population marginal distributions. It should be noted that this is the calibration property discussed in the previous section, and in Chapter 1 we showed that the GREG-estimator is also a calibration estimator. As presented here, logistic regression weighting does not ensure

that the distribution of the auxiliary variables that are used in the model for the response probabilities are calibrated to the population. Therefore, the same variables should be included in the GREG-model described in (6.31) and (6.32).

6.4.2 The propensity score method in internet panels

The idea of using response probabilities has received much attention in survey methodological literature lately due to the introduction of the propensity score method. The propensity score method originates from evaluation studies to estimate average treatment effects and was first introduced by Rosenbaum and Rubin (1983). 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 arise due to a different composition of the treatment and the control group. For this purpose, the propensity score is introduced as a way to balance the composition of the two groups.

The idea of using the propensity score method in web surveys was introduced by Harris Interactive, for example see Taylor et al. (2001). Harris Interactive used the propensity score method to solve the problems of under-coverage and self-selection in volunteer opt-in internet panels (for a description of internet surveys see Couper, 2000).

Schonlau et al. (2004) and Lee (2006) have described the technique of the propensity score method in this perspective. It begins with the availability of a reference survey that does not have the same problems of under-coverage and self-selection, ideally this should be a probability based survey with full response. This reference survey is used as a benchmark for the internet respondents by balancing the composition of the attributes of the web respondents so that they are similar to the composition of the attributes of the reference survey respondents. Therefore, these attributes have to be measured in both surveys. Furthermore, for the method to work well these attributes have to reflect disparities between the two modes, as well as differences between internet respondents and sample elements that would not respond to an internet panel request. Duffy (2005) therefore uses behavioural, attitudinal and socio-demographic information. These attributes are sometimes called webographics, see Schonlau et al. (2007).

The propensity score is determined as the conditional probability that a sample element responds to the internet panel, given the attributes and the fact that the element responded either in the reference survey or the internet panel.

It is assumed that sample elements with approximately the same propensity score are similar with respect to the attributes that are used in estimating the propensity scores. Based on this assumption the self-selection bias and the bias due to under-coverage are removed from the internet panel.

There are some disadvantages to this method. Bethlehem (2007) shows that the variance of the estimates based on this type of surveys is dominated by the smallest of the two, i.e. survey and panel. Thus, if the sample size of the reference survey is small, the variance of the estimates for the survey items in the internet panel is large, regardless of the size of the internet panel. In addition, it is not realistic to assume that the reference survey does not experience problems with under-coverage or nonresponse. Furthermore, the panel is usually an internet survey and the reference survey telephone of face-to-face. As such, there is also a risk of differential measurement errors.

6.4.3 The response propensity method

In the context of survey response, the propensity score is defined as the conditional probability that a sample element i with characteristics $\mathbf{X}_i$ responds to the survey, i.e.

$$\rho(\mathbf{X}_i) = P(R_i = 1 | \delta_i = 1; \mathbf{X}_i) \tag{6.36}$$

The response propensity $\rho(\mathbf{X}_i)$ is in fact an estimate for the response probability ρ_i using $\mathbf{X}_i$, see also Chapter 1. Therefore, we denote the response propensity by $\hat{\rho}_i$, see (1.10). Sample elements with the same characteristics $\mathbf{X}$ have the same response propensity. These propensities can be estimated based on the sample, hence unlike the method in the previous section no reference survey is needed as long as the characteristics $\mathbf{X}$ are known for all sample elements.

There are two important assumptions under which the propensity score method can be employed. The first one is the *conditional independence assumption*[1], which states that conditional on $\mathbf{X}_i$, the survey items are independent of the response behaviour. We denote this by

$$\mathbf{Y} \perp \mathbf{R} | \mathbf{X} \tag{6.37}$$

where $\perp$ indicates orthogonality, or independence. In other words, within sub populations defined by values of the observed characteristics, there is random response. This implies that all variables that affect either the survey item or the

[1]This assumption is also known as *selection on observables*, *unconfoundedness assumption* or the *ignorable treatment assumption*

response behaviour must be observed, i.e. there is no selection on unobservables. The second assumption is the *matching assumption*. It states that

$$0 < \hat{\rho}_i < 1 \tag{6.38}$$

This assumption ensures that for each value of $\mathbf{X}$ there are both sample elements that respond and sample elements that do not respond. We cannot compare sample elements when the response probability equals 0 or 1 because for these values there are no counterparts. Rosenbaum and Rubin (1983) show that the conditional independence assumption given $\mathbf{X}_i$ implies conditional independence given $\hat{\rho}_i$. That is

$$\mathbf{Y} \perp \mathbf{R}|\mathbf{X} \Longrightarrow \mathbf{Y} \perp \mathbf{R}|\hat{\boldsymbol{\rho}} \tag{6.39}$$

The conditional independence assumption corresponds to the missing-at-random (MAR)-assumption, see section 1.3.2.

The propensity score method thus balances the response behaviour on observed auxiliary variables. Rosenbaum and Rubin (1983) propose that

$$\mathbf{X} \perp \mathbf{R}|\hat{\boldsymbol{\rho}} \tag{6.40}$$

which implies that sample elements with the same response propensity have the same distribution of the auxiliary variables used in estimation of the response probability, regardless of whether they responded or not. The attractiveness of this method lies in the fact that matching can be based on the one-dimensional response propensity instead of all the auxiliary variables used in the estimation of the response probabilities.

In the context of survey methodology, the propensity score can be defined as the conditional probability that a sample element responds to a survey. The response propensity can, for instance, be defined as the product of the different individual probabilities in the response process, i.e.

$$\hat{\rho}_i = \hat{\xi}_i \times \hat{\gamma}_i \times \hat{\theta}_i \times \hat{\varrho}_i \tag{6.41}$$

In Chapter 7 we discuss approaches to estimate the different probabilities in the response process.

The response propensities can be used to adjust for nonresponse in different ways. We distinguish between using the response propensities directly in the estimation of the survey item and using them in combination with other nonresponse adjustment methods. Directly using the response propensities can be

done in two ways, *response propensity weighting* and *response propensity stratification*. These estimators use the available auxiliary information to model the relationship between response and auxiliary variables. In addition, the relationship between the survey item and the auxiliary information can be included. This can be achieved by updating the design weights in the GREG-estimator for nonresponse, hence obtaining the *response propensity GREG-estimator*. In the next sections, we discuss these three estimators.

6.4.3.1 Response propensity weighting

Following the suggestion of Särndal (1981) as presented in (6.4), the response propensity can be inserted in the modified Horvitz-Thompson estimator which then becomes

$$\overline{y}_{ht}^{r} = \frac{1}{N} \sum_{i \in r} \frac{d_i Y_i}{\hat{\rho}_i} \tag{6.42}$$

we refer to this estimator as the *response propensity weighting* estimator. The only difference with (6.3) is that we now use the response propensity instead of the response probability.

6.4.3.2 Response propensity stratification

Another possibility of directly using the response propensities in the estimation of survey item Y is to stratify the sample based on the response propensities. Then, within strata that have the same value of $\hat{\rho}$ the probability that $R_i = 1$ does not depend on the values of $\mathbf{X}$. This conclusion follows directly from (6.40). This implies that we should define strata based on the response propensity; within these strata the response behaviour is the same for respondents and nonrespondents.

Suppose we stratify the sample into F strata based on the response propensities. Cochran (1968) suggests that it is enough to use five strata, i.e. $F = 5$. The strata are denoted by $s_1, s_2, \ldots, s_F$. We introduce for each stratum a dummy variable, denoted by $L_1, L_2, \ldots, L_F$. For an element in a certain stratum f, the corresponding dummy variable L_f is assigned the value 1, and all other dummy variables are set to 0. The sample size of stratum f is denoted by n_f. These sample sizes are random variables and not fixed numbers. Post-stratification assigns the same weight to all elements in the same stratum. The correction

weight g_i for element i in stratum f is defined as

$$g_i = \frac{n_f}{n_{r,f}} \tag{6.43}$$

where $n_{r,f}$ is the number of responding sample elements in stratum f. Consequently, the post-stratification estimator can be expressed as

$$\begin{array}{ll} \overline{y}_{ps}^{\hat{\rho}} & = \frac{1}{N}\Big(\frac{n_1}{n_{r,1}}\sum_{i\in s_1} Y_i + \frac{n_2}{n_{r,2}}\sum_{i\in s_2} Y_i + \ldots + \frac{n_F}{n_{r,F}}\sum_{i\in s_F} Y_i\Big) \\ & = \frac{1}{N}\sum_{f=1}^{F} n_{r,f}\overline{y}_r^{(f)} \end{array} \tag{6.44}$$

where $\overline{y}_r^{(f)}$ is the (unweighted) response mean for the survey item in stratum f. The estimator (6.44) is referred to as the *response propensity stratification* estimator.

6.4.3.3 The response propensity GREG-estimator

In the response propensity weighting approach, the design weights are updated for the second phase in the sampling process by dividing them by the response propensities. The available information is used to model the relationship between the response and the auxiliary variables. We can extend this approach, by regarding the relationship between the survey item and the auxiliary information as well. One possibility to do this, is by updating the weights in the GREG-estimator presented in (6.33). The g-weights based on this estimator then can be expressed as

$$g_{i\hat{\rho}} = 1 + \Big(\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht}^{r}\Big)'\Big(\sum_{i=1}^{n_r}\frac{d_i}{\hat{\rho}_i}\mathbf{X}_i\mathbf{X}_i'\Big)^{-1}\mathbf{X}_i \tag{6.45}$$

in this expression, the design weights d_i are modified for the second phase which involves nonresponse. This use of the response propensities is therefore referred to as the *response propensity GREG-estimator*.

6.5 Concluding remarks

In practice, which estimator to use is for a large part dependent on the available level of information from the auxiliary variables. In table 6.1 we present an overview of estimators together with the type of auxiliary information.

The methods in section 6.4.3 all make use of response propensities. To

compute response propensities, auxiliary information for all sample elements is needed. Using the response propensities alone does not ensure that the distribution of characteristics is calibrated to known distributions in the sample. The response propensities do not account for the survey design. Therefore, the response propensities have to be combined with calibration estimators. In the next chapter we describe an application of these methods.

The use of response probabilities in adjustment for nonresponse is not new. However, the interest in using response probabilities has increased due to the introduction of the propensity score method. This is mostly caused by the rise of internet panels and corresponding commercial research on balancing the characteristics of respondents to volunteer internet panels with the characteristics of the population of interest, see for example Taylor et al. (2001), Duffy et al. (2005) and Lee (2006).

The use of propensity scores described in this chapter is not to be confused with the propensity score method used in adjustment for undercoverage and self-selection in volunteer opt-in panels, see Taylor et al. (2001), Duffy et al. (2005) and Lee (2006). The propensity score method exploits a reference survey that has full coverage, that is based on a probability sample and that has ignorable nonresponse. The reference survey is then used as a benchmark for the internet respondents. The composition of the attributes of the web respondents is balanced so that it is similar to the composition of the attributes of the reference survey respondents. The attributes have to be measured in both surveys. Duffy et al. (2005) propose a set of behavioural, attitudinal and socio-demographic attributes. The propensity score in volunteer opt-in panels reflects the response probability with respect to the set of attributes that is shared by the panel and the reference survey.

In this chapter we have discussed estimators that use a one-way classification of qualitative and/or quantitative variables. Two-way classification leads to other methods, for example raking or proportional fitting. A description of these methods is given in, for example, Särndal and Lundström (2005), Kalton and Flores-Cervantes (2003) or Holt and Elliot (1991).

Table 6.1: *Different estimators and their corresponding auxiliary information*

Estimator	*Auxiliary vector* $\mathbf{X}$	*Population total*	*Expression*
Post-stratification	$\mathbf{X}^* = (X_1, X_2, \ldots, X_H)'$	$\sum_{i\in U} X_i = (N_1, N_2, \ldots, N_H)'$	$\frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{ht,h}^r$
	$\mathbf{X}^\circ = (X_1, X_2, \ldots, X_H)'$	$\sum_{i\in s} d_i X_i = (\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_H)'$	$\frac{1}{N}\sum_{h=1}^{H} \hat{N}_h \overline{y}_{ht,h}^r$
		$\hat{N}_h = \sum_{i\in s_h} d_i$	
Ratio	$\mathbf{X}^* = X$	$\sum_{i\in U} X_i$	$\frac{1}{N}\Big(\sum_{i\in U} X_i\Big)\frac{\sum_r d_i Y_i}{\sum_r d_i X_i}$
	$\mathbf{X}^\circ = X$	$\sum_{i\in s} d_i X_i$	$\frac{1}{N}\Big(\sum_{i\in s} d_i X_i\Big)\frac{\sum_r d_i Y_i}{\sum_r d_i X_i}$
Regression	$\mathbf{X}^* = (1, X)'$	$\sum_{i\in U} \mathbf{X}_i^* = (N, \sum_{i\in U} X_i)'$	$\overline{y}_{ht}^r + (\overline{\mathbf{X}}^* - \overline{x}_{ht}^r)\beta^{(s)}$
			$\overline{\mathbf{X}}^* = (1, \frac{1}{N}\sum_{i\in U} X_i)'$
	$\mathbf{X}^\circ = (1, X)'$	$\sum_{i\in s} \mathbf{X}_i^\circ = (\hat{N}, \sum_{i\in s} d_i X_i)'$	$\overline{y}_{ht}^r + (\overline{\mathbf{X}}^\circ - \overline{x}_{ht}^r)\beta^{(s)}$
		$\hat{N} = \sum_{i\in s} d_i$	$\overline{\mathbf{X}}^\circ = (1, \frac{1}{N}\sum_{i\in s} d_i X_i)'$
GREG	$\mathbf{X}^* = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J)'$	$\sum_{i\in U} \mathbf{X}_i =$	$\overline{\mathbf{X}}^{*\prime}\boldsymbol{\beta}^{(s)}$
		$(\sum_{i\in U} \mathbf{X}_{1i}, \ldots, \sum_{i\in U} \mathbf{X}_{Ji})'$	$\overline{\mathbf{X}}^* = \frac{1}{N}\sum_{i\in U} \mathbf{X}_i$
	$\mathbf{X}^\circ = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J)'$	$\sum_{i\in s} d_i \mathbf{X}_i =$	$\overline{\mathbf{X}}^{\circ\prime}\boldsymbol{\beta}^{(s)}$
		$(\sum_{i\in s} d_i \mathbf{X}_{1i}, \ldots, \sum_{i\in s} d_i \mathbf{X}_{Ji})'$	$\overline{\mathbf{X}}^\circ = \frac{1}{N}\sum_{i\in s} d_i \mathbf{X}_i$
Response propensity	$\mathbf{X}^\circ = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J)'$		$\hat{\rho}_i = \rho(\mathbf{X}_i)$

Chapter 7

Adjustment for Undercoverage and Nonresponse in a Telephone Survey

In Chapter 6 we gave an overview of the calibration framework for nonresponse adjustment. Some conventional estimators can be expressed within this framework as adjustment weighting estimators. In addition, we described three approaches to use response propensities in the adjustment for nonresponse bias. In this chapter we modify these approaches so that they can be applied to adjust for errors caused by undercoverage in a CATI survey.

7.1 Introduction

For its social and demographic surveys, Statistics Netherlands often favours computer-assisted personal interviewing (CAPI) to cheaper modes that employ web or telephone. Due to the persuasive power and assistance of interviewers visiting selected persons or households, nonresponse in CAPI surveys is relatively low and data quality is high. However, the costs of this mode of interviewing are relatively high. A large group of trained interviewers is required that is distributed all over the country. Although the Netherlands is a small country,

travel costs make up a considerable proportion of the total costs.

To reduce costs, Statistics Netherlands is changing some of its CAPI surveys into computer-assisted telephone surveys (CATI) and web surveys. Here, we focus on CATI. By concentrating interviewers in one call centre, a smaller group is sufficient. No more time is spent on travel, and this also means no more travel costs are involved. Although a possible change from CAPI to CATI may substantially reduce the costs of surveys, there is also a potential drawback: it may reduce the quality of the produced statistics.

There is a vast amount of literature that compares telephone surveys to face-to-face surveys on different aspects, see also Chapter 9. De Leeuw (1992) performed a meta-analysis on face-to-face and telephone surveys, and considered several aspects of data quality. She concluded that differences in data quality between well conducted face-to-face and telephone surveys are small. This conclusion is in line with Groves (1989), who states that the most consistent finding in studies comparing responses in face-to-face and telephone surveys is the lack of difference between the two modes.

In Chapter 1 we described the possible sources of error in surveys. In this chapter we restrict ourselves to non-sampling errors caused by nonresponse and coverage. We assume that measurement errors are dealt with. Also, we regard the situation where telephone numbers are linked to a sample drawn from a population register. This is the current practice at Statistics Netherlands. The sample for a CATI survey is obtained by matching the sample from the population register to the Dutch telephone company KPN for listed telephone numbers. However, links will only be established for sample elements with a listed land-line telephone. Currently, the percentage of persons with a listed land-line telephone is estimated to be between 60% and 70%. This means that there is a substantial undercoverage of 30% to 40%.

The group without a listed land-line telephone actually consists of three categories: persons that have an unlisted land-line telephone, persons with only a mobile telephone, and persons that do not have a telephone at all. In our analysis, we cannot distinguish these groups although they are likely to be very different on socioeconomic variables (see e.g. Vehovar et al., 2004 and Callegaro and Poggio, 2004).

The differences between persons with and without a listed land-line telephone have been analysed extensively. Cobben and Bethlehem (2005) find an under representation of non-Western, non-native persons and regions where the percentage of non-natives is higher than 20%. Pickery and Carton (2005) analyse the representativeness of telephone surveys in Flanders. They find the same differences in ethnic group as Cobben and Bethlehem (2005). Furthermore, they

report differences related to the educational level, the employment situation and ownership of a house. Van Goor and Rispens (2004) find differences in the Netherlands with regard to the ethnic group, the household type and the employment situation. In Finland, Kuusela (2000) finds a difference between persons with and without a listed land-line telephone in the degree of urbanization and in the size and type of the household. Ellis and Krosnick (1999) report differences in telephone ownership with respect to education, income and ethnicity.

In the Netherlands, the coverage of listed land-line telephones decreased to such an extent that it seriously threatens CATI as a single data collection mode, irregardless of other errors. This motivates the two objectives of this chapter: first, we want to assess what the effect of a change from CAPI to CATI is on the estimation of population characteristics. As we will see in section 7.4.1, the estimates based on a CATI survey are different from the estimates in a CAPI survey. Therefore the second objective is to investigate if and how we can adjust for this effect. The aim of our research is to answer the following two questions

> *What is the effect on the quality of estimates of population characteristics, when changing from a CAPI to a CATI survey and, consequently, can we adjust for this effect?*

We thereby focus on coverage errors and nonresponse bias. The methods that we use are already introduced in the previous chapter. Here, we make some adjustments to the methods to fit to the situation of coverage errors and nonresponse bias. The data for the analysis come from the Integrated Survey on Living Conditions, denoted by its Dutch acronym POLS; 'Permanent Onderzoek LeefSituatie'. This is a CAPI survey. We artificially construct a CATI survey from the CAPI survey by removing persons without a listed land-line telephone. By doing so we can analyse directly the impact of undercoverage and we do not have any mode effects in response behaviour. Both the response rate and the composition of the response may be different in telephone surveys and face-to-face surveys. We implicitly assume that the CATI response behaviour will not be very different from the CAPI response behaviour.

The outline of this chapter is as follows: in section 7.2 we describe the data from the POLS 2002 survey. In section 7.3 we shortly outline the methods that we applied to the data to answer the research questions. Section 7.4 presents the results from the analysis and section 7.5 concludes.

7.2 Description of the data

The data that we use in the analysis is obtained by aggregating the monthly POLS-surveys for the year 2002. POLS is a continuous CAPI survey. Every month a sample of $3,000$ persons is selected and interviewed face-to-face. The survey has a modular structure; there is a base module with questions for all sampled persons and in addition there are a number of modules about specific themes (such as employment situation, health and justice). The sampled persons are selected for one of the thematic modules; the questions in the base module are answered by everyone.

The target population is not the same for every module. However, all target populations consist of persons of at least age 12 and older. Persons are selected by means of a stratified two-stage sample. In the first stage, municipalities are selected within regional strata with probabilities proportional to the number of inhabitants. In the second stage, an equal probability sample is drawn in each selected municipality. In this chapter, only persons of 12 years and older are regarded. These persons all have the same first-order inclusion probability. The focus of this research lies on the questions in the base module. It is difficult to distinguish the contributions of coverage errors and nonresponse bias in the total survey error. Figure 7.1 describes the situation graphically for the POLS 2002 survey. In the ideal situation where the sampling frame exactly covers the population, the bias will only be caused by nonresponse of both persons with and without a listed telephone. CAPI is close to this situation. The percentage response is $40.1\% + 16.6\% = 56.7\%$. In case of a CATI survey, the bias is caused both by undercoverage and nonresponse. Only 40.1% of the original sample will respond.

Note that the percentage response among the listed telephones is much higher (59.4%) than for no listed telephones (51.1%). Apparently, persons without a listed telephone behave differently from persons with a listed telephone.

As we already mentioned, POLS is a CAPI survey. For our research we need both a CAPI- and a CATI survey. We therefore construct the CATI survey from the CAPI survey. We can do so by matching the sample elements in the CAPI survey to the telephone register provided by the Dutch telephone company KPN. Deleting sample elements without a listed land-line telephone provides us with the sample that would have been used had the survey been performed by CATI. An advantage of this artificial way of generating the CATI survey is that possible mode effects caused by differences in face-to-face and telephone interviewing are avoided.

The CAPI survey sample consists of $35,594$ sample elements, $24,052$ of which

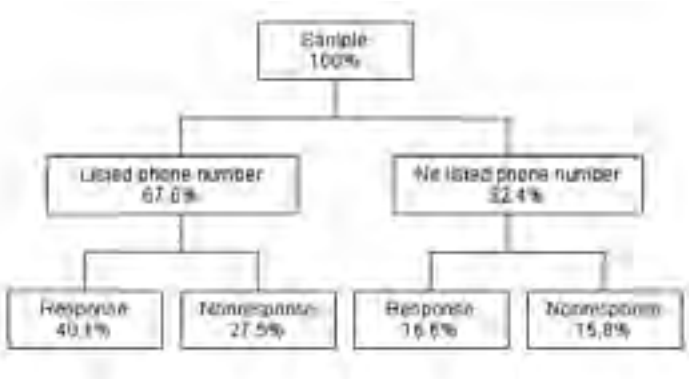

Figure 7.1: *Graphical representation of undercoverage and nonresponse in POLS 2002*

have a listed land-line telephone (67.6%). Figure 7.2 graphically displays the population and the two datasets. There are two types of variables that we use in the analysis: auxiliary variables and survey items. Auxiliary variables are available for both respondents and nonrespondents. These variables come from registers like the population register and the Centre for Work and Income (CWI). The survey items are the answers to the survey questions; these are only available for respondents. The variables that we use in our analysis are displayed in table 7.1. See Chapter 3 for a detailed description of the auxiliary variables.

Table 7.1: *Auxiliary variables and survey items in POLS 2002 survey*

Auxiliary variable	*Categories*
Gender	Male, Female
Age_2	12–34, 35–54, 55 +
Age_{15}	12–14, 15–17, ... , 70–74, 74 +
Marital status_2	Married, Not married
Marital status_4	Married, Not married, Divorced, Widowed
Ethnic group	Native, Moroccan, Turkish, Suriname, Netherlands Antilles/Aruba, other non-Western non-native, other Western non-native
Region_{15}	Province of residence and three largest cities
Region_4	North, East, South, West
Degree of urbanization	Very low, Low, Average, High, Very high
Household size	1, 2, 3, 4, 5+
Household type	Single, Couple, Couple with children, Single parent, Other
Interview month	January, February, ... , December

Continued on next page

Table 7.1: *Auxiliary variables and survey items in POLS 2002 survey - continued*

Auxiliary variable	*Categories*
Listed land-line telephone	Yes, No
Disability insurance	Yes, No
Social security	Yes, No
Average house value	Missing, 0, 0–50,000, 50,000–75,000, ... , 275,000–300,000, 300,000–350,000, > 350,000 (euro)
% non-natives in 6-digit postcode area	0–5%, 5–10%, ..., 40–50%, 50% and more
Survey item	*Categories*
Employment status	12 hours or more, less than 12 hours, unemployed
Educational level	Primary, Junior general secondary, Pre-vocational, Senior general secondary, Secondary vocational, Higher professional, University, Other
Religion	None, Roman-Catholic, Protestant, Islamic, Other

7.3 The methods

In this chapter, we slightly modify the methods from Chapter 6 so that they can be applied to adjust for coverage errors; as well as for coverage errors and nonresponse bias simultaneously. In section 7.3.1 we describe methods to adjust for undercoverage errors. Section 7.3.2 describes methods that adjust for both undercoverage errors and nonresponse bias. These methods will all be applied to the CATI response.

7.3.1 Telephone coverage propensity

Let the target population of the survey consist of N sample elements, $i = 1, 2, \ldots, N$. Furthermore, let Y denote one of the survey items described in table 7.1. The aim of the survey is to estimate the population mean

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \tag{7.1}$$

In addition, we denote by $\mathbf{X}$ a vector of auxiliary variables with values $\mathbf{X}_i$ for $i = 1, 2, \ldots, N$. The sample is selected without replacement from the population and can be represented by the N-vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_N)'$ of sample indicators, $(\delta_i = 1)$ when sample element i is selected in the sample, and $(\delta_i = 0)$ otherwise. The expected value of $\boldsymbol{\delta}$ is $E(\boldsymbol{\delta}) = \boldsymbol{\pi}$ where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)'$ is the vector

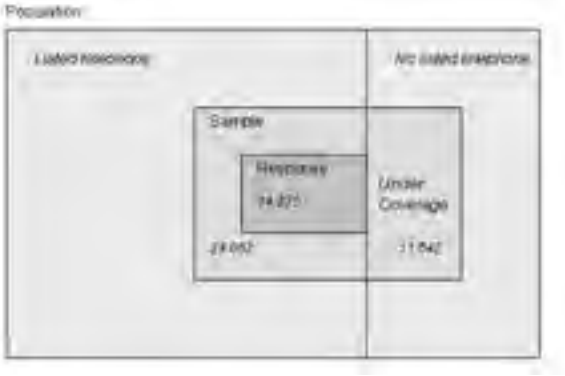

Figure 7.2: *Graphical representation of the division of the population with respect to telephone ownership*

of the first order inclusion probabilities. We denote by $d_i = 1/\pi_i$ the design weights. We assume that the values for $\boldsymbol{\pi}$ are known and nonzero. We can then estimate (7.1) without bias by the Horvitz-Thompson estimator, i.e.

$$\overline{y}_{ht} = \frac{1}{N} \sum_{i=1}^{N} \delta_i d_i Y_i \tag{7.2}$$

Now, let us assume that having a listed land-line telephone is the result of a random process. Each element i has a certain unknown probability τ_i of having a listed telephone, for $i = 1, 2, \ldots, N$. Let T denote an indicator, and $T_i = 1$ when element i is sampled and has a listed telephone, $T_i = 0$ if sample element i does not have a listed telephone. Now, the *telephone coverage propensity* is defined as

$$\tau(\mathbf{X}_i^t) = P(T_i = 1 | \delta_i = 1, \mathbf{X}_i^t) \tag{7.3}$$

where $\mathbf{X}^t = (\mathbf{X}_2^t, \mathbf{X}_2^t, \ldots, \mathbf{X}_n^t)'$ is the vector of auxiliary variables that is used to estimate the telephone coverage probability τ_i. The values τ_i are unknown. We can estimate τ_i by an appropriate method and the available auxiliary variables, thus obtaining the telephone coverage propensity $\hat{\tau}_i = \tau(\mathbf{X}_i^t)$. Some of the methods that we can use for this purpose are described in Chapter 1.

The telephone coverage propensities can be used to adjust for undercoverage, similar to the use of the estimated response propensities in nonresponse

adjustment. Therefore, we modify the response propensity estimators discussed in Chapter 6. To adjust for errors due to undercoverage in the CATI survey, we modify the propensity weighting estimator (6.42) and the propensity stratification estimator (6.44). The telephone coverage propensity weighting estimator becomes

$$\overline{y}_{ht}^{t} = \frac{1}{N} \sum_{i \in r} \frac{d_i Y_i}{\tau(\mathbf{X}_i^t)} \tag{7.4}$$

This is the adopted Horvitz-Thompson estimator for telephone coverage. To contrast this estimator with the adopted Horvitz-Thompson estimator for nonresponse $\overline{y}_{ht}^{r}$, we denote this estimator by $\overline{y}_{ht}^{t}$.

To obtain the post-stratification estimator, we stratify the sample based on the telephone coverage propensity score. Then, within strata that have the same value of $\tau(\mathbf{X}^t)$, $P(T_i = 1)$ does not depend on $\mathbf{X}^t$. We thus obtain F strata based on the telephone coverage propensities. Cochran (1968) suggests that it is enough to use five strata, i.e. $F = 5$. The strata are denoted by $s_1, s_2, \ldots, s_5$.

We introduce 5 dummy variables $X_1, X_2, \ldots, X_5$. For an element in a certain stratum f, the corresponding dummy variable X_f is assigned the value 1, and all other dummy variables are set to 0. The sample size of stratum f is equal to n_f. These sample sizes are random variables and not fixed numbers. Post-stratification assigns the same weight to all elements in the same stratum. The correction weight g_i for element i in stratum f is defined as

$$g_i = \frac{n_f}{n_{r,f}} \tag{7.5}$$

where $n_{r,f}$ is the number of responding sample elements in stratum f. Consequently, the telephone coverage propensity stratification estimator can be expressed as

$$\begin{aligned} \overline{y}_{ps}^{\tau(\mathbf{X}_t)} &= \tfrac{1}{N}\Big(\tfrac{n_1}{n_{r,1}} \textstyle\sum_{i \in s_1} Y_i + \tfrac{n_2}{n_{r,2}} \sum_{i \in s_2} Y_i + \ldots + \tfrac{n_F}{n_{r,F}} \sum_{i \in s_F} Y_i\Big) \\ &= \tfrac{1}{N} \textstyle\sum_{f=1}^{F} n_{r,f} \overline{y}_t^{(f)} \end{aligned} \tag{7.6}$$

where $\overline{y}_t^{(f)}$ is the adopted Horvitz-Thompson estimator for telephone coverage for the survey item in stratum f.

7.3.2 Simultaneous adjustment of undercoverage and nonresponse

Estimators (7.4) and (7.6) adjust for the errors caused by the undercoverage of persons without a listed land-line telephone. The question now arises how we can combine the adjustment for nonresponse and undercoverage. We use the GREG-estimator to adjust for nonresponse. We need to add the adjustment for undercoverage to this method. There are two ways to combine linear weighting and the telephone coverage propensity score method. The first approach is based on the idea of propensity score weighting and comes down to GREG-estimation with adjusted inclusion probabilities. We have described this approach in Chapter 6, (6.45), with response propensities. Here, we describe the approach with telephone coverage propensities. The second approach employs telephone coverage as an additional auxiliary variable in the weighting model for the GREG-estimator. We describe these two approaches separately. We distinguish between auxiliary variables that are used in estimating the telephone coverage propensity denoted by $\mathbf{X}^t$ and the variables that we use to adjust for nonresponse, which we will denote by $\mathbf{X}^r$.

In Chapter 6, equation (6.33), we expressed the GREG-estimator as

$$\overline{y}^r_{gr} = \sum_{i=1}^{n_r} \frac{d_i}{\hat{\rho}_i} g_{i\hat{\rho}} Y_i$$

This estimator only produces consistent estimates if the proper design weights d_i are used. Availability of data in a CATI survey is determined by both the sampling mechanism and the probability of having a listed telephone. Therefore, the d_i in (6.33) should be replaced by d_i/τ_i. Unfortunately, the τ_i are unknown, so they are replaced by their estimates $\hat{\tau}_i = \tau(\mathbf{X}^t_i)$. The g-weights based on the GREG-estimator then can be expressed as

$$g_{i\hat{\tau}} = 1 + \left(\overline{\mathbf{X}}^r - \overline{\mathbf{x}}^t_{r,ht}\right)' \left(\sum_{i=1}^{n_r} \frac{d_i}{\hat{\tau}_i} \mathbf{X}^r_i (\mathbf{X}^r_i)'\right)^{-1} \mathbf{X}^r_i \tag{7.7}$$

where $\overline{x}^t_{r,ht}$ is the adopted Horvitz-Thompson estimator for telephone coverage for the mean of the auxiliary variables $\mathbf{X}^r_i$. This method is referred to as the *telephone coverage propensity GREG-estimator*.

The other possibility is to include the telephone coverage propensity stratification variable in the weighting model. For the direct telephone coverage propensity stratification estimator we already divided the sample into 5 strata based

on the telephone coverage propensities by introducing the dummy variables $X_1, X_2, \ldots, X_5$. Now, we aggregate these variables into one ordered categorical variable $Propclass_5$ with 5 classes. Sample elements belong to class f if $X_f = 1$. This method is referred to as the *GREG-estimator with telephone coverage propensity variable*. In the following section we describe the analysis.

7.4 Analysis of the telephone POLS survey

When changing the survey interview mode from CAPI to CATI, there are two sources of missing data: nonresponse and undercoverage of sample elements with unlisted telephones, mobile only, or no telephone at all. Similar to nonresponse, undercoverage is only a threat to survey statistics if it is selective. Cobben and Bethlehem (2005) show that having a listed land-line telephone is especially selective with respect to the variables ethnic group, % non-natives and household type.

To obtain insight into the consequences of undercoverage, we apply the methods discussed in the previous sections to our data sets. First, we compare the response means of both the CAPI survey and the CATI survey. Therefore we apply the GREG-estimator based on the CAPI response, that is proposed by Schouten (2004). The variables used in the estimation are

$$\begin{aligned} &Age_{15} + Housevalue_{14} + \%Non-Natives_8 \\ &+Ethnicgroup_7 + Region_{15} + Householdtype_4 + Telephone_2 \end{aligned} \tag{7.8}$$

The subscripts denote the number of categories. Model (7.8) is applied to both surveys. When applying the model to the CATI survey, the variable $Telephone_2$ becomes redundant. The CAPI survey estimates are used as a benchmark to judge the quality of the survey estimates in the CATI survey. Therefore we need the CAPI estimates with their best nonresponse adjustment. The weighting model is applied to the CATI survey to see whether the survey estimates are biased if we do not account for undercoverage. Subsequently, we apply the direct telephone coverage propensity weighting estimator (7.4) and - stratification estimator (7.6) to the CATI survey to adjust for undercoverage.

And finally, the combination of linear weighting for nonresponse and the telephone coverage propensity is applied to the CATI survey to simultaneously adjust for nonresponse and undercoverage. For the telephone coverage propensity GREG-estimator described by (7.7) the variables for the nonresponse adjustment $\mathbf{X}^r$ are the same variables as in the GREG-estimator described in (7.8) applied to the CATI survey, thus without the variable $Telephone_2$. For

the GREG-estimator with telephone coverage propensity variable the weighting model becomes

$$\begin{array}{l} Age_{15} + Housevalue_{14} + \%Non - Natives_{8} \\ +Ethnicgroup_{7} + Region_{15} + Householdtype_{4} + Propclass_{5} \end{array} \tag{7.9}$$

The results are discussed in the following sections.

7.4.1 Nonresponse adjustment

In table 7.2 we show the response means and the adjusted estimates when applying the nonresponse adjustment; weighting model (7.8). In case the CATI survey estimates and CAPI survey estimates are similar, switching from CAPI to CATI does not introduce an additional bias. Otherwise, the selective undercoverage in CATI can not be ignored and has to be adjusted as well.
The standard errors are given in parentheses. For the response means, the standard errors are calculated by $\sqrt{p(1-p)/n}$, with p the percentage of the survey item category and n the number of respondents ($n = 20,168$ for CAPI and $n = 14,275$ for CATI). For the weighted estimates in columns 4 and 5, the standard errors are calculated by a first order Taylor approximation of the linear regression estimator.

The second and third column in table 7.2 display the unweighted response means for both surveys. There are quite some differences. Respondents with a listed, land-line telephone tend to work less, have a higher education and are more often non-religious than respondents without a listed land-line telephone. However, it is possible that adjusting for nonresponse also handles the selectivity of telephone coverage and reduces these differences. To see if this is indeed so we apply weighting model (7.8) to both surveys. The results are displayed in columns four and five in table 7.2.

To assess the quality of the adjusted estimates based on CATI, we compare these with the estimates from the CAPI survey. We assume that the results from weighting the CAPI survey are unbiased, i.e. equal to the true population characteristics. Table 7.2 shows that the estimates from CATI and CAPI are different. The CATI survey estimates tend to be adjusted in the right direction, but then over- or underestimate the characteristic when compared to the CAPI survey estimates. For instance the response mean of the percentage of persons that work 12 hours or more for the telephone survey is 51.8%. The CAPI survey estimate is 53.7%. The CATI survey estimate indeed increases the response mean but overestimates the CAPI survey estimate by 0.7%. The differences

Table 7.2: *Comparison of unweighted and weighted response means for the CAPI and the CATI survey (in %)*

Variable	*Response mean CATI*	*Response mean CAPI*	*GREG CATI*	*GREG CAPI*
Employment				
12 hours or more	51.8 (0.42)	52.4 (0.35)	54.4 (0.27)	53.7 (0.18)
Unemployed	7.0 (0.22)	6.7 (0.18)	6.6 (0.16)	6.4 (0.11)
Less than 12 hours	41.2 (0.41)	40.9 (0.35)	38.9 (0.25)	39.9 (0.17)
Education				
Primary	6.8 (0.21)	7.2 (0.18)	6.1 (0.13)	6.3 (0.09)
Junior general secondary	12.2 (0.27)	12.1 (0.23)	11.9 (0.22)	12.0 (0.15)
Pre-vocational	19.6 (0.33)	19.7 (0.28)	19.6 (0.27)	19.9 (0.18)
Senior general secondary	6.8 (0.21)	7.1 (0.18)	7.1 (0.18)	7.2 (0.12)
Secondary vocational	30.8 (0.39)	30.8 (0.32)	31.2 (0.31)	31.1 (0.21)
Higher professional	17.3 (0.32)	16.7 (0.26)	17.5 (0.26)	16.9 (0.17)
University	6.3 (0.20)	6.3 (0.17)	6.5 (0.17)	6.4 (0.11)
Other	0.2 (0.04)	0.2 (0.03)	0.2 (0.04)	0.2 (0.02)
Religion				
None	36.3 (0.40)	37.7 (0.34)	37.8 (0.32)	38.5 (0.21)
Roman-Catholic	35.4 (0.40)	33.5 (0.33)	32.7 (0.28)	32.4 (0.19)
Protestant	22.7 (0.35)	21.0 (0.29)	21.0 (0.25)	20.4 (0.17)
Islamic	1.2 (0.09)	2.5 (0.11)	3.2 (0.10)	3.3 (0.06)
Other	4.5 (0.17)	5.2 (0.16)	5.2 (0.16)	5.4 (0.01)

may seem minor, but 0.7% of 14 million persons (the population aged 12+) still are approximately 100,000 persons.

These differences are remarkable since model (7.8) incorporates the variables that cause the largest selectivity (Ethnic group, % non-natives and Region). The results suggest that answers to the survey questions for sample elements with a listed telephone on average are different than those for elements without a listed telephone, even when accounting for the variables that are correlated with telephone ownership. We thus have to account for the selective undercoverage in the CATI survey. Therefore, we applied the adjustment methods using the telephone coverage propensity discussed in section 7.3.1.

7.4.2 Undercoverage adjustment

We model the propensity score by means of a logit model as described in Chapter 5, appendix 5.A. Other models can be used too, but e.g. Dehija and Wahba (1999) conclude that different models often produce similar results. Using the CAPI survey sample, the propensity scores are modelled with the software package Stata. By stepwise excluding insignificant variables, the variables $\mathbf{X}^t$ in the final model are

$$\begin{aligned} &\%Non-Natives_8 + Region_{15} + EthnicGroup_7 + Urbanisation_5 + \\ &MaritalStatus_2 + HouseholdType_4 + HouseValue_{14} + Age_3 + \\ &DisabilityAll_2 + SocialAll_2 \end{aligned} \tag{7.10}$$

The subscripts denote the number of categories. We estimate the model parameters by Maximum Likelihood Estimation. The value of the pseudo R^2 for this model is 9.1%. This is rather low, which is an indication that there still is a lot of unexplained variance in this model.

Based on this model, the telephone coverage probabilities can be estimated. We then use the propensities $\hat{\tau}_i$ to adjust for errors due to undercoverage as described in section 7.3.1. The results are displayed in table 7.3.
The standard errors are given in parentheses. The first two columns are again the response mean for CATI and CAPI respectively, like in table 7.2. For the telephone coverage propensity methods in columns 4 and 5, the standard errors are calculated in R by non-parametric bootstrap estimation, see Chapter 5, with $B = 1,000$ bootstraps.

To see how the techniques perform, the results from telephone coverage propensity weighting and -stratification are compared to the response mean from the CAPI survey. Columns 2 and 3 display the response means for the CATI and

Table 7.3: *Estimates for the CATI survey based on telephone coverage propensity weighting and stratification*

Variable	*Response mean CATI*	*Response mean CAPI*	*Propensity weighting*	*Propensity stratification*
Employment				
12 hours or more	51.8 (0.42)	52.4 (0.35)	52.2 (0.50)	52.1 (0.48)
Unemployed	7.0 (0.22)	6.7 (0.18)	7.0 (0.24)	7.0 (0.22)
Less than 12 hours	41.2 (0.41)	40.9 (0.35)	40.7 (0.49)	40.9 (0.45)
Education				
Primary	6.8 (0.21)	7.2 (0.18)	7.1 (0.24)	7.1 (0.21)
Junior general secondary	12.2 (0.27)	12.1 (0.23)	12.2 (0.32)	12.2 (0.28)
Pre-vocational	19.6 (0.33)	19.7 (0.28)	19.4 (0.39)	19.5 (0.34)
Senior general secondary	6.8 (0.21)	7.1 (0.18)	6.9 (0.25)	6.9 (0.22)
Secondary vocational	30.8 (0.39)	30.8 (0.32)	30.6 (0.44)	30.7 (0.40)
Higher professional	17.3 (0.32)	16.7 (0.26)	17.1 (0.36)	17.2 (0.33)
University	6.3 (0.20)	6.3 (0.17)	6.3 (0.23)	6.3 (0.20)
Other	0.2 (0.04)	0.2 (0.03)	0.2 (0.04)	0.2 (0.04)
Religion				
None	36.3 (0.40)	37.7 (0.34)	37.2 (0.46)	37.0 (0.43)
Roman-Catholic	35.4 (0.40)	33.5 (0.33)	33.5 (0.46)	33.8 (0.41)
Protestant	22.7 (0.35)	21.0 (0.29)	21.5 (0.39)	21.7 (0.37)
Islamic	1.2 (0.09)	2.5 (0.11)	2.7 (0.11)	2.5 (0.09)
Other	4.5 (0.17)	5.2 (0.16)	5.1 (0.21)	5.1 (0.17)

CAPI survey respectively. In columns 4 and 5, the results from telephone coverage propensity weighting and -stratification are shown. Both telephone coverage propensity weighting and stratification perform well. The estimates based on the CATI survey are very close to the CAPI response means. Stratification based on the telephone coverage propensities has slightly lower standard errors than weighting. The method appears to be successful in adjusting for the selective undercoverage in CATI surveys.

7.4.3 Simultaneous adjustment

Now we can proceed to adjust the final estimates for nonresponse bias as well as coverage errors. We therefore apply the methods presented in section 7.3.2. The results are displayed in table 7.4.

The standard errors are given in parentheses and calculated by a first order Taylor approximation of the linear regression estimator. For the methods using the telephone coverage propensity in columns 3 and 4, we make the assumption that these propensities are deterministic values that introduce no additional variance.

The results are again compared to the estimates from the CAPI survey, displayed in the second column. The third and fourth column display the results from adjusting the CATI survey for nonresponse and undercoverage simultaneously with respectively the GREG-estimator for nonresponse adjustment with an additional telephone coverage propensity variable and the GREG-estimator to adjust for nonresponse bias with adjusted inclusion probabilities for telephone coverage presented in section 7.3.2.

Both methods seem to perform well. The standard errors for the combination of linear weighting and adjusted inclusion probabilities for telephone coverage are larger than for linear weighting including an additional telephone coverage propensity stratification variable. For the variables educational level and religion, the estimates for both methods are very similar. With respect to employment situation, the estimates based on linear weighting with an additional telephone coverage propensity stratification variable are closer to the benchmark estimates from the CAPI survey. The best adjustment technique to reduce the bias caused by telephone interviewing for this case, appears to be linear weighting including an additional telephone coverage propensity stratification variable.

Table 7.4: *Adjustment of the CATI survey estimates for both coverage errors and nonresponse bias*

Variable	*GREG-estimator CAPI*	*GREG-estimator + propensity variable*	*Propensity GREG-estimator*
Employment			
12 hours or more	53.7 (0.18)	54.0 (0.27)	54.8 (0.35)
Unemployed	6.4 (0.11)	6.7 (0.16)	6.6 (0.20)
Less than 12 hours	39.9 (0.17)	39.3 (0.26)	38.7 (0.33)
Education			
Primary	6.3 (0.09)	6.1 (0.14)	6.1 (0.17)
Junior general secondary	12.0 (0.15)	11.9 (0.22)	11.9 (0.29)
Pre-vocational	19.9 (0.18)	19.6 (0.27)	19.6 (0.36)
Senior general secondary	7.2 (0.12)	7.1 (0.18)	7.0 (0.23)
Secondary vocational	31.1 (0.21)	31.3 (0.31)	31.3 (0.41)
Higher professional	16.9 (0.17)	17.5 (0.26)	17.5 (0.34)
University	6.4 (0.11)	6.5 (0.17)	6.4 (0.22)
Other	0.2 (0.02)	0.2 (0.04)	0.2 (0.04)
Religion			
None	38.5 (0.21)	37.8 (0.32)	37.9 (0.42)
Roman-Catholic	32.4 (0.19)	32.7 (0.28)	32.8 (0.37)
Protestant	20.4 (0.17)	21.0 (0.25)	21.1 (0.34)
Islamic	3.3 (0.06)	3.2 (0.11)	3.1 (0.13)
Other	5.4 (0.01)	5.3 (0.17)	5.2 (0.21)

7.4.4 Summary

In this chapter we consider the influence of the data collection mode on errors related to undercoverage and nonresponse, and compare various techniques that aim at adjusting for the bias caused by these errors. These techniques are partly based on the GREG-estimator, and partly on using propensity scores. The aim of our research is to answer the two following questions: What is the effect on the quality of estimates of population characteristics, when going from a CAPI to a CATI survey and, consequently, can we adjust for this effect?

We explore to what extent adjustment techniques can reduce the bias caused by telephone interviewing. First, the CATI survey is adjusted for undercoverage of sample elements without a listed land-line telephone. Two methods are used: Telephone coverage propensity weighting and -stratification. No nonresponse bias is considered yet and the results are compared to the unweighted response mean of the CAPI survey. Second, the nonresponse bias is taken into account as well. Two combinations of the GREG-estimator and the telephone coverage propensities are considered.

A comparison of the response means in the CATI- and the CAPI survey shows that indeed these two surveys differ in the population characteristics of interest. Respondents that own a listed land-line telephone tend to work less, have a higher education and are more often non-religious than respondents that do not own a listed land-line telephone. Ignoring the undercoverage in the CATI survey, i.e. adjusting for nonresponse only, does not take away all the bias. We use the telephone coverage propensities to adjust for undercoverage in the CATI survey. Both telephone coverage propensity weighting and -stratification perform well. The estimates based on the CATI survey are very close to the CAPI response means. Stratification based on the telephone coverage propensities has slightly lower standard errors than weighting. Both methods appear to be appropriate to adjust for selective undercoverage in CATI surveys.

Subsequently, we simultaneously adjust for undercoverage and nonresponse bias. Therefore we applied two combinations of the GREG-estimator for nonresponse and the telephone coverage propensity score method. The best adjustment technique to reduce the bias caused by telephone interviewing for this case, appears to be the GREG-estimator with the inclusion of an additional telephone coverage propensity stratification variable. However, the ultimate estimates are still biased. There seems to be a relationship between telephone ownership and the questions in the survey that we cannot explain with the available auxiliary variables. We lack variables that are sufficiently informative to explain telephone ownership. This leads to biased estimates, especially for those survey questions

that are related to education and income. The variables that we try to estimate are actually the variables that we would like to use in our model.

7.5 Concluding remarks

The proportion of explained variance of the logit model that is used to estimate the telephone coverage propensities is only 9.1%. Despite this low level of explained variance, the telephone coverage propensity method still is able to adjust for selective undercoverage in the CATI survey. However, more research is needed to determine the importance of the model fit when using estimated propensities in the methods described in section 7.3.

The steady decrease in coverage of listed land-line telephones raises the question whether CATI, as conducted at Statistics Netherlands, is still a viable single data collection mode. The results of our research indicate that the omission of sample elements without a listed land-line telephone implies a bias for certain survey topics, like income and education, that cannot be adjusted for sufficiently. However, in a mixed mode data collection design the undercoverage of persons without a telephone can be compensated by another mode and, hence, CATI can still be a very important mode in a mixed mode design. A necessary condition for mixed mode data collection is that the survey questionnaire is cognitive equivalent for multiple modes, see Chapter 9. In that case, the advantages of CATI can be retained while at the same time the disadvantage of undercoverage can be overcome. The use of CATI in a mixed mode design has been analysed but deserves more attention, especially with respect to the coverage problem.

Chapter 8

Analysis and Adjustment Methods with Different Response Types

In Chapter 6 we presented an overview of nonresponse adjustment methods. The methods that we discussed ignore some important features of the response process. No distinction is made between different response types, for example contact and participation. In this chapter, we describe methods for the analysis of nonresponse and nonresponse adjustment that do account for the sequential nature of the response process by including different response types.

8.1 Introduction

The adjustment techniques that we discussed in Chapter 6 do not distinguish between different response types. As we showed in Chapter 3, different types of response are caused by distinctive phenomena. Furthermore, the effect of the response types on the bias may vary. This can be seen by looking at the nonresponse bias for a survey item Y, see equation (1.17). When we restrict ourselves to the two main nonresponse types, this bias can be attributed to two causes; one cause is non-contact and another is refusal. Let us introduce a nesting of these two parts by noticing that sample elements can only refuse once contacted, and non-contacted elements either would have refused, or participated would they

have been contacted. Equation (1.17) then becomes

$$\begin{aligned}\overline{Y}_r = \quad & \overline{Y}_n + \tfrac{n_{nc}}{n}\big(\tfrac{n_{r|nc}}{n}(\overline{Y}_r - \overline{Y}_{r|nc}) + \tfrac{n_{rf|nc}}{n}(\overline{Y}_r - \overline{Y}_{rf|nc})\big) \\ & + \tfrac{n_c}{n}\tfrac{n_{rf|c}}{n_c}(\overline{Y}_r - \overline{Y}_{rf|c})\end{aligned} \tag{8.1}$$

where $\overline{Y}_r$ is the respondent mean for survey item Y, $\overline{Y}_n$ is the total sample mean, $\frac{n_{nc}}{n}$ is the fraction of non-contacts in the sample. These non-contacts can be divided into sample elements that would have responded when contacted, the number of which is denoted by $n_{r|nc}$, and sample elements that would have refused participation once contacted, denoted by $n_{rf|nc}$. $\overline{Y}_{r|nc}$ is the mean for the first group, and $\overline{Y}_{rf|nc}$ for the second. Participation and refusal are conditional on being contacted. The last term in equation (8.1) reflects this nesting of refusals within response, with n_c the number of contacted sample elements, $n_{rf|c}$ the number of refusals among the contacted elements and $\overline{Y}_{rf|c}$ the mean for survey item Y for the sample elements that refused participation.

If the different groups display a different relationship with Y, for instance if $\overline{Y}_{r|nc} \geq \overline{Y}_r$ and $\overline{Y}_{rf|nc} \leq \overline{Y}_r$ the effects may cancel out. When the composition of the nonresponse varies over years, and the nonresponse types are not accounted for in the adjustment for nonresponse bias, the estimates for the survey items may differ due to a change in the composition of the response instead of real changes in society.

The literature describes several methods that take into account the different nonresponse types in the analysis of nonresponse. Durrant and Steele (2007) (see also Chapter 2) use a multilevel, multinomial logistic regression model for contact and participation. The multinomial model allows for different variables and coefficients in the equations for the distinct response types. A model is fitted for all households, for both contact and participation. However, this model does not allow for correlation between the response types. In addition, it does not take into account that households can only participate when they are contacted first. Multilevel models are used to analyse interviewer effects. This type of model accounts for the effect of clustering sample elements within interviewers. By combining the multilevel model with the multinomial model, as Durrant and Steele (2007) have done, it becomes possible to allow for a correlation between the unobserved interviewer influences on both contact and participation.

Lepkowski and Couper (2002) employ a nested logit model to analyse contact, and participation conditional on contact. The nested logit model does account for the sequential nature of the response process by separately modelling contact and participation, where participation is only modelled for those house-

holds that were successfully contacted. The analysis of the ELFS in Chapter 3 uses a similar nested logistic regression technique.

Both the multinomial model and the nested logit model assume independence between the response types. It is, however, more realistic to assume that the different response probabilities, for instance the contact probability and the participation probability, are dependent. A dependency between response types can be caused by misclassification. This occurs, for example, when households that have a low participation probability pretend to be absent (or plan to be absent, in case of an appointment) to avoid being confronted with the survey request. These households will be classified as non-contacts while in fact they refuse participation. Another reason for a dependency between processes arises when there are unobserved variables that affect different propensities in the response process.

The nested logit model accounts for the sequential nature of the response process by modelling each of the distinct stages using only the observed indicators for each response type. This allows the equations to be estimated separately. However, unobserved variables are not accounted for in the models. As a result, the underlying propensities are not defined if their corresponding indicators are missing. Nicoletti and Peracchi (2005) introduce a model that does define the propensities, also when the corresponding indicators are not observed, by explicitly modelling the censoring of the indicators. The model that they use is based on the bivariate probit model with sample selection (Van de Ven and Van Praag, 1981).

Groves and Couper (1998) present nested models that distinguish between contact and participation. They make the assumption that, conditional on the available background information, participation and contact are independent processes. This assumption is referred to as the conditional independence assumption and it corresponds to the missing-at-random assumption described in Chapter 1. The approach of Groves and Couper (1998) is aimed at constructing weights for the estimation of survey items in which the nonresponse due to non-contact and refusal is accounted for. They employ a weighted logistic regression technique to update the design weights, i.e. the inverse inclusion probabilities. They carry out a weighted logistic regression twice, first to adjust the design weights for nonresponse due to non-contact. In the second weighted logistic regression model they model the participation stage, where the initial design weights are replaced by the updated weights from the first model. The final weights reflect the selection in the two stages of the response process. A similar approach is described by Iannacchione (2003). The sequential weight adjustment approach, however, does not take into account the relationship between

the response behaviour and the survey items.

The survey items can be introduced into the bivariate probit model with sample selection described by Nicoletti and Peracchi (2005), if the survey item is bivariate. In case of a continuous survey item, the sample selection model proposed by Heckman (1979) can be applied. In fact, the bivariate probit model with sample selection is an adapted version of the sample selection model.

This chapter is outlined as follows. In section 8.2 we describe models to analyse nonresponse. The nested logit model and the bivariate probit model with sample selection are discussed. We focus on the difference between these models with respect to the sequential nature of the response process and the correlation between the different response types. These models can be refined to include interviewer effects by means of the multilevel model. We describe this approach in section 8.2.6. In section 8.3 we discuss how these nonresponse analysis models can be used to adjust for selective nonresponse. We describe the sequential procedure to adjust the design weights for the different types of nonresponse. Additionally, we describe the sample selection model proposed by Heckman (1979). We discuss issues regarding the identification and the estimation of this model. We present a number of methods to estimate the sample selection model. Furthermore, we discuss how these models can be extended to account for the different stages in the response process and to adjust the survey items for nonresponse bias at the same time. Section 8.4 ends this chapter with conclusions and suggestions for further research.

8.2 Nonresponse analysis models

8.2.1 Introduction

It is generally acknowledged in the survey literature that it is important to distinguish between different types of response. We already discussed this in Chapter 2, and demonstrated the distinct processes in the ELFS in Chapter 3. Moreover, an important aspect of the response process is its sequential nature. The aim of this section is to provide an overview of the different models that are being used to analyse nonresponse. We discuss to what extent the models distinguish between different response types, and how they account for the sequential nature of the response process.

We present the nested logit model (section 8.2.2) and the bivariate probit model with sample selection (section 8.2.3). The nested logit model and the bivariate probit with sample selection focus on the analysis of nested response

processes, whereby they allow for different assumptions regarding the dependence of the various processes. It is possible to extend these models to allow for interviewer effects. A model that is particularly useful for analysing interviewer effects is the multilevel model, which we introduce in section 8.2.6. It is, however, less straightforward to change the underlying assumptions of the model regarding the correlation between the processes.

Why do we want to allow for a correlation between the different processes? There are a number of reasons. In general, a correlation between processes arises when there are factors that affect both processes, but these factors are not included in the model. These factors are often referred to as unobservables and their influence can be accounted for by allowing the equations to be correlated. Another reason for the processes to be correlated arises due to misclassification. For instance, if a sample element does not want to participate in the survey and pretends to be absent when contacted, the sample element will be misclassified as a noncontact whereas in fact it is refusing to participate.

In the following three sections we describe the models. In section 8.2.7 we summarise the advantages and the disadvantages of the models.

8.2.2 The nested logit model

In figure 8.1 (see also Chapter 2) the response process is displayed as a sequence of different events (with corresponding probability): being processed (probability ξ), contacted (γ), able to participate (θ) and willing to participate (ϱ). Not every sample element passes through every stage of the response process. Whether or not sample elements proceed to a next stage depends on the outcome of the preceding stage. In table 8.1 we repeat the indicators presented in Chapter 1 (table 1.1) for the different stages: being processed U, contacted C, able to participate L and willing to participate P. These indicators are equal to 1 in case a sample element proceeds to the next stage (or participates in the survey), and zero if they drop out of the response process. Furthermore, if a sample element dropped out in an earlier stage, the subsequent indicators are not observed. Since the other characteristics for the sample elements are known, this missingness is referred to as *censoring*. In the nested logit model the underlying probabilities are not defined for the censored indicators. Other models, for instance the bivariate probit model with sample selection that we describe in section 8.2.3, assume that although the indicators are censored, the underlying probabilities are still defined. In Chapter 3 we analysed the nonresponse to the ELFS using logistic regression models for the different processes. The general

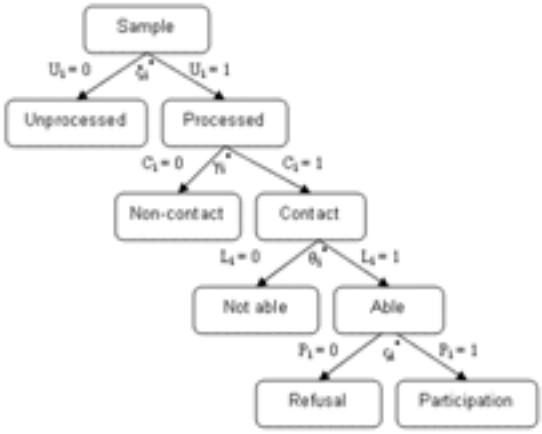

Figure 8.1: *Nested representation of the response process*

Table 8.1: Probabilities in the response process

Process	Indicator	Values	Probabilities
Being processed	U	1 processed, 0 unprocessed	ξ
Contact	C	1 contact, 0 no contact	γ
Being able (language)	L	1 able, 0 not able	θ
Participation	P	1 participation, 0 refusal	ϱ

logistic model for a probability ρ_i and corresponding indicator R_i is

$$logit(\rho_i) = \mathbf{X}_i^R \boldsymbol{\beta}^R \tag{8.2}$$

for $i = 1, \ldots, n$. $\mathbf{X}^R$ is a $n \times J$-matrix of variables, n is the number of sample elements for which the process R is defined, J the number of variables in the model and $\boldsymbol{\beta}^R$ is a $J \times 1$-vector of coefficients corresponding to the variables in $\mathbf{X}^R$. See also Chapter 5, appendix 5.A for a description and a discussion of models to estimate the response probability ρ. The estimated probabilities are referred to as propensities as described in Chapter 1, section 1.3.4.1. For example, the process probability ξ is estimated using $\mathbf{X}_i^U$ and the (estimated) process propensity is therefore denoted by $\hat{\xi}$.

We introduce the nested logit model by describing the models of the nonresponse analysis in the ELFS, presented in Chapter 3. The different stages in the

response process of the ELFS sample are analysed on the subset of observations that is available for each of the processes. For the process propensity $\xi(\mathbf{X}_i^U)$, the logit model is

$$logit(\xi(\mathbf{X}_i^U)) = \mathbf{X}_i^U \boldsymbol{\beta}^U \tag{8.3}$$

for $i = 1, \ldots, n^u$ with $n^u = n$, the number of households in the sample. This is the first stage in the response process, therefore for every household i in the ELFS sample the indicator U is defined. However, only households i for which $U_i = 1$ proceed to the next stage and are being contacted or not. For these households, the logit model for the contact propensity $\gamma(\mathbf{X}_i^{C|U=1})$ is defined as

$$logit(\gamma(\mathbf{X}_i^{C|U=1})) = \mathbf{X}_i^{C|(U=1)} \boldsymbol{\beta}^{C|(U=1)} \tag{8.4}$$

for $i = 1, \ldots, n^c$. The superscript $C|(U = 1)$ indicates that the households are being contacted, conditional on being processed and n^c is the number of households for which the contact indicator C is defined, thus the number of households for which $U_i = 1$. Again, households i for which $C_i = 1$ proceed to the next stage, where it is evaluated whether the household is able to participate or not. The logit model for the propensity to be able to participate, is

$$logit(\theta(\mathbf{X}_i^{L|C=1,U=1})) = \mathbf{X}_i^{L|(C=1,U=1)} \boldsymbol{\beta}^{L|(C=1,U=1)} \tag{8.5}$$

for $i = 1, \ldots, n^l$. n^l is the number of households for which the indicator L is defined. And finally, households i for which also $L_i = 1$ can decide whether to participate in the survey, or not. The corresponding logit model for the propensity $\varrho(\mathbf{X}_i^{P|L=1,C=1,U=1})$ is

$$logit(\varrho(\mathbf{X}_i^{P|L=1,C=1,U=1})) = \mathbf{X}_i^{P|(L=1,C=1,U=1)} \boldsymbol{\beta}^{P|(L=1,C=1,U=1)} \tag{8.6}$$

for $i = 1, \ldots, n^p$ where n^p is the number of households for which the participation indicator P is defined.

With these models, we obtain the propensities in the response process, i.e. $\hat{\xi}, \hat{\gamma}, \hat{\theta}$ and $\hat{\varrho}$. Passing through the response process, the number of observations decreases, i.e. $n^p < n^l < n^c < n^u = n$. In addition, the number of conditions under which the stages in the process are being analysed increases. The different stages in the response process are nested. Therefore, the combination of these models is referred to as a *nested logit model*. The assumption underlying the nested logit model is that the different response types are not correlated. As a consequence, equations (8.3) to (8.6) can be estimated separately.

8.2.3 Bivariate probit model with sample selection

In Chapter 5, appendix 5.A, we described logit and probit models for the estimation of response propensities. It does not really matter what functional form is used: probit or logit. In fact, even the linear model can be used as an approximation. The logistic distribution can be approximated by a normal distribution with mean 0 and variance $\frac{\pi}{3}$, the logistic distribution only has slightly more weight in the tails. In the former sections we used the logistic model to estimate the propensities in the response process. The logit model has as an advantage that the cumulative distribution function is more simple to evaluate, whereas the probit involves an unevaluated integral, see Chapter 1. Furthermore, the inverse transformation of the logit model is interpretable as the log-odds, whereas the inverse of the probit transformation does not have a direct interpretation. The estimated coefficients $\hat{\boldsymbol{\beta}}$ will be different in the two models, but roughly proportional so that the probit and the logit model produce rather similar results (see also example 8.2.3.1).

An essential difference between the two models that is important in the context of nonresponse, is the specification of the error term. In the interpretation of the latent variable regression model, the error term of the logistic model follows a logistic distribution. The error term in the probit model follows a standard normal distribution. The different distributions of the error term lead to different assumptions. The conventional logistic distribution function of the logit model, does not include a correlation between the different logit models in the nested logit. The error term distribution has to be customised to allow for a correlation with other error terms, but in the standard case there is none. Consequently, the equations in the nested logit model are independent from each other. In the probit model, the error terms follow a joint distribution that does allow for a correlation between the equations. The difference with the nested logit model, is that the propensities are now defined for all the households, regardless of whether or not the corresponding indicators are observed. In the nested logit model, the households for which the indicators in one stage of the response process are not observed, are not being used in that particular stage, nor are they being used in the following stages because by definition they are not observed there. In order to use these households, and estimate the underlying propensities even when no information is observed, these households can be regarded as censored, conditional on the outcome of the preceding stage (or stages). This censoring can be regarded as a form of sample selection. A model that can be used to introduce a correlation between the different response types, and that accounts for the censoring of unobserved households, is the bivariate

probit model with sample selection (Van de Ven and Van Praag, 1981).

The bivariate probit model with sample selection is also used by Nicoletti and Peracchi (2005). They apply the model to predict future contact and participation propensities in a panel survey. Jenkins et al. (2006) apply an extended version of the model to analyse patterns of consent, i.e. patterns in respondents' agreement to link additional data to the survey. They use a multivariate probit model for four bivariate outcomes, where two of the outcomes are subject to incidental truncation. Truncation occurs if no information at all is observed, i.e. no auxiliary variables nor the dependent variable, whereas censoring implies that only the dependent variable is not observed. Let us first consider the simplified response process that consists of the two stages: contact and participation, see figure 8.2. The bivariate model for this situation is described in terms of the latent variable regression models for the contact-, respectively the participation propensity, $\hat{\gamma}$ respectively $\hat{\varrho}$. The bivariate probit model is defined as

$$\begin{array}{rcl} \gamma_i^* = & \mathbf{X}_i^C \boldsymbol{\beta}^C + \epsilon_i^C \\ \varrho_i^* = & \mathbf{X}_i^P \boldsymbol{\beta}^P + \epsilon_i^P \end{array} \tag{8.7}$$

for $i = 1, \ldots, n$. Note that in the bivariate probit model the propensities are defined for all n, contrary to the nested logit model, where the participation propensity is only defined for contacted sample elements. We, however, do not observe the latent variables γ^* and ϱ^*, but we only observe the corresponding indicators C respectively P that can be defined as

$$C_i = \begin{cases} 1, & \text{if } \gamma_i^* > 0 \\ 0, & \text{otherwise} \end{cases} \tag{8.8}$$

$$P_i = \begin{cases} 1, & \text{if } \varrho_i^* > 0 \\ 0, & \text{otherwise} \end{cases} \tag{8.9}$$

The sample selection that occurs due to censoring of sample elements is introduced in the model by acknowledging that additionally, we only observe P_i if $C_i = 1$, i.e. if $\gamma_i^* > 0$.

To estimate this model, a distributional assumption on the error terms has to be made. A conventional distribution is the bivariate standard normal distribution, i.e.

$$\begin{pmatrix} \epsilon_i^C \\ \epsilon_i^P \end{pmatrix} \backsim N\left(\mathbf{0}, \begin{bmatrix} 1 & \zeta_{cp} \\ \zeta_{pc} & 1 \end{bmatrix}\right) \tag{8.10}$$

where $\zeta_{cp} = \zeta_{pc}$ is the correlation between the contact- and the participation equation. The variances are set equal to 1 because otherwise the model is only

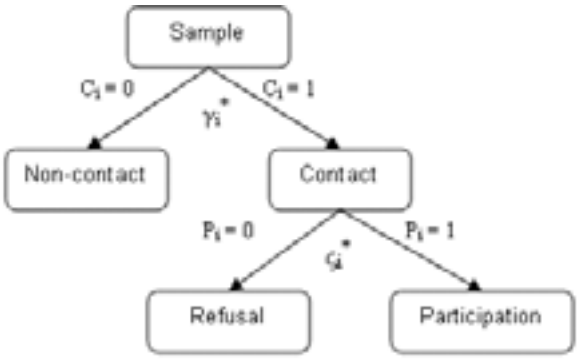

Figure 8.2: *Latent contact- and participation propensities*

identifiable up to scale, see Dubin and Rivers (1989). Van der Klaauw and Koning (2000) derive a test for the normality assumption. If the assumption is not correct, the parameters of the model are estimated with less precision. However, Van der Klaauw and Koning (2002) show that the increase in variance is little. Furthermore, their research also provides an alternative density function if normality is rejected.

The sample selection can be easily extended to more equations. The sequential nature of the response process is then reflected by the increasing degree of censoring. In the response process with four stages, C, L and P are censored. C is observed when $U = 1$, L is observed when $U = 1$ and $C = 1$ and P is only observed when $U = 1$, $C = 1$ and $L = 1$. The error terms follow a $N_4(\mathbf{0}, \mathbf{\Sigma})$ distribution, where

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \zeta_{uc} & \zeta_{ua} & \zeta_{up} \\ \zeta_{cu} & 1 & \zeta_{ca} & \zeta_{cp} \\ \zeta_{au} & \zeta_{ac} & 1 & \zeta_{ap} \\ \zeta_{pu} & \zeta_{pc} & \zeta_{pa} & 1 \end{bmatrix} \tag{8.11}$$

is a symmetric matrix, since $\zeta_{uc} = \zeta_{cu}$, $\zeta_{ua} = \zeta_{au}$, $\zeta_{up} = \zeta_{pu}$, $\zeta_{ca} = \zeta_{ac}$, $\zeta_{cp} = \zeta_{pc}$ and $\zeta_{ap} = \zeta_{pa}$. This is the bivariate probit model with multiple sample selection equations. There are several latent variables, each of which is observed as a dichotomous variable, and their underlying distribution is multivariate normal. Therefore, the model is often referred to as the multivariate probit model, for instance by Jenkins et al. (2006).

The bivariate model with sample selection can be estimated by Maximum Likelihood Estimation (MLE). Van der Klaauw and Koning (2000) show that, even when the distributional assumption is not correct, MLE still performs well. Other methods like Heckman's two step estimator or semi-parametric estimation methods can be extended to this situation too. We come back to these estimation methods when we discuss the sample selection model in section 8.3. The log-likelihood for model (8.7) can be constructed by noticing that sample elements can be divided into three mutually exclusive groups: those that are not contacted ($C_i = 0$), those that are contacted but refuse participation ($C_i = 1$ and $P_i = 0$) and those that are contacted and participate ($C_i = 1$ and $P_i = 1$). Each of these groups contributes to a specific part of the likelihood. Thus, the log-likelihood for the sample of n observations is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}^C, \boldsymbol{\beta}^P, \zeta_{cp}) = \textstyle\sum_{i=1}^{n} \quad & \Big((1 - C_i) \times \log P(C_i = 0|\mathbf{X}_i^C) \\ & + C_i(1 - P_i) \times \log P(C_i = 1; P_i = 0|\mathbf{X}_i^C, \mathbf{X}_i^P) \\ & + C_i P_i \times \log P(C_i = 1; P_i = 1|\mathbf{X}_i^C, \mathbf{X}_i^P)\Big)\end{aligned} \tag{8.12}$$

The conditional probabilities in (8.12) can be described as

$$\begin{aligned}P(C_i = 0|\mathbf{X}_i^C) &= \int_{-\infty}^{0} P(\gamma_i^* = z|\mathbf{X}_i^C)dz = \Phi\Big(-\mathbf{X}_i^C\boldsymbol{\beta}^C\Big) \\ P(C_i = 1; P_i = 0|\mathbf{X}_i^C, \mathbf{X}_i^P) &= \int_{0}^{\infty}\int_{-\infty}^{0} P(\gamma_i^* = z_C; \varrho_i^* = z_P|\mathbf{X}_i^C, \mathbf{X}_i^P)dz_P z_C \\ &= \int_{-\mathbf{X}_i^C\boldsymbol{\beta}^C}^{\infty} \phi(\tilde{z}_C)\Big(1 - \Phi\Big(\tfrac{-\mathbf{X}_i^P\boldsymbol{\beta}^P + \zeta_{cp}(\tilde{z}_C)}{\sqrt{1-\zeta_{cp}^2}}\Big)\Big)d\tilde{z}_C \\ P(C_i = 1; P_i = 1|\mathbf{X}_i^C, \mathbf{X}_i^P) &= \int_{0}^{\infty}\int_{0}^{\infty} P(\gamma_i^* = z_C; \varrho_i^* = z_P|\mathbf{X}_i^C, \mathbf{X}_i^P)dz_P z_C \\ &= \int_{-\mathbf{X}_i^C\boldsymbol{\beta}^C}^{\infty} \phi(\tilde{z}_C)\Big(\Phi\Big(\tfrac{-\mathbf{X}_i^P\boldsymbol{\beta}^P + \zeta_{cp}(\tilde{z}_C)}{\sqrt{1-\zeta_{cp}^2}}\Big)\Big)d\tilde{z}_C\end{aligned} \tag{8.13}$$

where we denote $\tilde{z}_C = z_C - \mathbf{X}_i^C\boldsymbol{\beta}^C$ for notational convenience. Maximising the likelihood in (8.12) thus results in estimates for the conditional probabilities. In the following section we show how the bivariate probit model with sample selection is applied to contact and participation in the Dutch LFS.

8.2.4 Bivariate probit analysis of nonresponse to the Dutch LFS

In Chapter 3 we analysed the nonresponse to the ELFS-sample from July to October 2005, using nested logistic regression models for the different response types in the process. Now, we apply the bivariate probit model with sample selection to the data from the LFS. We restrict the analysis to the two main response types; contact and participation. We refer to the response model with these two types of response as the simplified response process, and the response process as analysed in Chapter 3, with four types of response, as the elaborate response process. The bivariate probit model with sample selection additionally allows us to check whether there is a correlation between contact and participation in the ELFS-sample. Lynn and Clarke (2002) and Nicoletti and Peracchi (2005) find little evidence for a correlation between contact and participation.

We merge the response types being processed and contact, as well as being able and participation. See figure 8.2. Sample elements that are not processed are regarded as non-contacts. The total number of observations is $17,628$. The number of non-contacts is $1,438$. These observations are censored when regarding the participation process. For participation, the total number of observations reduces to $16,190$. The sample elements that are not able to participate are now regarded as refusals. We use the models for contact and participation that were obtained in Chapter 3. Because of the merging of the response types, we also merge the variables in the separate logit models. This has no consequence for the model for contact, but for participation the variables ethnic group and percentage non-natives are added to the model. The final models are summarised in table 8.2. We use Stata to perform the analyses by Maximum Likelihood Estimation. Hence, the likelihood in equation (8.12) is optimised. First, we fitted the bivariate probit model for contact and participation as displayed in table 8.2. Testing the variables for joint significance resulted in the omission of the

Table 8.2: *Summary of the different models for contact and participation*

Response type	*Variables*
Contact	Gender, type of household, number of persons, province and large cities, listed telephone, degree of urbanization, average age
Participation	Province and large cities, paid job, average house value ethnic group, percentage of non-natives

variables degree of urbanization ($\chi^2(4) = 4.13; p = 0.3886$) and number of persons ($\chi^2(5) = 8.13; p = 0.1494$) in the contact equation, and in the omission of the variable percentage of non-natives ($\chi^2(4) = 0.84; p = 0.9336$) in the participation equation. The final models are given in table 8.3. The results are given in tables 8.4 and 8.5. Our main interest lies in the estimated correlation between contact and participation, i.e. $\hat{\zeta}_{cp}$. It turns out that $\hat{\zeta}_{cp}$ is significant at a 5% confidence level. The estimated value equals -0.240 with a standard error of 0.116. The corresponding p-value for a χ^2-test equals $p = 0.0329 < 0.05$. Thus according to this model, contact and participation are slightly correlated. The negative value for $\hat{\zeta}_{cp}$ implies that the participation propensity is negatively influenced by the contact propensity. This result suggests that, once contacted, the participation propensity is smaller than the unconditional participation propensity. Maybe nice persons, i.e. persons with a high participation propensity, are more frequently away from home, i.e. have a low contact propensity?

Let us take a look at the propensities, i.e. the estimated probabilities. Figure 8.3 displays two histograms with the density of the propensities. The first histogram displays the marginal participation propensity in the bivariate probit model with sample selection, i.e. $\hat{\varrho}_i^{marg} = P(P_i = 1)$. The second histogram shows the conditional participation propensity in the bivariate probit model with sample selection, i.e. $\hat{\varrho}_i^{cond} = P(P_i = 1|C_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P)$. In the histograms we also show a fitted kernel density with a bandwidth of 0.6 to approximate the density of the propensities. The mean estimated marginal participation propensity, $\overline{\hat{\varrho}}^{marg}$ equals 0.689, which is higher than the mean estimated conditional participation propensity $\overline{\hat{\varrho}}^{cond} = 0.677$. This confirms the negative correlation between contact and participation. Once contacted, the participation propensity is lower. The difference, however, is not significant.

A possible explanation may be the very high contact rate. The mean estimated contact propensity equals $\overline{\hat{\gamma}} = 0.937$. The results suggest that the few sample elements that are not contacted would have easily participated. Indeed

Table 8.3: *Final models in the bivariate probit model with sample selection*

Response type	*Variables*
Contact	Gender, type of household, province and large cities, listed telephone, average age
Participation	Province and large cities, paid job, average house value ethnic group

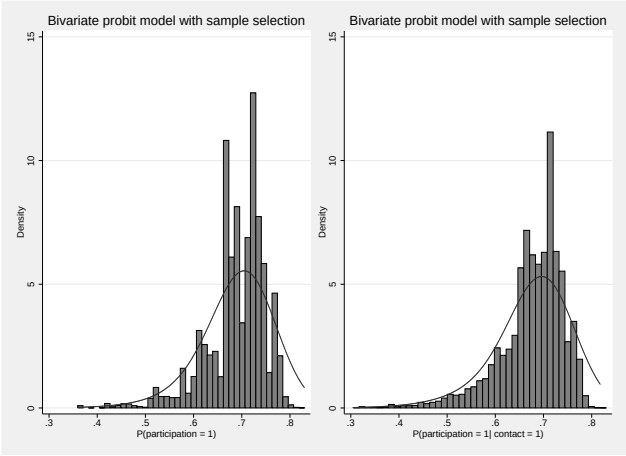

Figure 8.3: *Propensities for participation, and participation conditional on contact in the bivariate probit model with sample selection*

nice persons could be away from home frequently. However, remember that the response type being processed is merged with non-contact. Another hypothesis that follows from this analysis, is that unprocessed cases occur frequently among sample elements with a participation propensity that is higher than average.

8.2.5 A comparison of the bivariate probit model with sample selection and the nested logit model

To compare the bivariate probit model with sample selection and the nested logistic regression technique we apply the latter technique to the simplified response process consisting of contact and participation. We can compare the estimated coefficients in the two models by noticing that the logit and the probit model produce estimates for $\boldsymbol{\beta}$ that are distinct, but roughly proportional. Johnston and Dinardo (1997) show that, when we consider the derivatives of the probabilities of contact or participation with respect to $\mathbf{X}$ at the mean of the sample, then

$$\beta_i^{logit} \approx \frac{\phi\left[\Phi^{-1}(\overline{p})\right]}{\overline{p}(1-\overline{p})}\beta_i^{probit} \tag{8.14}$$

where $\overline{p}$ is the proportion of 1s (contact or participation) in the sample. In case of contact, $\overline{p} = 0.918$ and $\beta_i^{logit} \approx 2.0\beta_i^{probit}$. In case of participation, $\overline{p} = 0.625$ and $\beta_i^{logit} \approx 1.6\beta_i^{probit}$. For the interpretation of the coefficients we focus on the sign and significance of the reported estimated coefficients. Tables 8.4 and 8.5 display the results for the two models for contact respectively participation. We report $\hat{\boldsymbol{\beta}}$ and the corresponding standard error *se*. An asterisk denotes that the variable is significant at a 5%-level, double asterisks denote significance at a 1%-level. General model fit statistics are the Nagelkerke pseudo R^2 (see (3.2)), the χ^2 and the degrees of freedom (df) of the model. Furthermore, for the bivariate model with sample selection the estimated correlation between contact and participation $\hat{\zeta}_{cp}$ is reported too.

The parameter estimates for the nested logit models and the bivariate probit model with sample selection are roughly the same (when accounting for the proportionality of the parameters described in (8.14)) for both contact and participation. Although the estimated correlation $\hat{\zeta}_{cp}$ between contact and participation is significant at a 5%-level, the parameter estimates vary little. This result is in line with Nicoletti and Peracchi (2005), who also find a significant correlation but no distinct parameter estimates. Concluding, there is a (weak) correlation between contact and participation. However, allowing for the correlation does not change the parameter estimates.

8.2.6 Multilevel model

Surveys can be conducted using interviewers, or by means of self administration. In Chapter 9 we discuss the advantages and disadvantages of different data collection modes. In interviewer-assisted surveys, the interviewer communicates with the respondent either face-to-face or over the telephone. In these type of surveys, the interviewer plays an important role in the response process. Biemer and Lyberg (2003) provide a discussion on the role of the interviewer. They distinguish between standardised interviewing and conversational interviewing. In standardised interviewing, the interviewer has to follow a protocol that tries to ensure that the interviewers behave in a standard way, thus reducing the risk of measurement error in responses caused by the interviewer. An example of this is the fieldwork strategy of Statistics Netherlands, described in Chapter 3. In conversational interviewing, interviewers are more free to interact with respondents, to optimise the survey situation for the respondent in order to minimise measurement error in all error sources, not just the interviewer's error.

Whatever style of interviewing, the interviewers will affect the response pro-

Table 8.4: *Nested logit model and bivariate probit model with sample selection estimates for the contact propensity* $\hat{\gamma}$

		Logit		*Probit with selection*	
Variable	*Category*	β^{logit}	se	β^{probit}	se
Gender	All women	0.545**	0.082	0.296**	0.044
(Reference all men)	Mixed	0.633**	0.130	0.324**	0.069
Type of household	Unmarried no/c	0.339*	0.141	0.184*	0.075
(Reference single)	Married no/c	0.427**	0.153	0.230**	0.079
	Unmarried with/c	0.734**	0.197	0.375**	0.100
	Married with/c	0.679**	0.144	0.367**	0.075
	Single parent	0.506**	0.122	0.274**	0.063
	Other	0.321	0.291	0.153	0.156
	Mixed	0.480**	0.165	0.252**	0.086
Province + large cities	Friesland	1.682**	0.198	0.941**	0.104
(Reference Groningen)	Drenthe	1.365**	0.487	0.783**	0.241
	Overijssel	1.178**	0.153	0.644**	0.085
	Gelderland	1.402**	0.180	0.784**	0.097
	Utrecht (except Utrecht city)	1.730**	0.181	0.933**	0.095
	Noord-Holland (except Amsterdam)	1.066**	0.140	0.622**	0.079
	Zuid-Holland (except Den Haag, Rotterdam)	1.403**	0.146	0.785**	0.081
	Noord-Brabant	1.879**	0.165	1.017**	0.088
	Limburg	2.035**	0.181	1.091**	0.094
	Amsterdam	0.293*	0.144	0.181*	0.084
	Rotterdam	1.168**	0.153	0.671**	0.087
	Den Haag	1.825**	0.210	0.992**	0.110
	Utrecht	1.434**	0.212	0.791**	0.114
Listed telephone	yes	1.281**	0.069	0.637**	0.033
(Reference no)					
Average age	>= 35 and < 55	0.183**	0.071	0.086*	0.038
(Reference >= 15; < 35)	>= 55	0.398**	0.089	0.198**	0.047
Constant		-0.358**	0.134	-0.120	0.078
Pseudo R^2		0.1466			
χ^2		1460.57			
df		25			

Table 8.5: *Nested logit model and bivariate probit model with sample selection estimates for the participation propensity* $\hat{\varrho}$

		Logit		*Probit with selection*	
Variable	*Category*	β^{logit}	se	β^{probit}	se
Province + large cities	Friesland	0.169	0.153	0.016	0.100
(Reference Groningen)	Drenthe	0.030	0.276	-0.058	0.170
	Overijssel	0.151	0.145	0.023	0.093
	Gelderland	0.125	0.154	0.005	0.099
	Utrecht	-0.045	0.149	-0.103	0.087
	Noord-Holland	-0.117	0.142	-0.133	0.090
	Zuid-Holland	-0.116	0.142	-0.141	0.092
	Noord-Brabant	0.036	0.144	-0.059	0.094
	Limburg	0.225	0.146	0.051	0.096
	Amsterdam	-0.574**	0.155	-0.364**	0.094
	Rotterdam	-0.094	0.149	-0.119	0.095
	Den Haag	-0.062	0.161	-0.118	0.104
	Utrecht	0.273	0.177	0.101	0.111
Paid job (Reference No)	Yes	0.249**	0.038	0.149**	0.23
Average house value[a]	0 - 50	-0.036	0.135	-0.002	0.082
(Reference missing)	50 - 75	-0.187*	0.087	-0.103	0.053
	75–100	-0.219**	0.079	-0.130**	0.048
	100–125	-0.213**	0.078	-0.131**	0.047
	125–150	-0.101	0.080	-0.069	0.048
	150–200	0.045	0.079	0.017	0.048
	200–250	0.058	0.094	0.023	0.056
	250–300	0.079	0.125	0.034	0.075
	300–400	-0.066	0.133	-0.055	0.080
	400 and more	0.074	0.202	0.029	0.122
Ethnic group	Morrocan	-0.678**	0.137	-0.412**	0.085
(Reference native)	Turkish	-0.404**	0.126	-0.244**	0.078
	Surinam, NL Antilles	0.242*	0.119	0.157*	0.071
	other non-Western	-0.275**	0.074	-0.159**	0.046
	other Western	-0.311**	0.111	-0.182**	0.068
	mixed	0.035	0.054	0.017	0.032
Constant		0.716**	0.150	0.548**	0.102
Pseudo R^2		0.0148			
χ^2		300.77			
df		30			
$\hat{\zeta}_{cp}$				-0.240*	0.116
LR test of independent equations ($\hat{\zeta}_{cp} = 0$): $\chi^2(1) = 4.55$, p-value = 0.033					

[a] × 1,000 euro

cess in the survey, as well as the answers provided to the survey. An interviewer effect occurs, when the interviewers systematically affect the probability that sample elements participate, or are contacted. Also, the interviewers may affect the survey answers. For instance, when the interviewer thinks positive of voting and sample elements therefore overreport voting behaviour (i.e. a social desirability bias, see Chapter 9). Since multiple sample elements are assigned to one interviewer, the effect of the interviewer on the response process and the survey answers is clustered within interviewers. To analyse interviewer effects and to use the lowest level of available information, i.e. the sample element level, this clustering has to be accounted for. Similarly, an analysis of respondents effects, with the inclusion of interviewer data, should take into account the clustering and thus the hierarchical structure of the data too. In this section we show how the models in the previous sections can be refined to include interviewer variance into the analysis models.

The type of model that can account for interviewer variance appears in the literature under a variety of different names: hierarchical regression model (Hox, 1994), random effects model (Agresti, 2002), variance component model (Anderson and Aitkin, 1985), multilevel model (Hox, 1995; Durrant and Steele, 2007). We will refer to this model as the multilevel model, because this term is at this moment most commonly used in the survey literature.

The multilevel model can be used to describe different forms of nesting. For instance the nesting of persons within families, pupils within schools, or interviewers within geographical areas as described in Durrant and Steele (2007). Here, we describe how the model can be used to analyse the effect of interviewers on data quality. In survey sampling, the sample elements are drawn randomly from the sampling frame. In case of a face-to-face survey, clusters of elements are assigned to interviewers based on geographic location. The characteristics of the interviewer are then confounded with the characteristics of the respondents and there is a correlation between the characteristics of the interviewer and the respondents. For instance, persons that live in highly urbanized areas will be interviewed by an urban interviewer. This may improve the response, since similarity leads to liking (see the compliance principles discussed in Chapter 2), but it can also introduce an unwanted interviewer effect. In case of a CATI survey using a centralised system (see Chapter 9), sample elements are distributed randomly to interviewers. The confounding of characteristics does not apply to this situation, but the interviewers still affect the response process and survey answers.

A detailed description of the multilevel model can be found in Hox (1995). We start first with the logistic regression model for the response indicator R

and response propensity ρ, i.e. equation (8.2). Extending this model to include interviewer effects, implies introducing an extra index for the interviewers and a corresponding random term for the interviewer effect, which we denote by ϵ. Suppose there are $m = 1, \ldots, M$ interviewers. Then the multilevel model becomes

$$logit(\rho_{im}) = \mathbf{X}^R_{im}\boldsymbol{\beta} + \epsilon_m \tag{8.15}$$

for $i = 1, \ldots, n$ and for $m = 1, \ldots, M$. The random effect for interviewer m, ϵ_m, is included in the equation, sample elements with the same interviewer are thus assigned the same interviewer effect. We assume that $\epsilon_m \sim NID(0, \sigma^2_m)$, where NID denotes normally and independently distributed.

This model can be extended in several ways. For example, the variance can be made heterogeneous by introducing a sample element specific variance, i.e. σ^2_{im}. This will, however, introduce a large number of additional model parameters. Depending on the size of n, model estimation will be computationally intensive or even impossible. Another extension involves integrating the multilevel model in the nested logit model or the bivariate probit model with sample selection, This would enable the analysis of correlated interviewer effects between the various response types. The nested logit models described in equations (8.3) to (8.6) are extended with an additional index for the interviewers and an additional term ϵ_m for the interviewer variance, as described in (8.15). The probit model with sample selection for contact and participation, described in (8.7) then becomes

$$\begin{array}{rcl} \gamma^*_{im} = & \mathbf{X}^C_{im}\boldsymbol{\beta}^C + \epsilon^C_{im} + \epsilon^C_m \\ \varrho^*_{im} = & \mathbf{X}^P_{im}\boldsymbol{\beta}^P + \epsilon^P_{im} + \epsilon^P_m \end{array} \tag{8.16}$$

for $i = 1, \ldots, n$ and $m = 1, \ldots, M$. ϵ^C_m is the interviewer variance for interviewer m in the contact participation, and likewise ϵ^P_m is the interviewer variance for interviewer m in the participation equation. The random effects ϵ^C_{im} and ϵ^C_m are mutually independent with different variance components, the same holds for ϵ^P_{im} and ϵ^P_m. For instance, ϵ^C_{im} accounts for the variation among sample elements in their contact propensity that is not measured in the variables $\mathbf{X}^C_{im}$, whereas ϵ^C_m accounts for the variation among interviewers in making contact to sample elements. Because the different error terms are mutually independent, we can describe the corresponding error term distributions as follows

$$\begin{pmatrix} \epsilon^C_{im} \\ \epsilon^P_{im} \end{pmatrix} \backsim N\left(\mathbf{0}, \begin{bmatrix} 1 & \zeta_{cp} \\ \zeta_{pc} & 1 \end{bmatrix}\right) \tag{8.17}$$

and

$$\begin{pmatrix} \epsilon_m^C \\ \epsilon_m^P \end{pmatrix} \backsim N\left(\mathbf{0}, \begin{bmatrix} \sigma_c & \zeta_{m,cp}\sigma_c \\ \zeta_{m,pc}\sigma_p & \sigma_p \end{bmatrix}\right) \tag{8.18}$$

where $\zeta_{cp} = \zeta_{pc}$ is the correlation between the contact- and the participation equation, σ_c respectively σ_p is the variance of the interviewer effect on contact respectively participation and $\zeta_{m,cp} = \zeta_{m,pc}$ is the correlation between the interviewer effect on contact and the effect on participation. This correlation is positive when interviewers have the same systematic effect on making contact as on gaining participation, either positive of negative. The correlation between the respondent level effect and the interviewer effect is zero; these effects are assumed to be mutually independent.

8.2.7 Advantages and disadvantages

The models described in section 8.2.2 and 8.2.3 can be employed to analyse response behaviour. The underlying assumptions of the models with respect to the correlation structure of the distinct stages in the response process are different. Furthermore, they differ in the way that they account for the sequential character of the response process. Table 8.6 summarises characteristics of the different models.

The multinomial model allows for a distinction between different response types in a sense that the dependent variables for each of the response types, and the corresponding coefficients, are allowed to be different. This type of model, however, does not acknowledge the sequential nature of the response process, nor does it allow for a dependency between response types. The nested logit model does account for the sequential nature of the response process. Nevertheless, because the different indicators are only defined on subsets of the sample, the truncation that arises due to unobserved indicators is not accounted for.

Table 8.6: *Characteristics of the nonresponse analysis models*

Model	*Distinguish response types*	*Sequential nature*	*Dependence assumption*
Multinomial	yes	no	no
Nested logit	yes	yes	no
Bivariate probit with sample selection	yes	yes	yes

Hence the underlying assumption of the nested logit model is that the processes are independent from each other. The bivariate probit model with sample selection regards the censoring of sample elements as a form of sample selection. It makes the censoring explicit by acknowledging the fact that certain response types are not observed, conditional on previous stages in the response process. The underlying propensities are being estimated based on their observed indicators, but unlike the nested logit model the propensities are defined for every sample element regardless of whether the corresponding indicator is observed. In addition, the joint distribution of the error terms in the equations allow for a correlation and therefore allow for a dependency between the different stages in the response process.

The multilevel model tackles another type of nesting, namely that of the clustering of sample elements within interviewers. Therefore, this model is a refinement of the other models to allow for interviewer effects. The clustering causes confounding of characteristics from the interviewer and the sample elements. When analysing interviewer effects, the multilevel model distinguishes these two levels of information. Furthermore, when the data are analysed on the respondent level and data on the interviewers is included in the models, the cluster effect has to be accounted for. Otherwise the standard errors are much lower than they should be, leading to results that appear more significant than they are.

The most elaborate nonresponse analysis model thus is the bivariate model with sample selection refined with interviewer effects. In this model, the response types are allowed to be correlated, defined for all sample elements and the effects of the respondents and the interviewers are separated.

8.3 Alternative methods for nonresponse adjustment

8.3.1 Introduction

The objective of nonresponse adjustment methods is to reduce the bias in the estimated survey items for errors due to selective nonresponse. In the previous section we discussed a number of models that describe the relationship between the characteristics of sample elements and their response behaviour. In this section we describe how these models can be extended to adjust for nonresponse bias.

To describe the relationship between response behaviour and auxiliary infor-

mation, it can be sufficient to include demographic or socio-economic characteristics that are not survey specific but approximately the same in all surveys, as we did in Chapter 3. However, the relation between auxiliary information and the survey item Y can be different for each item, and for different surveys. For instance, the topic of the survey can influence survey participation and the name of the survey alone can cause nonresponse. In that case, changing the name of the survey will influence the response rate. For example, the Dutch 'Fertility Survey' had a low response rate. Changing the name of the survey to 'Fertility and Family Survey' increased the response. Furthermore, there are situations in which distinct response types may have a different influence on the survey item. In those situations especially, the composition of the response influences the estimated survey items and therefore it is important to model and take into account these specific relationships.

Example 8.3.1 Relationship between employment and non-contact
Daalmans et al. (2006) have analysed the relationship between an increasing response rate and estimates for the number of employed persons in the Dutch LFS in 2004. In 2004, the composition of the response changed due to changes in the fieldwork strategy (described in Chapter 3). Especially the non-contact rate has decreased. Daalmans et al. (2006) analyse the number of employed persons, adjusted for nonresponse with the linear regression estimator using the standard LFS weighting model, for consecutive contact attempts. It turns out that, after each consecutive contact attempt, the estimated number of employed persons increases. See figure 8.4. They conclude that there is a very strong relation between contact and employment. ■

There may be a correlation between the survey items and the different response types. Again, if the response types display a different relationship with the survey item Y, the effect on the nonresponse bias is different as shown in equation (8.1).

Another issue that arises, is the number of survey items. A survey usually consists of a large number of survey items. The effect of nonresponse can be different for each of these variables and it would be interesting to adjust each of the survey items separately, i.e. use different adjustment models for different survey items. This is not desirable in practice as there are also some other considerations for nonresponse bias adjustment, see Chapter 6. The adjustment weights calculated by the linear regression estimator can be described in terms of the $\mathbf{X}$-variables, thereby not relying on a specific survey item as we have shown in Chapter 6, but using a fixed set of $\mathbf{X}$-variables for every survey item.

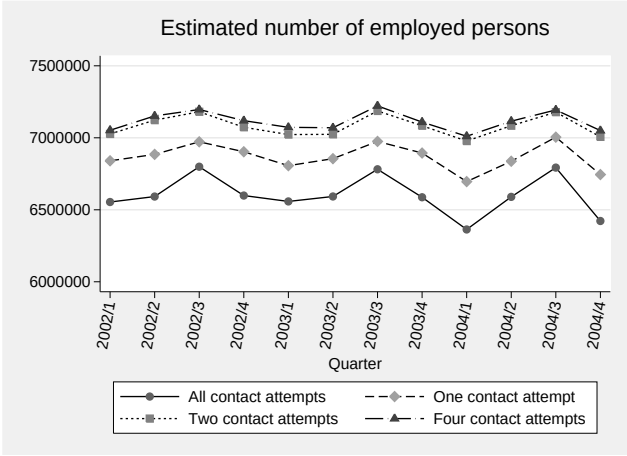

Figure 8.4: *Relationship between contact and employment in the Dutch LFS 2001 - 2004*

In this section, we discuss two weighting adjustment methods that take into account different response types and the sequential character of the response process. We present the sequential weight adjustment method, and the sample selection model (with as a special case the bivariate probit model with sample selection). The sample selection model additionally allows for a correlation between the response types, and a correlation between response types and survey items. Section 8.3.2 describes the sequential weight adjustment technique. In section 8.3.3 we present the sample selection model. We show how it can be used in the situation of survey nonresponse. Furthermore, we discuss issues regarding the identification and the estimation of the sample selection model. In section 8.3.3.4 we give suggestions for the extension of the sample selection model for the inclusion of multiple selection equations for different response types. Section 8.3.3.5 discusses how the sample selection model can be customised to allow for categorical survey items. Finally, in section 8.3.4 we summarise the models presented in this section.

8.3.2 Sequential weight adjustment method

Groves and Couper (1998) and Iannacchione (2003) describe a sequential weight adjustment method for a two stage response process consisting of contact, and

participation conditional on contact. The nested logit models described in section 8.2 are used to construct nonresponse adjustment weights. We first describe the sequential weight adjustment method. Then we outline the advantages and disadvantages of the method.

The method consists of sequentially fitting a number of logistic regression models. First, a logistic model for the contact probability is constructed for all sample elements, i.e.

$$logit(\gamma(\mathbf{X}_i^C)) = \mathbf{X}_i^C \boldsymbol{\beta}^C \tag{8.19}$$

for $i = 1, \ldots, n$. The model parameters $\boldsymbol{\beta}^C$ are estimated by MLE. The contact propensity can be computed as

$$\hat{\gamma}_i = \frac{exp(\mathbf{X}_i^C \hat{\boldsymbol{\beta}}^C)}{1 + exp(\mathbf{X}_i^C \hat{\boldsymbol{\beta}}^C)} = \frac{1}{1 + exp(-\mathbf{X}_i^C \hat{\boldsymbol{\beta}}^C)} \tag{8.20}$$

Based on this regression, sample elements that have successfully been contacted ($C_i = 1$) receive as a weight $w_i = 1/\hat{\gamma}_i$ so that they represent the sample. In the second step, a weighted logistic regression is fitted for the participation propensity using only the subsample of contacted sample elements, i.e.

$$logit(\varrho(\mathbf{X}_i^{P|C=1})) = w_i \mathbf{X}_i^{P|C=1} \boldsymbol{\beta}^{P|C=1} \tag{8.21}$$

for $i = 1, \ldots, n^c$ where n^c are the sample elements that have been contacted, i.e. for which $C_i = 1$. Again, $\boldsymbol{\beta}^{P|C=1}$ can be estimated by MLE and the participation propensity can be calculated as

$$\hat{\varrho}_i = \frac{1}{1 + exp(-\mathbf{X}_i^{P|C=1} \hat{\boldsymbol{\beta}}^{P|C=1})} \tag{8.22}$$

The final weights w_i^* are obtained by multiplying w_i with

$$1 + exp(-\mathbf{X}_i^{P|C=1} \hat{\boldsymbol{\beta}}^{P|C=1}) \tag{8.23}$$

leading to the following expression for w_i^*

$$w_i^* = w_i \times (1 + exp(-\mathbf{X}_i^{P|C=1} \hat{\boldsymbol{\beta}}^{P|C=1})) = \frac{w_i}{\hat{\varrho}_i} \tag{8.24}$$

Both Groves and Couper (1998) and Iannacchione (2003) already use a weighted logistic regression model in the first step, to account for the survey design by including the design weights d_i (the inverse of the inclusion probability, $d_i = \pi_i^{-1}$).

The weights that are obtained by the sequential weights adjustment procedure can be used in the nonresponse adjustment methods that we discussed in Chapter 6, for instance in the Horvitz-Thompson estimator or the propensity score adjustment methods.

The sequential weight adjustment method thus accounts for the sequential nature of the response process. Furthermore, it allows for a distinction between the different response types. However, when the estimated propensities are directly used as weights the variance of the estimator can become large. This effect increases with the number of nested processes (Little and Vartivarian, 2005). In addition, similar to the nested logit model, a disadvantage of the method is that it does not allow for correlation between response types. Moreover, the participation propensity is only defined for the subset of sample elements that have been successfully contacted.

We can overcome this disadvantage by using probit models instead of logit models. The probit model with sample selection that we presented in section 8.2 defines an underlying, latent participation propensity that is also defined for sample elements that are not contacted. Furthermore, the error distribution of the probit model allows for correlation between the distinct stages. Because of these properties of the bivariate probit model with sample selection, it is not necessary to explicitly weight the observations for the different types of response, as in the nested logit model approach described above. Nicoletti and Peracchi (2005) have applied the bivariate probit model with sample selection described in section 8.2.3.

To adjust for nonresponse bias, as argued in Chapter 6, the relationship with the survey item(s) should also be included in the models. Nicoletti and Peracchi (2005) do not include this relationship, but instead they focus on the response process solely. Therefore, in the next section we discuss an extension of the probit model with sample selection that also includes the relationship with the survey item(s).

8.3.3 The sample selection model

8.3.3.1 The sample selection model in its original context

The sample selection model was first proposed by Heckman (1979) to determine women's wages and labour force supply, see also Heckman (1974) and Gronau (1974). The outcome that the researchers were interested in concerned the women's wages. It was argued that the observed wages for women were not a random sample of women's wages. The selection was the decision of woman to

become active in the labour force. If the decision of a woman to participate in the labour force depends on characteristics that are also influential on her wage, then there is a selectivity bias. When all variables that determine the participation decision are observed, these can simply be inserted in the wage equation. Conditional on these variables, the wages for non-working and working women are the same. This corresponds to the MAR-assumption, or the conditional independence assumption. If, however, there are unobservable characteristics, and these characteristics influence both the decision to participate in the labour force and the wage, then there is a relationship between the selection (the participation decision) and the outcome (women's wages). In this case it is not sufficient to include the observed characteristics in the wage equation, because through the unobserved characteristics the wage is also influenced by the decision to participate in the labour force.

In this section, we first translate the sample selection model to the context of survey nonresponse. If the decision to participate in a survey is dependent on characteristics that are also related to the survey item(s), there is a selectivity bias, or nonresponse bias. We discuss issues of identification of the model and present a number of ways to estimate sample selection models. Next, we extend the model to account for different response types, as well as categorical survey items.

8.3.3.2 Sample selection due to nonresponse

The sample selection model can be applied to the situation of survey nonresponse. In the context of survey nonresponse, the sample selection arises due to self-selection of respondents, either explicit (refusal) or implicit (not able, noncontact) and in some cases also due to actual sample selection (unprocessed cases). We refer to both types as *sample selection.*

In this section, we describe the sample selection model for survey nonresponse in the simplified form where we only consider the final response and the outcome for the survey item, see figure 8.5. The generalisation to more response stages is discussed in section 8.3.3.4. The sample selection model described in figure 8.5 consists of two equations. The first equation models the response propensity. The outcome of the first equation determines whether the survey item is observed. This equation is therefore referred to as the *selection equation.* The second equation accounts for the censoring of sample elements that do not respond, and models the survey item. Therefore the second equation is referred to as the *regression equation.*

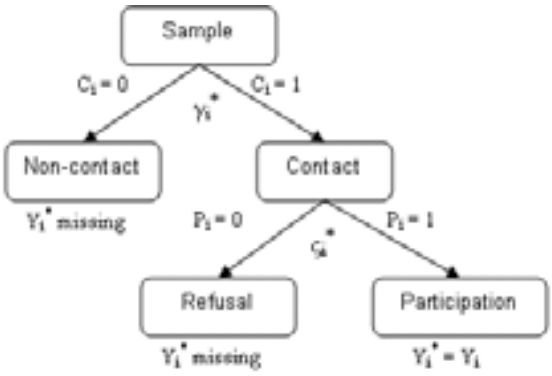

Figure 8.5: *The sample selection model for survey nonresponse*

The sample selection model for survey nonresponse consists of two equations

$$\begin{array}{rl} \rho_i^* = & \mathbf{X}_i^R \boldsymbol{\beta}^R + \epsilon_i^R, \\ Y_i^* = & \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \epsilon_i^Y \end{array} \tag{8.25}$$

for $i = 1, \ldots, n$. Both equations are latent variable regression equations. We do not observe ρ_i^*, but instead we observe the binary response indicator R_i. If $R_i = 1$, we observe Y_i^*, otherwise Y_i^* is missing, i.e. censored. Thus, we can define Y_i as

$$Y_i = \left\{ \begin{array}{ll} Y_i^*, & \text{if } R_i = 1 \\ \text{missing}, & \text{if } R_i = 0 \end{array} \right. \tag{8.26}$$

The bivariate model with sample selection described in the previous section is a modification of the sample selection model for discrete outcomes. In case of a binary categorical survey item, for instance having a job yes/no, the bivariate probit model with sample selection is the appropriate model to use.

In the sample selection model the error term distribution is assumed to be bivariate normal with variance σ^2, which leads to the following joint distribution of the error terms

$$\left(\begin{array}{c} \epsilon_i^R \\ \epsilon_i^Y \end{array} \right) \backsim N\left(\mathbf{0}, \left[\begin{array}{cc} 1 & \zeta\sigma \\ \zeta\sigma & \sigma^2 \end{array} \right] \right) \tag{8.27}$$

with ζ the correlation between the two error terms.

Recall that we are still interested in estimating the survey items, adjusted for nonresponse bias. In Chapter 1 we showed that the Horvitz-Thompson estimator $\overline{y}_{ht}$ is an unbiased estimator for the population mean of a survey item Y

$$\overline{y}_{ht} = \frac{1}{N}\sum_{i=1}^{N}\delta_i\frac{Y_i}{\pi_i}$$

Due to nonresponse, this estimator becomes biased. The modified Horvitz-Thompson estimator $\overline{y}^r_{ht}$ deals with nonresponse bias by adjusting the inclusion probabilities for the response probability, see Chapter 6,

$$\overline{y}^r_{ht} = \frac{1}{N}\sum_{i=1}^{n}\frac{R_iY_i}{\pi_i\rho_i}$$

In terms of the sample selection model, we can modify the Horvitz-Thompson estimator by replacing Y_i with the expected value of Y_i given the model and $R_i = 1$. We denote by $\overline{y}^{sel}_{ht}$ the modified Horvitz-Thompson estimator for the mean of survey item Y based on the sample selection model.

$$\overline{y}^{sel}_{ht} = \frac{1}{N}\sum_{i=1}^{n}\frac{\hat{E}\Big[Y_i|R_i = 1, \mathbf{X}^R_i, \mathbf{X}^Y_i\Big]}{\pi_i} \tag{8.28}$$

In the next section we show how we can calculate (8.28) by estimating the parameters of the sample selection model.

8.3.3.3 Estimation

There are a number of methods that can be employed to estimate the parameters in the sample selection model. Vella (1998) provides an overview of the different methods. We discuss the most commonly used methods, which are Maximum Likelihood Estimation and the two-step estimator proposed by Heckman (1979). Both methods depend heavily on the underlying distributional assumptions. To avoid distributional assumptions, the literature suggests to use semi-parametric estimation methods. For example Gallant and Nychka (1987) discuss a general semi-parametric estimation strategy.

Let us first consider the expected value of Y_i conditional on $R_i = 1$, and $\mathbf{X}^R_i, \mathbf{X}^Y_i$

$$\begin{aligned} E[Y_i|R_i = 1, \mathbf{X}^R_i, \mathbf{X}^Y_i] &= \mathbf{X}^Y_i\boldsymbol{\beta}^Y + E[\epsilon^Y_i|R_i = 1, \mathbf{X}^R_i, \mathbf{X}^Y_i] \\ &= \mathbf{X}^Y_i\boldsymbol{\beta}^Y + E[\epsilon^Y_i|\epsilon^R_i > -\mathbf{X}^R_i\boldsymbol{\beta}^R, \mathbf{X}^R_i, \mathbf{X}^Y_i] \end{aligned}$$

From this conditional expectation we can see that there is a selection bias if ϵ_i^Y and ϵ_i^R are dependent, or $\mathbf{X}_i^R$ and $\mathbf{X}_i^Y$ are correlated, i.e. $E[\epsilon_i^Y|\epsilon_i^R] \neq 0$. A test for the presence of a selection bias involves testing whether the correlation coefficient $\zeta = 0$. If $\zeta = 0$, there is no selection bias and the two equations of the sample selection model can be estimated separately.

Maximum Likelihood Estimation (MLE) amounts to finding specific parameter values that optimise the probability of finding the data that have actually been observed. The likelihood function is a function of the parameters, conditional on the observed data, i.e. $L(\boldsymbol{\beta}^R, \boldsymbol{\beta}^Y, \sigma^2, \zeta; Y)$. It is usually more simple to maximize the log of the likelihood, denoted by $\mathcal{L}(\boldsymbol{\beta}^R, \boldsymbol{\beta}^Y, \sigma^2, \zeta; Y)$. For the sample selection model (8.25) the log-likelihood function can be derived as in the bivariate probit model with sample selection (see (8.12)). There are two mutually exclusive groups, respondents $R_i = 1$ and nonrespondents $R_i = 0$. For sample elements that respond, we observe Y_i so their contribution to the likelihood is $P(R_i = 1; Y_i = y_i|\mathbf{X}_i^R, \mathbf{X}_i^Y)$. For nonresponding sample elements we do not observe Y_i, and their contribution to the likelihood equals $P(R_i = 0|\mathbf{X}_i^R)$. Thus, the log-likelihood for the sample of n observations is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}^R, \boldsymbol{\beta}^Y, \sigma^2, \zeta; Y) = \textstyle\sum_{i=1}^n \quad & \Big((1 - R_i) \times \log P(R_i = 0|\mathbf{X}_i^R) \\ & + R_i \times \log P(R_i = 1; Y_i = y|\mathbf{X}_i^R, \mathbf{X}_i^Y)\Big)\end{aligned} \tag{8.29}$$

where the conditional probabilities are defined as

$$P(R_i = 0|\mathbf{X}_i^R) = \int_{-\infty}^{0} P(\rho_i^* = r|\mathbf{X}_i^R)dr = \Phi\Big(-\mathbf{X}_i^R\boldsymbol{\beta}^R\Big) \tag{8.30}$$

and

$$P(R_i = 1; Y_i = y|\mathbf{X}_i^R, \mathbf{X}_i^Y) = \int_{0}^{\infty} P(\rho_i^* = r; Y_i^* = y|\mathbf{X}_i^R, \mathbf{X}_i^Y)dr \tag{8.31}$$

$$= \Phi\Big(\frac{-\mathbf{X}_i^R\boldsymbol{\beta}^R - \frac{\zeta}{\sigma}(y - \mathbf{X}_i^Y\boldsymbol{\beta}^Y)}{\sqrt{1 - \zeta^2}}\Big) \times \frac{1}{\sigma}\phi\Big(\frac{y - \mathbf{X}_i^Y\boldsymbol{\beta}^Y}{\sigma}\Big) \tag{8.32}$$

The estimates for the coefficients that are obtained by MLE are consistent and efficient, but not necessarily unbiased. The results become inconsistent if the underlying distributional assumptions are violated. Furthermore, full likelihood estimation may become cumbersome if the number of parameters is large.

Heckman (1979) has developed a two-step procedure that avoids some of the complications of full MLE. Again, consider the conditional expectation of Y_i

$$\begin{aligned} E[Y_i|R_i = 1, \mathbf{X}_i^R, \mathbf{X}_i^Y] &= \mathbf{X}_i^Y \boldsymbol{\beta}^Y + E[\epsilon_i^Y | \epsilon_i^R > -\mathbf{X}_i^R \boldsymbol{\beta}^R, \mathbf{X}_i^R, \mathbf{X}_i^Y] \\ &= \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \zeta\sigma \frac{\phi(-\mathbf{X}_i^R \boldsymbol{\beta}^R)}{1 - \Phi(-\mathbf{X}_i^R \boldsymbol{\beta}^R)} \\ &= \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \zeta\sigma \frac{\phi(\mathbf{X}_i^R \boldsymbol{\beta}^R)}{\Phi(\mathbf{X}_i^R \boldsymbol{\beta}^R)} \end{aligned}$$

The term $\phi(\mathbf{X}_i^R \boldsymbol{\beta}^R)/\Phi(\mathbf{X}_i^R \boldsymbol{\beta}^R)$ is referred to as the *inverse Mills ratio*, denoted by λ_i. In this context, λ_i is a monotone decreasing function of the probability that a sampled element responds to the survey request. The interpretation of the inverse Mills ratio λ_i is that observations with a higher response probability have a lower probability of introducing selection bias. The first step of the Heckman two-step procedure involves estimation of $\boldsymbol{\beta}^R$ by applying a probit model to the selection equation. The log-likelihood function for the probit model (see also Chapter 5, appendix 5.A) is

$$\mathcal{L}(\boldsymbol{\beta}^R) = \sum_{i=1}^{n} (1 - R_i) \times \log \Phi\Big(- \mathbf{X}_i^R \boldsymbol{\beta}^R \Big) + R_i \times \log \left(1 - \Phi\Big(- \mathbf{X}_i^R \boldsymbol{\beta}^R \Big)\right)$$

The ML-estimate for $\boldsymbol{\beta}^R$ is denoted $\hat{\boldsymbol{\beta}}^R$. The inverse Mills ratio can be computed with the estimated values by

$$\hat{\lambda}_i = \frac{\phi(\mathbf{X}_i^R \hat{\boldsymbol{\beta}}^R)}{\Phi(\mathbf{X}_i^R \hat{\boldsymbol{\beta}}^R)} \tag{8.33}$$

The rationale behind the two-step estimator arises from regarding the conditional expectation of ϵ_i^Y as an omitted variable from the regression equation in the sample selection model. The selection bias is caused by this omission. Therefore, in the second step of Heckman's procedure the omitted variable $\hat{\lambda}_i$ is included in the regression equation to adjust for the selectivity. The regression equation can be estimated straightforwardly by OLS, based on the subset of observations for which $R_i = 1$. The regression equation becomes

$$Y_i = \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \sigma\zeta\hat{\lambda}_i + \epsilon_i^{Y^*} \tag{8.34}$$

From this equation we can see that testing for sample selectivity bias reduces to testing whether $\zeta = 0$. The two-step procedure provides consistent estimators,

but it ignores the fact that the error term $\epsilon_i^{Y^*}$ now is heteroskedastic because of the inclusion of the estimated inverse Mills ratio. The standard errors in the regression equation have to be adjusted to account for the first step estimation (see, for example, Greene 1981). As a consequence, under the same distributional assumptions MLE is more efficient than the two-step estimator.

Consider the case when $\mathbf{X}_i^R = \mathbf{X}_i^Y$, i.e. the same variables are used in the selection equation and the outcome equations. Identification of the parameters $\boldsymbol{\beta}^Y$ in the two-step method now depends on the non-linearity of the inverse Mills ratio. The inverse Mills ratio is not linear in $(\mathbf{X}_i^R\boldsymbol{\beta}^R)$ but the function that maps $(\mathbf{X}_i^R\boldsymbol{\beta}^R)$ into the inverse Mills ratio is linear for certain ranges of $(\mathbf{X}_i^R\boldsymbol{\beta}^R)$. If $\mathbf{X}_i^R = \mathbf{X}_i^Y$ and the inverse Mills ratio is linear in $\mathbf{X}_i^R\boldsymbol{\beta}^R$, then there is perfect collinearity in the second step

$$Y_i = \mathbf{X}_i^Y\boldsymbol{\beta}^Y + \sigma\zeta(\mathbf{X}_i^Y\hat{\boldsymbol{\beta}}^R) + \epsilon_i^{Y^*}$$

For this reason, it is common to make an exclusion restriction. This involves including at least one of the regressors in $\mathbf{X}_i^R$ that is not included in $\mathbf{X}_i^Y$. Identification of the sample selection models hinges on the assumption of a bivariate normal distribution of the error terms. This causes a serious lack of robustness against misspecification.

The exclusion restriction only identifies $P(Y_i|\mathbf{X}_i^Y)$ if the instrumental variable predicts well whether or not Y_i is observed but the variable must be unrelated to the value of Y_i. This is no ideal situation, however it can be argued that there are variables influencing the participation decision that have no relationship with the survey item.

The two methods discussed for estimating the sample selection model, MLE and Heckman's two-step estimator, are fully parametric and have been widely criticised for their dependence on the normality assumption (Maddala, 1991, Johnston and DiNardo, 1997, Vella, 1998). A lot of research has been devoted to the relaxation of the distributional assumptions. Dubin and Rivers (1989) discuss and present alternative parametric methods to relax the normality assumption, for instance using a logit rather than a probit form of the regression equation. Vella (1998) presents a survey on estimation methods and discusses departures from the fully parametric models in the direction of relaxation of the distributional assumptions. Maddala (1991) proposes to use more generalized distributions, or semi-parametric methods. Nonparametric estimation methods are presented in Das et al. (2003). Gallant and Nychka (1987) propose a semi-parametric method that appears to be very practical. They use a Hermite series approximation for the joint distribution of the error terms.

A detailed description of the semi-parametric estimation method is beyond the scope of this thesis. Vella (1998) presents an overview of the various semi-parametric estimations methods. For an empirical application of the approach from Gallant and Nychka (1987), see Melenberg and van Soest (1993). We describe a semi-parametric approach to the sample selection model that is based on Heckman's two-step estimator, using the method from Gallant and Nychka (1987). In the first step, a semi-parametric estimator for the selection equation results in estimates for $\hat{\boldsymbol{\beta}}^R$. In the second step, the inverse Mills ratio is replaced by series approximation, i.e.

$$Y_i = \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \delta_1(\mathbf{X}_i^R \hat{\boldsymbol{\beta}}^R) + \delta_2(\mathbf{X}_i^R \hat{\boldsymbol{\beta}}^R)^2 + \ldots + \delta_K(\mathbf{X}_i^R \hat{\boldsymbol{\beta}}^R)^K + \epsilon_i^{Y^*} \quad (8.35)$$

Now $\boldsymbol{\beta}^Y$ and $\delta_1, \delta_2, \ldots, \delta_K$ can be estimated by OLS. Gallant and Nychka (1987) show that the estimates for $\boldsymbol{\beta}^Y$ and $\delta_1, \delta_2, \ldots, \delta_K$ are consistent if the number of approximations K increases with an increasing sample size. In practice, the choice of K can be guided by model selection criteria, see De Luca and Peracchi (2007). Van der Klaauw and Koning (1996) note that higher K lead to a lack of precision of the estimates, as well as a quick increase in computation times. This estimator does not have underlying distributional assumptions, however it does depend on the exclusion restriction due to the risk of collinearity in regression equation (8.35).

8.3.3.4 Multiple selection equations

In the sample selection model as described above, the selection equation that determines the censoring of the outcome is a function of one single index, namely the response probability ρ_i. However, we have argued and showed that it is important to distinguish between the different types of (non)response. In this section we modify the selection equation in the sample selection model to account for the sequential nature of the response process. Vella (1989) describes additional forms of censoring rules in the sample selection model. One of the forms Vella (1998) considers is based on more than one indicator or selection equation. This approach can be applied, in a slightly adapted way, to the sample selection model in the context of survey nonresponse with distinctive nonresponse types resulting in multiple selection equations. We illustrate this approach by regarding the simplified form of the response process, see figure 8.6. We only consider the response stages contact and participation. The generalisation to more response types is straightforward but cumbersome in notation. Now consider the modified form of the sample selection model in (8.25). The selectivity

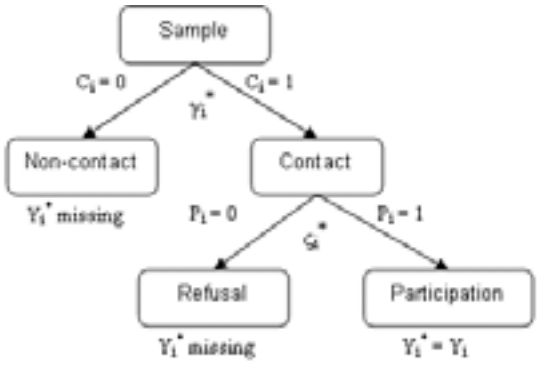

Figure 8.6: *The sample selection model with multiple selection equations*

bias is defined as a function of two distinct processes, being contact and survey participation. The modified sample selection model becomes

$$\begin{aligned} \gamma_i^* = &\quad \mathbf{X}_i^C \boldsymbol{\beta}^C + \epsilon_i^C, \\ \varrho_i^* = &\quad \mathbf{X}_i^P \boldsymbol{\beta}^P + \epsilon_i^P, \\ Y_i^* = &\quad \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \epsilon_i^Y \end{aligned} \tag{8.36}$$

for $i = 1, \ldots, n$. Recall that γ_i^*, ϱ_i^* are latent variables. We do not observe them, but instead we observe the sign of their underlying process. If $\gamma_i^* > 0$, the corresponding indicator C_i equals 1 and else $C_i = 0$. The same holds for P_i and ϱ_i^*. Furthermore, P_i is only observed when $C_i = 1$, otherwise it is censored. For the outcome equation holds that Y_i^* is a latent variable that is only observed when $C_i = 1$ and $P_i = 1$. In this situation we define Y_i as

$$Y_i = \begin{cases} Y_i^*, & \text{if } C_i = 1; P_i = 1 \\ \text{missing}, & \text{if } C_i = 1; P_i = 0 \text{ or } C_i = 0 \end{cases}$$

The error terms are assumed to follow a multivariate $N_3(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, or

$$\begin{pmatrix} \epsilon_i^C \\ \epsilon_i^P \\ \epsilon_i^Y \end{pmatrix} \backsim N\left(\mathbf{0}, \begin{bmatrix} 1 & \zeta_{CP}\sigma_Y & \zeta_{CY}\sigma_Y \\ \zeta_{PC}\sigma_Y & 1 & \zeta_{PY}\sigma_Y \\ \zeta_{YC}\sigma_Y & \zeta_{YP}\sigma_Y & \sigma_Y^2 \end{bmatrix}\right) \tag{8.37}$$

The covariance between the survey item Y and the contact different participation equation is denoted by $\zeta_{YC}\sigma_Y$ different $\zeta_{YP}\sigma_Y$. The correlation between

Y and contact different participation equals ζ_{YC} different ζ_{YP}. Furthermore, $\mathbf{\Sigma}$ is symmetric, i.e. $\zeta_{CP} = \zeta_{PC}$, $\zeta_{CY} = \zeta_{YC}$ and $\zeta_{PY} = \zeta_{YP}$.

Poirier (1980) describes a form of partial observability that we can use in the situation of survey nonresponse. The model that Poirier (1980) uses, does not observe C_i nor P_i but only their product $C_i \times P_i$. From his results follows that the conditional expectation of Y_i equals

$$E[Y_i|C_i \times P_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P, \mathbf{X}_i^Y] = \mathbf{X}_i^Y \boldsymbol{\beta}^Y + E[\epsilon_i^Y | C_i \times P_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P, \mathbf{X}_i^Y] \quad (8.38)$$

where

$$\begin{aligned} &E[\epsilon_i^Y | C_i \times P_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P, \mathbf{X}_i^Y] = \\ &\zeta_{YC}\sigma_Y \left(\frac{\phi(\mathbf{X}_i^C \boldsymbol{\beta}^C)\Phi(\mathbf{X}_i^P(\boldsymbol{\beta}^P - \zeta_{PC}\boldsymbol{\beta}^C)/\sqrt{1-\zeta_{PC}^2}))}{\Phi_2(\mathbf{X}_i^C \boldsymbol{\beta}^C, \mathbf{X}_i^P \boldsymbol{\beta}^P, \zeta_{PC})} \right) \\ &+ \zeta_{YP}\sigma_Y \left(\frac{\phi(\mathbf{X}_i^P \boldsymbol{\beta}^P)\Phi(\mathbf{X}_i^C(\boldsymbol{\beta}^C - \zeta_{PC}\boldsymbol{\beta}^P)/\sqrt{1-\zeta_{PC}^2}))}{\Phi_2(\mathbf{X}_i^C \boldsymbol{\beta}^C, \mathbf{X}_i^P \boldsymbol{\beta}^P, \zeta_{PC})} \right) \end{aligned} \quad (8.39)$$

That the results of Poirier (1980) are applicable to the survey nonresponse situation can be easily seen by realising that if $C_i = 1$ and $P_i = 1$ then $C_i \times P_i = 1$ and in all other cases $C_i \times P_i = 0$.

In the sample selection model with one selection equation, the two-step estimator results in the inclusion of the inverse Mills ratio in the regression equation. The additional terms in this case are more complicated due to the correlated structure of the multiple selection equations. This can be seen by looking at the case of no correlation between the contact and the participation equation. The last term in the condition expectation of Y_i, i.e. (8.39), then reduces to

$$E[\epsilon_i^Y | C_i = 1; P_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P, \mathbf{X}_i^Y] = \zeta_{YC}\sigma_Y \frac{\phi(\mathbf{X}_i^C \boldsymbol{\beta}^C)}{\Phi(\mathbf{X}_i^C \boldsymbol{\beta}^C)} + \zeta_{YP}\sigma_Y \frac{\phi(\mathbf{X}_i^P \boldsymbol{\beta}^P)}{\Phi(\mathbf{X}_i^P \boldsymbol{\beta}^P)}$$

This model can be implemented by separately estimating the two selection equations by probit estimation to obtain estimates for $\boldsymbol{\beta}^C$ and $\boldsymbol{\beta}^P$. Consequently, the corresponding two inverse Mills ratios can be computed and inserted in the regression equation for Y_i as omitted variables. Likewise, the modified sample selection model (8.36) can be implemented. First, the two selection equations are estimated by a bivariate probit model with sample selection, presented in the previous section. From this regression, the estimates for $\boldsymbol{\beta}^C$ and $\boldsymbol{\beta}^P$ are obtained. These can then be used to compute the two terms in the conditional expectation of ϵ_i^Y in equation (8.39). Insertion of these two terms in the regression equation for Y_i, adjusts the survey item for the sample selection bias.

8.3.3.5 Categorical regression equation

The sample selection model is designed for continuous survey items. The bivariate probit model with sample selection is designed for dichotomous variables. Nevertheless, most survey items have categorical outcomes with more than two categories. For the application of the sample selection model to adjustment for nonresponse bias in social surveys it is beneficial to develop methods that deal with categorical survey items, i.e. to develop a categorical regression equation. We distinguish between two types of categorical variables: ordered and unordered. We describe how the regression equation in the sample selection model can be adjusted for categorical variables in both cases. The development of the sample selection model for categorical survey items can use these suggestions but the actual adjustment of the model remains the topic of further research.

Categorical variables are variables that have a measurement scale consisting of a set of categories (Agresti, 2002). These variables are often encountered in social research for measuring attitudes and opinions. There are two types of scales: *nominal* and *ordinal*. Nominal categorical variables do not have naturally ordered categories. The order of listing of the categories is not relevant for statistical inference. The ethnic group a person adheres to is an example of a nominal categorical variable. Variables that do have an ordering are referred to as ordinal categorical variables. For instance, degree of urbanization is ordered, with categories based on a classification of the surrounding address density. In this case, the variable has a quantitative nature, with an underlying continuous distribution. Nominal variables are qualitative, the distinct categories differ in quality and not in quantity.

The methods to deal with these two types of categorical variables are different. For ordinal categorical variables, statistical inference can use the underlying continuous distribution. By regarding the value of the underlying continuous variable, it can be determined to what category a sample element belongs to. The continuous variable is used to form intervals for the different categories. Each of the categories is represented by a dummy variable that is one if the underlying continuous variable takes a value in the interval corresponding to the category, and zero otherwise. This corresponds to a regression where the response is in the form of grouped data.

Example 8.3.2 Employment status as an ordinal categorical variable
To illustrate the approach to ordinal categorical variables, we use an artificial example of the employment status as an ordinal representation of a propensity

to work. Let Y_i^* be a latent variable that is defined as the propensity to work. This latent variable can be modelled with the regression equation in the sample selection model (8.25), i.e.

$$Y_i^* = \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \epsilon_i^Y$$

The variable employment status consists of three categories: Employed, unemployed and not a member of the labour force. For each of these categories we introduce a dummy variable that takes a value 1 if the sample elements belongs to the corresponding category, and zero otherwise. Thus,

$$\begin{aligned} Y_i^{nl} &= \begin{cases} 1 & \text{if sample element i is not a member of the labour force} \\ 0 & \text{else} \end{cases} \\ Y_i^{u} &= \begin{cases} 1 & \text{if sample element i is unemployed} \\ 0 & \text{else} \end{cases} \\ Y_i^{e} &= \begin{cases} 1 & \text{if sample element i is employed} \\ 0 & \text{else} \end{cases} \end{aligned}$$

To determine which of the categories sample element i belongs to, we regard the work propensity Y_i^*. If the propensity to work is below some threshold c_1, then sample element i does not belong to the labour force. Above a certain threshold c_2, sample element i belongs to the active labour force and is employed. Between c_1 and c_2, sample element i does belong to the labour force, but the propensity is too low to become active. In that case, sample element i is unemployed. Put differently,

$$\begin{array}{ll} Y_i^{nl} = 1 & \text{if } Y_i^* < c_1 \\ Y_i^{u} = 1 & \text{if } c_1 < Y_i^* < c_2 \\ Y_i^{e} = 1 & \text{if } Y_i^* > c_2 \end{array}$$

If the underlying continuous variable follows a normal distribution, this model is the ordered probit model. The probabilities to belong to different categories of employment status are

$$\begin{aligned} P(Y_i^{nl} = 1) &= \Phi\Big(\frac{c_1 - \mathbf{X}_i^Y \boldsymbol{\beta}^Y}{\sigma}\Big) \\ P(Y_i^{u} = 1) &= \Phi\Big(\frac{c_2 - \mathbf{X}_i^Y \boldsymbol{\beta}^Y}{\sigma}\Big) - \Phi\Big(\frac{c_1 - \mathbf{X}_i^Y \boldsymbol{\beta}^Y}{\sigma}\Big) \\ P(Y_i^{e} = 1) &= 1 - \Phi\Big(\frac{c_2 - \mathbf{X}_i^Y \boldsymbol{\beta}^Y}{\sigma}\Big) \end{aligned}$$

■

In case of nominal categorical variables, the regression equation has to be substituted by a multinomial probit model where the categories of the dependent variable are unordered. Suppose that we observe a multinomial variable Y_i with $j = 1, \ldots, J$ categories. Y_i is modelled in terms of a latent variable $Y_i^* = (Y_{i1}^*, \ldots, Y_{iJ}^*)$ via

$$Y_i = \begin{cases} 0, & \text{if } max(Y_i^*) < 0 \\ j, & \text{if } max(Y_i^*) = Y_{ij}^* \end{cases} \tag{8.40}$$

with

$$Y_{ij}^* = \mathbf{X}_{ij}^Y \boldsymbol{\beta}^Y + \epsilon_{ij}^Y \tag{8.41}$$

for $i = 1, \ldots, n$ and $\epsilon_{ij}^Y \backsim N_J\Big(\mathbf{0}, \boldsymbol{\Sigma}\Big)$. Now, for sample element i we observe one of J possible outcomes with corresponding probabilities $p_{i1}, \ldots, p_{iJ}$. The multinomial probabilities are given by

$$p_{ij} = P(\mathbf{X}_{ij}^Y \boldsymbol{\beta}^Y + \epsilon_{ij}^Y > \mathbf{X}_{ik}^Y \boldsymbol{\beta}^Y + \epsilon_{ik}^Y; \forall k \neq j) \tag{8.42}$$

Estimation of these probabilities involves the calculation of multiple integrals of the multivariate normal distribution and is computationally intensive. Literature suggest a number of alternative estimation methods, for instance the Bayesian Gibbs sampler proposed by Albert and Chib (1993) or the Markov Chain Monte Carlo technique using Bayesian statistics as described in Imai and Van Dyk (2005).

Example 8.3.3 Employment status as a nominal categorical variable
Now, let us regard the variable employment status as a nominal variable, instead of an ordinal variable as described in example 8.3.2. We introduce a multinomial variable employment status Y_i with three categories, i.e.

$$Y_i = \begin{cases} 1, & \text{if person i is employed} \\ 2, & \text{if person i is unemployed} \\ 3, & \text{if person i does not belong to the labour force} \end{cases} \tag{8.43}$$

Furthermore, we introduce three latent variables Y_{i1}^*, Y_{i2}^* and Y_{i3}^* as

$$\begin{aligned} Y_{i1}^* &= \mathbf{X}_{i1}^Y \boldsymbol{\beta}^Y + \epsilon_{i1}^Y \\ Y_{i2}^* &= \mathbf{X}_{i2}^Y \boldsymbol{\beta}^Y + \epsilon_{i2}^Y \\ Y_{i3}^* &= \mathbf{X}_{i3}^Y \boldsymbol{\beta}^Y + \epsilon_{i3}^Y \end{aligned}$$

for $i = 1, 2, \ldots, n$. For sample element i we now observe category j of Y_i for which the latent variable Y_{ij}^* takes the largest value. In this situation, we have three underlying propensities corresponding to the different categories of the survey item employment situation whereas in the situation of an ordinal categorical variable Y as described in example 8.3.2 we only had one underlying propensity. We now observe the category of employment situation for which the underlying propensity is the largest. ■

8.3.4 Summary

The evaluation of nonresponse analysis models described in section 8.2.5, as well as in Chapter 3, motivate the incorporation of different response types, as does a lot of the recent survey literature (for example Groves et al. 2002, Bethlehem and Schouten 2004, Stoop 2005, Durrant and Steele 2007). Until now, most of the literature on different response types has focussed on methods to analyse nonresponse.

In this section, we describe how the methods described in section 8.2 can be used to adjust for nonresponse bias. The nested logit models allow for a sequential weight adjustment approach as suggested by Groves and Couper (1998) and Iannacchione (2003). However, this approach uses only the relationship between the response and background variables. It does not take into account the relationship with the survey items. We consider the case of discrete and continuous survey items. For discrete bivariate variables, the bivariate probit model with sample selection can be used to adjust for nonresponse bias. Standard statistical software is available to analyse situations where there is one selection equation, and a bivariate survey item. The sample selection model proposed by Heckman (1979) describes a model that can be used for continuous survey items. We discuss how this model can be estimated for the situation of survey nonresponse based on one selection equation. We discuss three estimation methods, Maximum Likelihood Estimation, Heckman's two-step estimator and a semi-parametric estimation procedure. The parametric estimation methods are highly dependent on the distributional assumption of the models. Therefore, we shortly discuss one semi-parametric procedure. However, this method is not frequently applied in practice.

There are no standard procedures to use the sequential response process with several response types. Therefore, we discuss how the bivariate probit model with sample selection and the sample selection model can be extended to multiple selection equations. Furthermore, a lot of survey items are categorical.

We present extensions to the models to allow for categorical survey items. We present models for both ordinal and nominal categorical variables.

8.4 Concluding remarks

Intuitively, the methods described in section 8.3.3 should be preferred over methods that do not account for different response types, or independence between the response types. Unfortunately, this remains an intuition since we have yet to develop software for these models. Recent developments in econometrics describe a number of ways to estimate these models, for instance with Markov Chain Monte Carlo methods or Bayesian estimation. This will avoid complicated evaluations of multiple integrals of the multivariate normal distribution.

Chapter 9

Nonresponse Adjustment in Mixed Mode Surveys

Nowadays it is common practice to combine data collection modes in one survey. In this chapter we provide an overview of different data collection modes and their combined use, so-called mixed mode surveys. In this chapter, we first describe mixed mode data collection. Next, we discuss the use of combined data from mixed mode surveys for statistical inference. We thereby focus on methods to adjust for nonresponse bias in mixed mode surveys.

9.1 Introduction

In Chapter 3 we analysed the nonresponse to the first round of the Dutch Labour Force Survey. This is a CAPI survey. We re-approached nonrespondents to the Dutch LFS using a combination of CATI, mail- and web surveys, see Chapter 4. These are some examples of the various data collection modes that can be used to conduct household surveys.

Modes differ in various aspects. In Chapter 7 we showed how CAPI and CATI differ in coverage. With CATI, only households with a listed land-line telephone can be reached whereas CAPI does not have this restriction. However, CAPI is much more expensive due to the assistance of interviewers that travel to the selected addresses. The mail survey in the re-approach of the nonrespondents to the Dutch LFS (Chapter 4) had a very low response rate. Also, the response to the web survey was low. Moreover, for web surveys it is known that not

everybody has an internet connection and/or the ability to use the internet to answer surveys. However, the mail and web surveys were chosen because of their lower costs compared to CAPI and CATI.

Because each individual data collection mode has its shortcomings and its benefits, mixing data collection modes provides an opportunity to compensate for the weakness of each individual mode. This can reduce survey costs and at the same time increase the response. It may even be possible to reduce the selectivity of the response beforehand. For this purpose, sampled persons or households can be allocated to a specific mode based on known background characteristics. If a specific group in the population does not participate in one mode and these persons are willing to participate in another mode, this will reduce the selectivity of the response.

This chapter begins with an introduction to mixed mode surveys. First, we discuss mode differences and mode effects in section 9.2. In section 9.3 we introduce a number of designs for mixed mode surveys. Next, we discuss the integration of data from different modes. We describe how we can combine data collected by different modes for nonresponse adjustment in section 9.4. We do not regard measurement errors and coverage errors, but instead we focus on nonresponse bias alone. In section 9.5 we present three approaches that can be followed to adjust for nonresponse bias in mixed mode surveys. The approaches vary from straightforward linear weighting to a simultaneous approach that makes use of the available paradata in each mode. A description of paradata is provided in section 9.4.2. In section 9.6 we illustrate the approaches by an application to the pilot Safety Monitor. The three approaches are not exhaustive, and there are some unresolved issues. Therefore, section 9.7 ends this chapter with concluding remarks and suggestions for further research.

9.2 Data collection modes

9.2.1 Introduction

The *data collection mode* is the means of communication by which the questionnaire is presented to the sample elements and by which the answers are registered. An important aspect of data collection modes, is how they are administered. A distinction is made between interviewer-assisted modes and self-administered questionnaires. Another main aspect is the way the data is collected (laptop, telephone, mail or web). Based on these aspects, we can make the following distinction between modes, see table 9.1. The addition of an 'I' stands

Table 9.1: *Different data collection modes*

	Administration	
Data collection by	*Interviewer*	*Self*
Laptop	CAPI	CASQ
Telephone	CATI	IVR
Mail	PAPI	mail survey
Web	CAWI	web survey

for Interview and refers to interviewer-assisted modes. 'CA' refers to Computer Assisted. 'P', 'T' and 'W' stand for different Personal, Telephone and Web. 'SQ' means Self-administered Questionnaire. 'PAPI' stands for Paper And Pencil Interview, and 'IVR' for Interactive Voice Response, which is an automated telephone interviewing system, Couper (2005).

The most common modes in survey research are CAPI, CATI, mail and web surveys (see also Pierzchala, 2006). In this chapter we focus on these four modes. *CAPI* stands for Computer Assisted Personal Interview. The interviewer visits respondents at home and administers the computerised questionnaire from a laptop. CAPI surveys are often referred to as face-to-face surveys, however a face-to-face survey can also be a personal interview where the interviewer answers the questions and records the respondents' answers on a paper questionnaire. CAPI surveys allow for a long, complex routing structure of the questionnaire because the routes are controlled by the computer. Furthermore, the data can be checked and corrected during the interview. Besides that, more complex surveys can be conducted due to the assistance of the interviewer that can explain the questions and probe for accurate answers. Furthermore, the interviewer can use tactics to gain cooperation from the respondents, for example by employing the leverage-salience theory described in Chapter 2. Visual aids can be used when needed, for example response cards. In general, CAPI interviews have a high response rate due to the interaction between the interviewer and the respondent. The interviewer can build rapport and confidence in the interaction with the respondent. CAPI surveys have a high coverage of the population.

Thus, the advantages are manifold and, therefore, in general CAPI has the highest data quality. However, there are also disadvantages associated with CAPI. These disadvantages are mainly caused by the interviewers. First, there are high costs associated with the interviewers travelling to the respondents. Furthermore, some types of questions are sensitive to social desirability bias caused by the interviewer presence, see also section 9.2.3. In addition, the interviewer

behaviour may lead to interviewer effects; often referred to as interviewer variance. Sample elements are clustered within interviewers. As interviewers have their own way of acting and reacting this may cause a cluster effect, or interviewer variance. To prevent interviewer variance, interviewer procedures can be standardized. An example of a standardized procedure is the contact strategy presented in Chapter 3.

CATI stands for Computer Assisted Telephone Interviewing. Like in CAPI, the questionnaire is computerised, but in CATI it is administered over the telephone. Originally, it is implemented by use of a central call centre, where a call management system is used to distribute telephone numbers over the interviewers present. This allows for an optimal exploitation of the sample, in contrast to CAPI where there is usually a higher proportion of unprocessed cases due to interviewer illness or vacation. In CAPI it is more cumbersome to re-distribute addresses to other interviewers in case the interviewer becomes ill, whereas with a centralised CATI system the call management system keeps track of all sample elements and can therefore easily re-distribute sample elements. Technologies that facilitate decentralised CATI surveys are, for example, the interviewer performing a CATI survey from home as is done in Sweden since the mid 1980s (Bergman et al. 1994). Because of the interviewer assistance, CATI is similar to CAPI. However, there are some important distinctions between these two modes. First of all, CATI is less prone to interviewer variance and social desirability bias because of the less personal communication channel. Some research seems to indicate that respondents are more inclined to answer sensitive questions in CATI (than in CAPI). Also, CATI surveys are more easy to set up and hence can be performed more quickly than CAPI. They are less expensive because no travel costs are involved.

Compared to CAPI, CATI is however less flexible in the sense that the questions cannot be very complicated, no visual aids can be used and the optimal duration of an interview is shorter than in CAPI surveys (approximately 30 minutes, Biemer and Lyberg 2003). The interviewers in CATI can make less rapport to the respondents, which leads to less probing for accurate answers and less persuading power to enhance participation. As a consequence, the response rates in CATI surveys are usually lower. And last but not least, the coverage of CATI surveys is lower than CAPI surveys. Sample elements without a listed telephone are not covered by CATI surveys.

Mail surveys are frequently referred to as PAPI. This is not deemed fit as the extension 'I' refers to an interviewer present and this is not the case with a mail survey. When an interviewer is present, mail surveys should be referred to as face-to-face surveys. In a mail survey, a paper questionnaire is sent to a sample

element, who answers the questions without interviewer assistance and sends it back over the mail. Mail surveys are self-administered. The main difference with CAPI and CATI is the absence of an interviewer. Therefore, the quality in a mail survey depends heavily on the questionnaire design. Compared to CAPI and CATI surveys, mail surveys are inexpensive. Furthermore, there is a reduced risk of social desirability bias. Compared to CATI surveys, an advantage of mail surveys is the possibility to use visual aids. Plus, the respondents can answer the questions whenever they like and they have more time to think about the questions and give more thoughtful answers. Like CAPI surveys, mail surveys have a high coverage of the population. However, mail surveys lack the advantages of interviewer assistance. Therefore, the response rate to a mail survey is lower than for interviewer-assisted surveys. The questionnaire cannot be too complicated and there is a larger risk of item nonresponse. Additionally, mail surveys need a longer fieldwork period as well as a longer processing time due to the data entry and editing afterwards. There is no control over the response process, which complicates the calculation of response rates and the adjustment for nonresponse. The type of response is unknown, it is only observed whether completed questionnaires are returned, or not.

Web surveys are often mistakenly referred to as CAWI. In a web survey the respondent is invited to a web page where the questionnaire is found. There is no interviewer present. A survey is CAWI when assistance is provided by means of a web cam, but this situation is very uncommon. Web surveys are self-administered surveys like mail surveys, therefore sharing the advantages of a lower risk of social desirability bias and lower costs. Likewise, the disadvantages of self-administered surveys concern the lack of interviewer assistance in gaining cooperation and probing for accurate answers. Compared to mail surveys, web surveys have more questionnaire design choices. Also, the data collection process is more controlled by using computerised questionnaires. This facilitates routing and data editing during the fieldwork which reduces the risk of measurement errors and makes it possible to use more complex questionnaires. However, compared to mail surveys web surveys have a lower coverage. The part of the population that has no access to the internet cannot be reached. Another issue in web surveys concerns the confidentiality and security of sending information over the internet. Finally, the risk of drop-outs is high because there is no interviewer present to stimulate participation.

Another development in survey research is the use of administrative data, or registers. Statistics Netherlands is by law obliged to use data from registers for the production of statistics, unless these data fail to meet certain quality requirements (timeliness, completeness, accuracy), see the act of Statistics' Netherlands

(2004) (in Dutch: Wet op het Centraal Bureau voor de Statistiek). A lot of research is devoted to the use of registers in the production of statistics, see for example Wallgren and Wallgren (2007). The advantage of registers is, that the data already exists and is hence relatively inexpensive to use. Furthermore, the response burden is greatly reduced by not asking respondents the same information twice. However, registers may have some disadvantages as well. As the data in registers is collected, updated and coded by an organisation other than the survey organisation, the data collection process is not controlled by the survey organisation. This change of roles makes it hard to judge the quality of the obtained data. The quality of the data strongly depends on the importance for and use by the register holder, as well as its means to process and control data. Furthermore, there may be conceptual differences leading to other definitions of data. As most organisations that have a register are not primarily aimed at producing statistics, there usually is no knowledge of statistics involved in the data collection process which may lead to imperfect data (erroneous or not-missing-at-random). Registers can also be employed as one of the modes in a mixed mode data collection design. For instance, if the register does not cover the entire population and a survey is conducted to complete the information. However, describing the use of registers is beyond the scope of this thesis and in the remainder of this chapter we do not regard registers anymore.

To summarise, CAPI, CATI and web surveys share advantages of CAI as described in the introduction to this chapter. CATI and CAPI surveys have the advantage of interviewer assistance. Mail and web surveys are cheaper, and do not have the disadvantage of interviewer variance and social desirability bias caused by the interviewers. In self-administered data collection modes like mail and web surveys, the survey topic plays a larger role in the response process, possibly leading to a more selective response. As De Leeuw (1992) notes, the largest mode difference is caused by the difference in administration: interviewer-assisted versus self-administered. We describe mode differences in more detail in the following section. Then, in section 9.2.3 we focus on the effects of these differences when combining modes in a mixed mode survey setting.

9.2.2 Mode differences

De Leeuw (1992) performed a meta-analysis on face-to-face and telephone surveys and considered several aspects of data quality. She concluded that differences in data quality between well conducted face-to-face and telephone surveys are small. This conclusion is in line with Groves (1989), who states that the most consistent finding in studies comparing responses in face-to-face and tele-

phone surveys is the lack of difference between the two modes. Snijkers (2002) compares CATI to CAPI with respect to cost and quality. The advantages of CATI concern the efficient use of hardware and software (in case of centralised telephone interviewing), the immediate availability of the data and the reduced costs because the interviewers do not make travel costs and a smaller number of interviewers is needed for the same number of interviews.

The main difference between a CAPI and a CATI survey is the way of communicating. Groves (1989) distinguishes two differences: the 'channel capacity' and the 'intimacy of the social interaction'. A face-to-face survey is capable of carrying more types of messages (e.g. non-verbal). Telephone calls received from strangers tend to be short interactions (more frequently used for business purposes) and the intimacy of the social interaction is less personal. Snijkers (2002) states that due to the use of the telephone as a communication medium, telephone interviewing is only adequate for simple questions that can be answered instantaneously and that need little time to input the answers. This implies a questionnaire with simple question wordings, short lists of response categories in closed-ended questions and a shorter duration of the questionnaire.

De Leeuw (2005) presents an extensive overview of the advantages and disadvantages of mixed mode survey designs, thereby distinguishing between the use of multiple modes in the contact phase and the actual data collection phase. In this chapter, we focus solely on multiple modes for data collection. In her research, De Leeuw (2005) emphasizes the difference in *cognitive burden*, i.e. the burden of the processing of information related to the decision making, of the modes. She thereby adheres to the stimulus-response model of survey responding (Tourangeau and Rasinski, 1988, Snijkers, 2002, Ariel et al., 2008). According to this model, the cognitive process of survey response can be divided into four phases. In each of the phases, the respondent has to perform cognitive tasks that can be different across modes. The first phase concerns the interpretation and comprehension of the question. This depends on question wording, syntax, length, order, perceived sensitivity and the response categories. Then, the information has to be retrieved from either the respondents' memory or other sources. This second phase depends on the type of information that is being asked, whether this is for instance factual, attitudinal or proxy. Phase three concerns the judgement. In this phase the respondent uses the retrieved information to form an internal answer. The retrieved information is integrated and evaluated. The judgement of the answer depends on the consistency with prior answers, or with social desirability. These two steps together, the information retrieval and judgement, can be seen as the information transmission phase. And last, the respondent reports the answer by comparing the internal answer

with the response options or by reporting the answer in the requested quantity. For this phase it is crucial that the respondent comprehends the wording of the response options.

Based on the response-stimulus model, Pierzchala (2006) identifies attributes of surveys that affect cognition and response. Seven aspects of modes are considered, being: presentation, transmission of response, segmentation of the questionnaire, dynamic or passive medium, type of administration, who/what determines the pace of the interview and the channel of communication. De Leeuw (2005) groups these factors into three classes: medium-related factors (presentation, transmission, segmentation and dynamic/passive), factors concerning the information transmission (administration, pace and channel of communication) and interviewer effects. Interviewer effects interact with medium-related factors and information transmission factors. Pierzchala (2006) investigates what aspects are responsible for most mode effects. He identifies three so-called critical dimensions. These are: an aural versus a visual presentation, self-administered versus interviewer-administered, and dynamic versus passive nature of the questionnaire[1]. The first and third dimension relate to the information transmission, and the second is medium related. 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, the higher the risk of mode effects. See table 9.2. In this table, 'na' stands for not applicable.

CAPI and CATI are similar in presentation, administration and in the dynamic nature of the questionnaire. Accordingly, we describe the degree of disparity of the other modes with respect to both CAPI and CATI. Web surveys also have a dynamic questionnaire. But the presentation is visually 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, the degree of disparity therefore is 3. Mail surveys and web surveys are similar in self-administration and visual presentation, however mail

[1]CAI-questionnaires are dynamic and paper questionnaires are passive.

Table 9.2: *Degree of disparity between data collection modes*

Mode combination	CAPI/CATI	Mail	Web
CAPI/CATI	na	3	2
Mail	3	na	1
Web	2	1	na

surveys have a passive questionnaire. Their degree of disparity is 1. Hence, the largest mode differences are expected for a combination of mail surveys and CAPI or CATI, followed by web surveys and CAPI or CATI surveys. 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.

However, there are other aspects that play a role when choosing which modes to mix. Roberts (2007) identifies budget restrictions, timeliness and the available infrastructure to facilitate mixed mode designs. Biemer and Lyberg (2003) discuss factors that influence the decision process in choosing modes. These are: the desired level of data quality, the available budget and time, and the specific content of the survey (types of questions, length of survey, complexity of the questions, need for visual aids). Furthermore, they acknowledge that choosing an optimal design is especially difficult in situations where there are a lot of options.

Example 9.2.1 Choosing a mixed mode design
In 2005 - 2007, Statistics Netherlands has performed four mixed mode pilots. The aim of this research was to determine an optimal mixed mode design for the surveys from Statistics Netherlands. See Janssen et al. (2007) for a preliminary outcome of the research. Statistics Netherlands' primary interest was reducing the costs of data collection, while preserving the quality of the data. Especially, the possibilities for using a web survey as one of the data collection modes were investigated. The pilots were conducted on three different surveys. Three of the pilots were linked to a reference survey which was conducted approximately at the same time. In all surveys, the sample elements were persons. Table 9.3 gives an overview of the pilots together with their basic design. Where 'ICT' refers to the survey about Information- and Communication Technology, 'SM' stands

Table 9.3: *Details of the four pilots at Statistics Netherlands in 2005 to 2007*

Pilot	*Modes in pilot*	*Size*	*Original mode*	*Period*	*Population*
ICT	I: Web, mail	8,691	CATI	2005	12 - 74 y
	II: CATI, Web, mail	1,832		2005	
SM-1	Web, CATI, CAPI	3,750	CATI/CAPI	2006	$>$ 14 y
SM-2	Web, CATI, CAPI	3,000	CATI/CAPI	2007	$>$ 15 y
IE	CAPI	1,857	No reference	2006	$>$ 15 y
	Web, mail $\rightarrow$ CATI	1,867		2006	

for Safety Monitor and 'IE' refers to the Informal Economy Survey. In table 9.4 we present the mixed mode strategies used in the pilots. In three of the four pilots the mixed mode strategy consisted of a sequence of attempts with different modes. In the pilots a maximum of four attempts can be distinguished. In case of a web survey, the respondent was sent an advance letter containing a login to a secured Statistics Netherlands' website. In case mail was combined with web, the respondent could also apply for a paper questionnaire by a return postcard. The first pilot was linked to the ICT survey. This survey is CATI and deals with the usage and knowledge of ICT within households and by individuals. The pilot was directed at the design of mixed mode strategies including web as a mode, the composition and size of response in such strategies and the logistics that are concerned with the data collection. In the pilot two groups were formed. The first group was a new sample. In the third attempt this group received a reminder either by mail or by telephone. The second group consisted of nonrespondents (refusals and non-contacts) in the regular ICT survey. Hence, their first attempt was CATI. The subsequent attempts mimicked that of the first group, but without telephone reminders. From the results of the ICT pilot, Janssen (2006) concluded that it is technically and logistically possible to incorporate web surveys into survey designs of Statistics Netherlands. Furthermore, he found that the telephone reminder resulted in a small, cost-ineffective increase in response rate.

The ICT pilot formed the basis for logistic infrastructure in the other pilots. The second and third pilot both have the Safety Monitor as the survey of reference. The main topics of the Safety Monitor are the perception of security and police performance. The survey itself is a mix of CAPI and CATI. Persons with a listed land-line telephone are interviewed in CATI, the remaining persons are interviewed CAPI. We will refer to the pilots as SM-1 and SM-2. The pilots

Table 9.4: *The mixed mode strategies in the four pilots*

Approach	*Pilot*			
	ICT		SM 1 & 2	IE
	I	II		
1	Web/mail	CATI	Web	Web/mail
2	Web	Web/mail	Web	Web
3	Web/mail	Web	Web	Web/mail
4		Web/mail	CAPI/CATI	CATI

differ from the regular survey since CAPI/CATI is preceded by a web survey. SM-2 was designed following the conclusions and recommendations of the pilot SM-1 and is, therefore, a continuation of that pilot. In both pilots all persons first received an advance letter with a login that gave access to a website where they could fill in a web questionnaire. Nonrespondents were assigned to CATI in case they had a listed land-line telephone, and otherwise were assigned to CAPI.

The main difference between the two pilots, is that in SM-2 respondents were made aware of the (interviewer assisted) follow-up survey in case of nonresponse to the web survey. This resulted in a higher initial response to the web survey, and less refusals in the follow-up survey. Apparently, the announcement of the follow-up led to less irritation with the respondents, which resulted in a higher response in the follow-up. Furthermore, it is possible that respondents wanted to avoid the visit of an interviewer and therefore the announcement led to an increase in the response to the web survey.

The third pilot, Informal Economy, was not linked to an existing survey. It deals with undeclared labour, or black wages. Within the pilot it was decided to allocate one half of the sample to CAPI and one half of the sample to a mix of web/mail and CATI. Like the pilots SM-1 and SM-2, the second half of the sample received a letter with a web login and an application form for a paper questionnaire. Nonrespondents were assigned to CATI, but only if they had a listed land-line telephone. No attempts were made to contact nonrespondents without a telephone. Hence, for this pilot we compare the two strategies where the CAPI survey can be regarded as the reference survey. This pilot is referred to as IE.

In general, it was possible to realise a (direct) cost reduction of 25%. However, the response obtained in the web survey did not exceed 35%. Therefore, it was decided that web surveys should always be combined with other data collection modes. This conclusion is in line with Biemer and Lyberg (2003). Janssen et al. (2007) suspect though that it is possible to obtain a higher response. A large number of persons indicated that they forgot to answer the questionnaire, did not feel like it or were too busy. They recommend further fine tuning of the fieldwork strategy for web surveys to convert these nonrespondents to respondents, for instance by using incentives. ■

9.2.3 Mode effects

Biemer and Lyberg (2003) distinguish between two sorts of mode effects: A pure mode effect is a difference in survey outcomes that originates solely from the choice of data collection mode. A design mode effect is a difference in survey outcomes that is caused by the combination of differences in total design, that are inherent to combining different modes. For example, the fieldwork strategy for interviewer-assisted surveys can cause a difference in survey estimates besides the effect that interviewers have (i.e. interviewer variance, social desirability bias, higher response, better quality answers). Pure mode effects are difficult to assess, therefore our focus lies on design mode effects, from now on referred to as mode effects. Differences in survey outcomes can be caused by the error sources that we described in Chapter 1. The choice of data collection mode can influence three non-sampling error sources (Roberts, 2007, Ariel et al. 2008): coverage error, nonresponse bias and measurement error.

Ariel et al. (2008) performed a literature study on mode specific measurement errors. They concluded that measurement errors indeed exist, and should be accounted for. The main causes for measurement error are social desirability, satisficing and context effects. *Social desirability bias* can occur in phase three of the response-stimulus model, if respondents change their 'internal' answer to an answer that they perceive as more social desirable. This behaviour leads to under reporting of perceived social undesirable behaviour such as alcohol consumption, and to over reporting of perceived social behaviour such as voting. *Satisficing* implies that persons reduce the cognitive burden of optimal question answering by short cutting the response process (Krosnick, 1991). The degree of satisficing is related to respondents ability, motivation and task difficulty. In satisficing, one or more phases in the response-stimulus model are skipped. Satisficing can manifest in respondents more often choosing no opinion answers, social desirable answers, non-differentiation between survey items when asked for rating, or acquiescence response bias, i.e. the tendency to agree with any statement regardless of its content. *Context effects* can arise due to external factors that influence phase three and four of the response-stimulus model, where the respondent comes up with an answer to the survey question.

Coverage errors arise due to the sampling frame for data collection modes. When sample elements do not have an entry in the sampling frame, their selection probability is zero and we speak of *undercoverage*. When sample elements are not in the population, but they do have an entry in the sampling frame we speak of *overcoverage*. The target population of Statistics Netherlands' surveys is the population that can be found in the population register, with the excep-

tion of the institutionalized population, see Chapter 3. The population register functions as the sampling frame for all surveys. Hence, CAPI and mail surveys have a low risk of coverage errors. For CATI and web surveys, however, sample elements need a telephone or access to the internet. The sampling frame for a telephone survey is obtained by matching the records from the population register with the register from the Dutch telephone company (KPN). Originally KPN distributed all land-line telephones. However, nowadays there are also other providers of land-line telephones (mostly digital). Therefore, the sample is also matched with the telephone book, where numbers from other providers are also listed too. As a consequence, sample elements without a listed land-line number (i.e. mobile only, an unlisted telephone or no land-line telephone) do not appear in the sampling frame for the telephone survey.

An alternative to sampling from a telephone directory is Random Digit Dialing (RDD). This method randomly generates telephone numbers from the frame of all possible telephone numbers. Lepkowski (1988) provides a review of RDD and other telephone survey sampling methods. The advantage of RDD compared to sampling from a telephone directory is that it provides coverage of both listed and unlisted numbers (and possibly mobile numbers as well). However, there are some disadvantages. A large problem of RDD is the number of ineligible listings. Numbers can belong to businesses, or households occur in the sample more than once because they own more than one number. It is thus unknown whether the generated telephone numbers correspond to eligible sample elements. With respect to adjustment methods, a disadvantage of RDD is that there is no auxiliary information at all available on the nonrespondents. Sometimes there is no difference between non-existing telephone numbers and nonrespondents, in which case non-existing numbers are wrongfully regarded as nonrespondents.

Web surveys at Statistics Netherlands are performed by sending an advance letter to sample elements, with a login to a secured website where the questionnaire can be found. Sample elements are asked to fill in the form on the internet. If they have no access to the internet, they cannot respond to the survey. CATI and web surveys, hence, have to do with undercoverage of the part of the population that does not have a listed land-line telephone, or internet access. The number of listed land-line telephones in the Netherlands is decreasing, whereas the number of households that have access to the internet increases, see example 9.2.2. It seems that the roles are reversing. This is also caused by the larger proportion of households that have no listed land-line telephones but mobile telephones only. However, having access to the internet does not imply that persons actually know how to use the internet to answer ques-

tionnaires. For example, persons can have access to the internet at the public library, or at their work where they may not be able or allowed to answer questionnaires during work time. A solution to undercoverage of persons without internet access is provided in the MESS-panel (MESS stands for: Measurement and Experimentation in the Social Sciences, see Das et al. (2006)). The sample for the MESS-panel is probability based, and the restriction of internet access for participation is overcome by providing respondents without internet access with a WebTV.

Example 9.2.2 Internet access and telephone ownership
Figure 9.1 displays the number of persons that have access to the internet and the number of persons that have a listed land-line telephone as measured in Statistics Netherlands' surveys. The number of persons with a listed land-line telephone is decreasing, while more and more persons have internet access. The statistics for internet access are based on the ICT survey in 2005 to 2007 where the following question was asked: does anybody in your household have access to the internet? The ICT survey is a telephone survey and it appears that the reported internet coverage is lower for persons without a listed land-line telephone (Janssen et al. 2007). This effect is related to age: the younger, the larger the difference in internet coverage between listed and unlisted persons. Younger persons more often have access to internet, and less often a listed land-line telephone. The reported internet coverage should thus be regarded with caution.

The statistics for telephone ownership are based on the percentage of linked telephone numbers to the personal surveys from Statistics Netherlands in 2005 – 2007. So it seems that telephone surveys with listed land-line telephones only become less attractive whereas the coverage of web surveys is increasing. However, the composition of the internet population is selective with respect to some important demographic and socio-economic characteristics. See the figures in 9.2. These figures confirm the general expectations with regard to the selectivity of the internet population. In graph 1 we see that especially persons above 65 less frequently have access to the internet. However, the internet access in this group is increasing more rapidly then the access in the other age groups. Furthermore, access to the internet increases with the educational level, as well as with the disposable income (graph 2 resp 3). ■

We already showed that nonresponse is often selective with respect to demographic characteristics like age and household composition, and socio-economic characteristics like income, educational level and employment situation, see

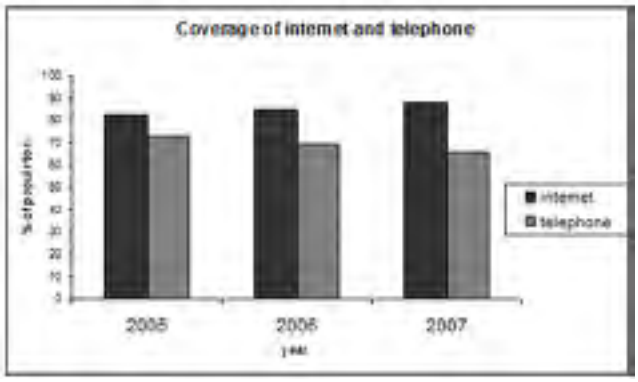

Figure 9.1: *The percentage of persons with access to internet and a listed landline telephone in the Netherlands 2005 - 2007*

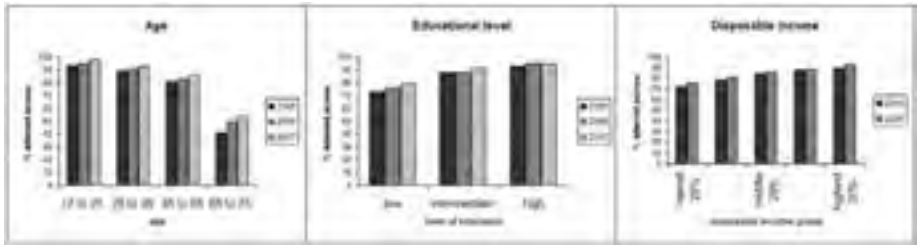

Figure 9.2: *Access to internet according to age, educational level and disposable income*

Chapter 3. Different modes lead to different response levels and thus to different compositions of the respondent pools within modes. As a consequence, nonresponse bias may have a different impact when differentiated to the mode of data collection. Mode comparisons are usually made between two modes. Roberts (2007) reports on a comparison of CAPI to CATI (done by Holbrook et al. in 2003) that finds that CATI respondents are higher educated, have a higher income, are younger and less frequently non-native. Web/mail respondents are usually better educated and more literate than CATI respondents. Kwak and Radler (2002) compare web and mail respondents in a survey among students. They find that web respondents more often are male and in general younger than mail respondents. Their restricted target population complicates

the generalisation of their results to the population, however they do find consistent evidence for their findings in the literature. De Leeuw (1998) compares mail, face-to-face, telephone and CATI surveys respondents. She compares the respondents to the population and finds that the selectivity is the same for all modes. The only difference between modes is found for gender and marital status, in mail versus face-to-face surveys. In mail surveys, there are more male and married respondents. In face-to-face surveys, there are more women and divorced and widowed respondents. It should be noted that the results of these analyses may be country-specific.

Example 9.2.3 Measurement error and nonresponse bias in mixed mode pilots

In example 9.2.1 we presented four mixed mode pilots conducted at Statistics Netherlands. We used the pilots to investigate measurement errors in the different mixed mode strategies. It should be noted, however, that the pilots were not designed to investigate measurement errors. The research shows that respondents may present answers that are significantly different for different modes.

For the IE pilot, differences were observed in the weighted percentage of undeclared employment for the two mixed mode strategies that were followed; CAPI and the mix of web, mail and CATI, see Gouweleeuw and Eding (2006). These differences are significant at the 5% level. Moreover, observed differences are also significant for a subset of undeclared employment activities, namely the weighted percentage of undeclared employment for which a financial compensation is offered. However, for this pilot we cannot completely disentangle measurement errors from nonresponse bias. The survey estimates are adjusted for nonresponse bias by weighting with a set of auxiliary variables. Within the strata that are formed by the weighting adjustments, nonrespondents may still be different from respondents. We can, therefore, not exclude that observed differences are attributable to a missing data mechanism that is not-missing-at-random.

According to the literature, see e.g. Dillman (2000), self-administered modes are expected to produce more honest answers to sensitive questions. We regard the main topics of the IE pilot as sensitive to the respondents. The IE mode strategy contained two self-administered modes; web and mail. However, surprisingly, both web and mail lead to estimates of the percentage of undeclared employment that are lower than the estimates following from the CAPI survey. We found that the CAPI estimates resemble the web estimates most. Adding mail and CATI answers only increases the differences in estimates for the proportion of individuals that have income from undeclared employment. We do

not have an explanation for this phenomenon other than selective response.

The SM-1 pilot shows that incorporating web in the mixed mode strategy significantly (at a 5% level) affects the answer patterns to scalar questions. Scalar questions are derived from a number of underlying statements. For instance, the scale variable that measures the opinion on the availability of the police is based on five statements. Respondents have to indicate whether they agree, disagree, are neutral or don't know with respect to each statement. A different score is attached to each answer category, in this case score 0 for agree, score 1 for neutral and don't know and score 2 for disagree. Then, the total score for the scalar question 'availability of the police' is obtained by summing the individual scores of the statements.

For other types of questions we found no systematic effects. However, we have to note that the sample size of the pilot is relatively small in comparison to the sample size of the reference survey. The scalar questions include questions on traffic inconveniences, degradation of the neighbourhood and the availability of police. The pilot sample showed an evident increase of the percentage of respondents that chose a neutral answer. This shift is primarily caused by the web respondents. In some questions the shift occurs at the expense of both the positive and negative answer categories; in other questions only at the expense of the positive answer category. A more specific problem in the SM-1 pilot was the 'don't know' answer category, which in CAPI or CATI questionnaires is mostly implicitly deduced by the interviewer. The interviewer does not present this answer category to the respondent, but fills in this answer in case the respondent lingers to give an answer. This approach cannot easily be translated to self-administered questionnaires. Although we found no significant increase of 'don't know' answers, more research is necessary.

On the basis of the SM-1 pilot it also is hard to find an unambiguous explanation for the observed response bias. It is possible that through web interviewing a specific divergent group of respondents is interviewed that normally would not participate in CAPI or CATI surveys. For some questions the difference in answers may be explained by the so-called recency effect, i.e. persons begin by processing the last option they heard and as a consequence, the latter response options are more often chosen, see De Leeuw (2005). For other questions the presence of an interviewer may lead to more positive answers and, thus, may form an explanation, see e.g. Christian et al. (2006). However, these explanations cannot be extrapolated to all questions where differences in answer patterns were found.

Hence, in both studies we observed differences that are at least partially attributable to measurement errors. In general it is difficult to disentangle mea-

surement errors and nonresponse bias. By changing the mode also the response behaviour is altered. In the IE pilot we cannot distinguish the two types of biases beyond the resolution of the auxiliary variables that are used in the weighting of survey estimates. For the SM-1 pilot we are able to investigate the overall effect of incorporating web as a data collection mode when we make the assumption that the response has approximately the same composition.

Table 9.5 gives an overview of the response rates in the four pilots for the total sample and for the sample that has a listed land-line telephone, different. The general picture is that response rates increase when the sample is restricted to sample elements with a listed land-line telephone. This result confirms earlier findings, e.g. Cobben and Bethlehem (2005) concluded that households without a telephone more often refuse to participate; even when a data collection mode other than CATI is used. Table 9.5 confirms the finding that interviewer-assisted strategies lead to higher response rates. In all cases the response rates to mixed mode strategies that only employ web and mail surveys, are low compared to strategies that also encompass CAPI and CATI. This implies that in the self-administered strategies the risk of nonresponse error is much higher. However, it does not mean that the nonresponse bias is indeed larger. To get insight into the impact on nonresponse bias we need to consider the composition of the response. For the analysis of the nonresponse bias, we restrict ourselves to the SM-1 pilot. We fitted a multivariate logistic regression model to the response indicators, see also Chapter 3. The auxiliary variables in this response model are age, ethnicity, type of household and degree of urbanization. These variables are available from administrative data for both respondents and nonrespondents. With respect to these variables, we can calculate the representativeness with the R-indicator, see also Chapter 5. Table 9.6 displays the results for the SM-1 pilot. The obtained response in the web survey alone is less representative with respect to the auxiliary variables. Also, the maximal absolute bias $\hat{B}_m$ is large due to the low response rate. When the response in the CATI and CAPI group is added, the representativeness increases and approaches the representativeness of the reference survey, even though the sample size of this survey is much larger. The reduction in costs that is realized by employing a mixed mode data collection with web as one of the modes has not caused a decrease in representativeness with respect to these auxiliary variables. ■

Table 9.5: *Response rates in the four pilots (total and restricted to listed land-line telephones)*

			Response rate	
Pilot	*Group*	*Mode strategy*	*Total*	*Telephone*
ICT	Reference	CATI	na	65%
	Pilot	I: web/mail	38%	41%
		II: CATI, web/mail	55%	69%
SM-1	Reference	CATI/CAPI	69%	72%
	Pilot	web	30%	30%
		web, CATI/CAPI	65%	65%
SM-2	Reference	CATI/CAPI	67%	71%
	Pilot	web	30%	32%
		web, CATI/CAPI	65%	67%
IE	Pilot	CAPI	61%	65%
		web/mail	37%	39%
		web/mail, CATI	53%	62%

Table 9.6: *Composition of the response to the regular Safety Monitor, the response to the pilot web survey only, and the total response to the pilot Safety Monitor*

Response	*Size*	*Response rate*	$\hat{R}$	$CI^{bt}_{0.05}$	$\hat{B}_m$
Reference	30,139	68.9%	81.4%	(80.3; 82.4)	6.8%
web	3,615	30.2%	77.8%	(75.1; 80.5)	18.3%
web, CAPI/CATI	3,615	64.7%	81.2%	(78.3; 84.0)	7.3%

9.3 Mixed mode designs

In this section, we describe a number of designs for the implementation of mixed mode surveys. In section 9.3.1 we discuss different views on how to design the questionnaire, i.e. the instrument design. Then we present two ways of implementing the actual data collection using multiple modes: concurrent mixed mode design and sequential mixed mode design (section 9.3.2). One additional way to perform a mixed mode survey is to let the respondent choose the mode. It seems polite to leave the respondents the choice, but in fact an additional opportunity to refuse participation is introduced (Dillman and Tarnai, 1991). We therefore do not regard this option here.

We restrict ourselves to mixed mode data collection, and do not regard the use of multiple modes in the contact phase (described by De Leeuw (2005) as mixed- or multi mode systems). The emphasis in this section lies on the way the data are collected. We thus regard mixed mode surveys from a data analysis point of view. Deciding on how to design a mixed mode survey is complicated. For a discussion see Biemer and Lyberg (2003).

9.3.1 Instrument design

Within mixed mode data collection, there are three perspectives on how to design mixed mode surveys. The first perspective is based on the work by Dillman (2000). It is commonly referred to as unimode design, but this term is also being used to indicate design for single mode surveys (Biemer and Lyberg, 2003). Therefore we will refer to it as a *uniform* mode design. The uniform mode design seeks to optimize the design for one mode, keeping the design the same across all modes. For example, the use of don't know, refusal or empty are the same across modes. The aim of uniform mode design is to provide the same stimulus in all modes.
The second perspective is initiated by De Leeuw (2005), and referred to as a *generalized* mode design. Generalized mode design aims at presenting the same stimulus in each mode, instead of the same question, in an attempt to keep the cognitive burden of the respective modes the same. This design may result in different questionnaires for different modes. An example of generalized mode design can be found in Pierzchala et al. (2004), where the CATI-question 'Are you {name}?' was found too conversational for a web survey, so for the web survey they changed it to 'Is your name {name}?'.
The third perspective is *mode-specific* design, which aims to optimize the survey implementation in each mode separately. For example, in a telephone survey the number of answer categories cannot be as large as in a face-to-face survey due to recency effects. To avoid this effect, questions with a large number of answer categories in face-to-face will be split into two nested questions in a telephone survey. Because each mode is optimally designed we refer to this perspective as the *tailored* mixed mode design.

9.3.2 Concurrent and sequential design

In a *concurrent* mixed mode design, the sample s is divided into k disjunct subsamples $s^{(1)}, s^{(2)}, \ldots, s^{(k)}$ and these are all approached in a different data collection mode but at the same time. An illustration is provided in example 9.3.1.

In a *sequential* mixed mode design, the entire sample is initially approached in the same mode. At time $t + 1$, sample elements that did not respond to the initial mode are followed-up in another mode. This process can be repeated a number of times. In this design, at every time t the data is collected in one particular mode. In sequential approaches, the nonrespondents to the preceding mode are followed-up in another mode. Therefore this situation is referred to as a sequential mixed mode design, see example 9.3.2. Typical of this process is that respondents in subsequent modes (i.e. other than the initial mode) are nonrespondents to earlier modes; and the final nonrespondents did not respond to any of the previous modes. It is also possible to mix these two designs, as illustrated in example 9.3.3.

Example 9.3.1 The basic-question approach
In Chapter 4 we described two techniques of re-approaching nonrespondents in the Dutch Labour Force Survey (LFS). One technique was the basic-question approach. This approach was implemented using CATI and a combination of a web and a mail survey. Persons with a listed telephone number were approached by CATI, whereas persons without a listed number received a letter with a paper questionnaire and a login to a secured website from Statistics Netherlands where they could also fill-in the questionnaire, see figure 9.3. Here the sample s is split up into two subsamples. We denote by $s^{(1)}$ the sample that is approached by CATI, and by $s^{(2)}$ the combination of paper and web questionnaires. Note that we knew in advance which persons would go to subsample $s^{(1)}$ or $s^{(2)}$ because the subdivision was made based on telephone ownership. ■

Example 9.3.2 Re-approaching nonrespondents to the Dutch LFS

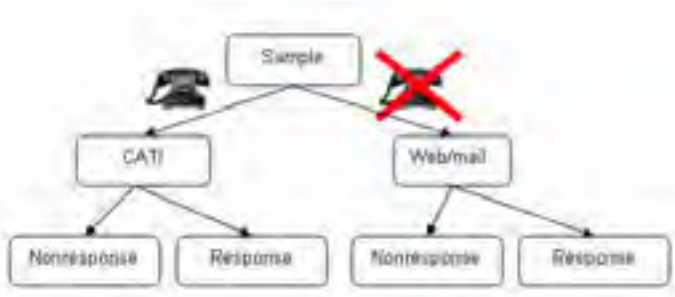

Figure 9.3: *Concurrent mixed mode design in the basic-question approach*

In Chapter 4 we described the re-approach of nonrespondents to the Dutch LFS with the basic-question approach. A sample of the original nonrespondents that had a listed, land-line telephone was re-approached by CATI, see figure 9.4. Here the sample s is first approached by CAPI at time t. The subsample of respondents to CAPI at time $t+1$ is denoted by $s^{(1)}$. The nonrespondents to CAPI are re-approached by CATI at time $t+1$. The fieldwork period ends at time $t+2$. Thus at $t+2$ we have a subsample denoted by $s^{(2)}$ of sample elements that were originally nonrespondents to CAPI but that did respond in the re-approach with CATI. The last subsample denoted by $s^{(3)}$ consists of sample elements that responded neither in CAPI, nor in CATI, i.e. the final nonrespondents. ■

Example 9.3.3 The mixed mode design in the pilot Safety Monitor
The pilots on the Safety Monitor (see example 9.2.1 and 9.2.3, and Chapter 5, section 5.5.3) follow a combination of a concurrent and a sequential mixed mode design. See figure 9.5. First, the entire sample is approached by a web survey. Next, nonrespondents to the web survey are re-approached by a combination of CAPI and CATI, based on telephone ownership. A small group of nonrespondents to the web survey was not eligible for a follow-up, they are the final web nonrespondents. We come back to this in section 9.6.2. ■

9.4 Combining data from different data collection modes

9.4.1 Introduction

For the remainder of this chapter, we impose some restrictions on the scope of our research. We restrict ourselves to data collection during the response phase. However, we do regard the contact phase when including data about the data collection process (or paradata, see section 9.4.2). Furthermore, we do not regard the mixed mode variant where the choice of data collection mode is left to the respondent.

In section 9.2.3 we discussed mode effects, in particular measurement errors, coverage errors and nonresponse bias. Research on measurement errors is aimed at minimising them. The occurrence of differential measurement errors across modes severely undermines the data quality in mixed mode surveys. Research on minimising measurement errors, i.e. developing questionnaires for mixed mode surveys, is crucial for the further development of mixed mode data collection.

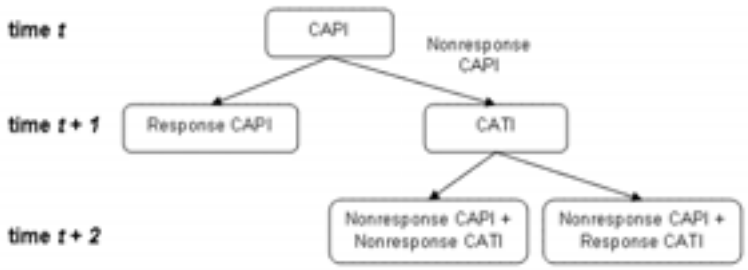

Figure 9.4: *Sequential mixed mode design in BQA re-approach of nonrespondents to the Dutch LFS*

Survey researchers are aware of this problem and therefore a lot recent work is devoted to measurement errors, for instance the work by Dillman et al. (1996), Dillman (2000), Dillman and Christian (2005), Roberts (2007) and Ariel et al. (2008). In our research we focus on adjustment for nonresponse bias. We, therefore, make the assumption that there are no measurement errors.

With regard to coverage errors, we assume that we know for all sample elements whether we can reach them by telephone, i.e. whether they have a listed land-line telephone. So, we only regard telephone surveys based on a sampling frame from listed land-line telephones. Another restriction on this research is imposed by assuming that web surveys are not used as a single mode, i.e. they are always combined with a data collection mode that does cover the entire population for example a mail survey. Thereby, we exclude coverage errors related to mixed mode surveys with a web survey. We make this assumption because we do not know whether sample elements have access to the internet, and when they have access, whether they are acquainted with the use of internet.

Having excluded measurement errors, and restricted the presence of coverage errors, our research is aimed at coping with differential nonresponse bias in mixed mode surveys. The question of interest, is how we can combine the data from different modes into one survey estimate adjusted for differential nonresponse bias.

In Chapter 8, we described a number of methods to adjust for nonresponse bias that incorporate different types of response. These methods account for the sequential nature of the response process and allow for mutual dependent response types, as well as a dependency between the different types of response

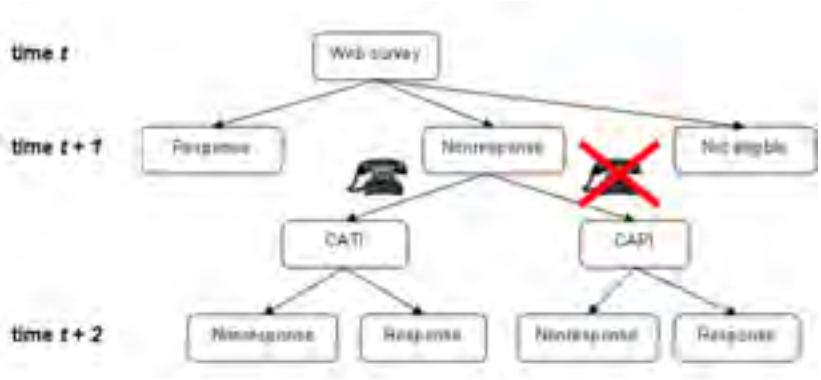

Figure 9.5: *The mixed mode design of SM-1 and SM-2 pilots*

and the survey items. As we have shown in Chapters 2 and 3, by distinguishing response types valuable information about the sample elements is revealed.

We recommended in Chapter 8 that the different response types (e.g. contact, participation) should be incorporated into adjustment methods for selective nonresponse. However, the amount of available information on the response process is distinct in the different data collection modes. In CAPI and CATI surveys, we observe all response types. But in web and mail surveys, without follow-ups we only observe whether sample elements return the completed questionnaires. Furthermore, recent research on nonresponse in surveys has brought to the attention so-called paradata (Couper, 1998). Paradata are auxiliary data about the process of data collection. We elaborate on the use of paradata in section 9.4.2. The available paradata is different across modes.

The question thus arises whether we should use all the available information in each mode separately, or unify the information used across modes. Put in the line of reasoning followed by Dillman (2000) and De Leeuw (2005), should we perform a uniform nonresponse bias adjustment, or a tailored nonresponse bias adjustment? This question reduces to the question whether we should model the response process in each mode separately, making optimal use of the available information. Or should we ignore the fact that the data are collected by distinct data collection modes, combine all the data into one dataset and proceed as normal thereby ignoring extra information about the data collection process. In section 9.5 we elaborate on this question by considering three approaches that differ in the way that they account for the mixed mode design; distinguish

between different response types over modes; and estimate the survey items. But first, we discuss the use of paradata.

9.4.2 Paradata

Couper (1998) first introduced the concept of paradata. Paradata are defined as auxiliary data about the process of data collection in both the contact and the response phase. Therefore, often the term process data is used to indicate the same. Paradata is a by-product from the data collection process, see also Couper and Lyberg (2005).

A number of levels on which paradata can be collected are identified by Kreuter (2006). She distinguishes between Contact Protocol Data (CPD) and data about the Interview Process (IP). The Contact Protocol Data is only available for interviewer-assisted surveys. This type of data can be divided into two types, the first of which is collected for every element in the sample that has been allocated to an interviewer (CPD I). This data comprehends information about each contact attempt, like day, time, mode and the outcome of each attempt. Also aggregate information about the contact process, like the number of contact attempts or the sequence of outcomes of the contact attempts fall into this category. The second type of Contact Protocol Data (CPD II) comprehends data that has only been collected for elements where contact is established. This is recorded information on the reaction of respondents, and other interviewer observations like type of dwelling, access impediments. Finally, the interviewer process provides data on the actual interview like item nonresponse and response time. Some interviewer process data are also available for web surveys (item nonresponse, response time) and mail surveys (item nonresponse).

These categories can be extended to include information about the interviewer (e.g. age, gender, years of experience), classification of response types (processed, non-contact, able, refusal, response). If interviewer data is used in the models to adjust for selective nonresponse, the clustering of sample elements within interviewers has to be accounted for. This can be done by refinement of the nonrespondents adjustment models with an additional level, as described by the multilevel model in Chapter 8.

The use of paradata is twofold. In a number of studies, these data have been used to analyse response behaviour. Durrant and Steele (2007) recommend to collect paradata for the guidance of contact strategies and interviews and for the identification of groups in the population that allow for tailoring. Heerwegh (2003) collects paradata in web surveys. These data comprise information about the number of changes made to answers, the sequence in which the questions are

answered and the time spent on each question. These paradata serve to describe response behaviour and could possibly be an indication of data quality. They can also serve as feedback to questionnaire designers, e.g. when respondents have returned to earlier questions in the questionnaire and changed their answers so that they could follow a different routing, this might be an indication that the questionnaire is not designed optimally. Another application of paradata is to inform the fieldwork. Henly and Bates (2005) use contact history information in a panel survey to inform interviewers about the previous response process. They provide them with information about the time, day and result of previous contact attempts, the number of attempts needed to obtain a response as well as remarks made by the respondents. They hypothesize that this information can be used by the interviewer to tailor their approach thus lowering the number of contact attempts needed to establish contact and at the same time increasing response. They, however, do not find convincing evidence for this hypothesis. This is possibly caused by seasonal effects, as well as interviewers not being accustomed to the use of contact history data. Groves and Heeringa (2005) describe how paradata can be used to inform fieldwork strategies by a so-called responsive design.

Beaumont (2005) investigates whether paradata can be used for the adjustment of nonresponse bias. Whereas most variables that are being used for nonresponse adjustment are non random, fixed demographic or socio-economic characteristics of the sample elements or the geographical location, the paradata are random because they can change if the data collection process were to be repeated for the same sample. Furthermore, paradata are only available on the sample level, and not for the entire population. With these restrictions in mind, Beaumont (2005) shows that paradata can be used for nonresponse adjustment without introducing additional bias or an additional variance component in the estimates for population totals.

It may seem far-fetched to include information that at first sight has so little to do with the survey items. However, we feel that this information aids to grasp what happens during the fieldwork. Furthermore, the response behaviour can only partly be explained by socio-economic and demographic variables. In general, logistic models for the response propensity have an explained variance of at most 20%. That implies that 80% of the variation cannot be explained by the standard information[2].

Paradata display the behaviour of respondents in the data collection strategy.

[2]This effect is partly caused by the fact that we are trying to estimate response behaviour based on one observation only.

Hence, there is a causal relationship between response behaviour and paradata. The strategy influences response behaviour and paradata makes the behaviour visible, but not the other way around. The variables that we have considered in our research thus far, are all fixed attributes of the sample persons. Paradata are not fixed, but random. In Chapter 8 we gave an example of the relationship between the number of contact attempts and the employment situation. If such eminent relationships exists between survey items and paradata, we should include them in the adjustment models.

9.5 Nonresponse adjustment methods in mixed mode surveys

In Chapter 6, we described a number of traditional adjustment methods for nonresponse bias. Methods to adjust for nonresponse including different response types are discussed in Chapter 8. In this chapter, we focus on nonresponse adjustment that accounts for the data collection design. First, we discuss three approaches to adjust for nonresponse in the concurrent mixed mode design in section 9.5.1, as in this situation the adjustment methods prove to be quite straightforward. Next, we discuss the same three approaches in a sequential design in section 9.5.2, where the situation is more complicated because of the sequence of data collection modes that may occur.

9.5.1 Concurrent mixed mode and nonresponse adjustment

Let us regard a concurrent mixed mode setting as in figure 9.6. We describe three approaches to adjust for nonresponse in this setting. The approaches vary in the way they incorporate the mixed mode design, different types of auxiliary information and whether they include variables in different modes or in all modes. Consequently, they differ considerably in their complexity, ease of implementation and computation time. We describe these approaches in terms of the resulting respondents' weights w_i in the calibration framework.

9.5.1.1 First approach - Linear weighting

The first approach is straightforward linear weighting. The data from the different modes are combined into one dataset. For calculating the weights w_i the respondents are compared to the nonrespondents and we do not distinguish

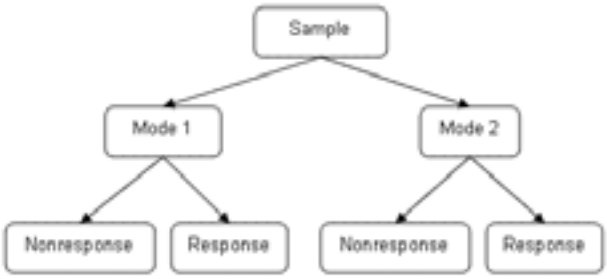

Figure 9.6: *A concurrent mixed mode design with two modes*

between modes, see figure 9.7. Information on the sample level or on the population level can be used and the resulting estimator is the general calibration estimator described in Chapter 6, i.e.

$$w_i = d_i(1 + (\overline{\mathbf{X}} - \overline{\mathbf{x}}_{ht}^r)'(\sum_r d_i \mathbf{X}_i \mathbf{X}_i')^{-1}\mathbf{X}_i) \tag{9.1}$$

as in (6.27). For the calculation of w_i, socio-economic and demographic variables are used. However, we do not yet include information on the data collection mode or -process.

This method, however, disregards the fact that the data are collected in multiple modes. All respondents are combined, and compared to the combined nonrespondents. Based on this comparison, the weights w_i are constructed. If the relationship between the auxiliary variables and the response behaviour in one mode is very different from this relation in another mode, this difference will not be reflected in the weights w_i. This relationship can, for example, be affected

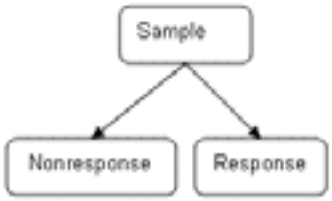

Figure 9.7: *Approach 1 - straightforward linear weighting*

by coverage errors. For instance, in a telephone survey and a face-to-face survey the relationship between ethnic group and response is different due to a lower coverage of listed land-line telephones among non-natives (see Chapter 7).

9.5.1.2 Second approach - Linear weighting per mode

Approach 1 does not allow for differences in response behaviour across modes. Therefore, in approach 2 we expand the linear weighting model by including the mode in the weighting model. This can be done by constructing a variable *mode* $= M$ that indicates the mode in which a sample element is approached. In the mixed mode setting displayed in figure 9.6, the variable M has two categories, i.e.

$$M_i = \begin{cases} 1, & \text{if person i is approached in mode 1} \\ 0, & \text{if person i is approached in mode 2} \end{cases} \tag{9.2}$$

The inclusion of the variable M automatically restricts the auxiliary information to the sample, $\mathbf{X}$, since M only has a meaning for the persons in the sample[3]. The expression for the calibration weights w_i is the same as for the first approach, see (9.1). The only difference is the use of the auxiliary variable vector $\mathbf{X}$ which is extended to $(\mathbf{X}, M)'$.

If the variable mode is added to the weighting model as a main effect, the model allows for an overall difference with respect to the mode. However, different relations between auxiliary variables and the response behaviour within modes will only be reflected in the weights if the auxiliary variables $\mathbf{X}$ include interactions between the mode and the socio-economic or demographic information for which a different relationship exists. If there are indeed different relationships between the auxiliary variables and the response behaviour across modes, there are two options to extend the linear weighting approach. The first option, approach 2.1, is to cross the variable mode with the variables for which a difference in response behaviour is expected or observed. The disadvantage of this option is that the weighting model can become very large, with the risk of empty cells and large variances. The second option, approach 2.2, is to split the sample into subgroups that are approached in different modes, i.e. samples $s^{(1)}, \ldots s^{(k)}$, and to model the response behaviour in each mode separately, see figure 9.8. Approach 2.2. constructs a weighting model for each mode separately, thus allowing the models to include different auxiliary variables that explain response behaviour. This implies that coefficients for the same auxiliary

[3]It is possible to calibrate to the population level after the variables are calibrated to the sample. The final calibration weights are then constructed in two steps.

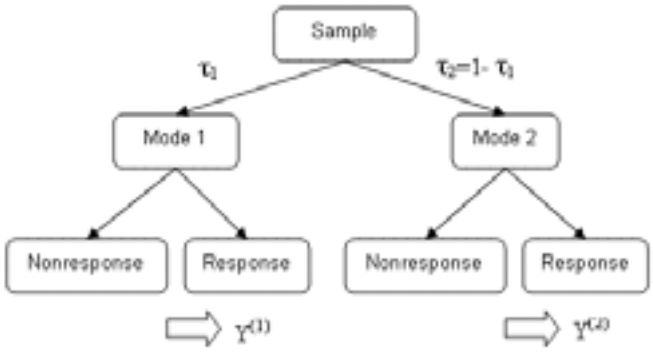

Figure 9.8: *Approach 2.2 - nonresponse adjustment in each mode separately*

variable can be different for distinctive modes (if included in the corresponding weighting model). Approach 2.2 then results in two estimates for the survey item, one based on the response behaviour in mode 1, and one based on the response behaviour in mode 2. In addition, we introduce the probability τ_1 for the assignment to mode 1, and consequently $\tau_2 = 1 - \tau_1$ is the probability that a sample element is assigned to mode 2. If for instance mode 1 is CATI, then τ_1 would be the probability that sample elements have a listed land-line telephone.

This situation is analogous to stratified sampling, in which the modes determine the strata. In all strata, a separate weighting adjustment is performed. The estimates in the strata can be combined into one single estimate for the entire survey using the mode allocation probabilities.

If, however, in every separate response model the same variables are included, approach 2.2 is not different from approach 2.1. Furthermore, in that situation approach 2.1 is superior to approach 2.2 because in approach 2.1 all the modes are estimated simultaneously. Approach 2.2 uses a smaller number of observations for each response model in a different mode. If the more detailed stratification was not necessary, this may result in a higher variance for the estimate thus obtained, see Little and Vartivarian (2005).

Example 9.3.2.1 The basic-question approach - continued
In the control group for the basic-question approach as displayed in figure 9.3, we have a number of auxiliary variables at our disposal, denoted by $\mathbf{X}$. These variables are linked from external sources, e.g. the population register or the tax

office, see Chapters 3 and 4. This information is available for every person in the sample, both respondents and nonrespondents. Additionally, it is also available for the entire target population.

Similar to approach 1, we use $\mathbf{X}$ in the general calibration estimator. However, for approach 2.1, we include an extra auxiliary variable for the mode in which a person was approached. The information is than confined to the sample so that we obtain $(\mathbf{X}, M)'$. According to approach 2.2, the sample is divided into subsamples $s^{(1)}$ and $s^{(2)}$ as in example 9.3.2.1. Next, the response behaviour in $s^{(1)}$ and $s^{(2)}$ is modelled separately. In this approach the information is not confined to the sample level anymore and we could also use information on the population level. However, in this example the information is available on both the sample- and the population level. Let us denote the auxiliary variables that explain the response behaviour in the CATI sample $s^{(1)}$ by $\mathbf{X}^{(1)}$, and in the sample for the paper/web questionnaires, $s^{(2)}$, by $\mathbf{X}^{(2)}$. The sample average for survey item Y, is then estimated in both samples separately.

For sample $s^{(k)}$ we obtain an estimate $\overline{y}_w^{(k)}$ $(k = 1, 2)$ by solving

$$\begin{aligned} \overline{y}_w^{(k)} &= \frac{1}{N} \sum_{i \in s^{(k)}} w_i y_i \\ \text{subject to} \qquad & \frac{1}{N} \sum_{i \in s^{(k)}} w_i \mathbf{X}_i^{(k)} = \overline{\mathbf{X}}^{(k)} \end{aligned}$$

Subsequently, the final estimate $\overline{y}_w$ is obtained by mixing the estimates in the two subsamples according to the mix in which the total sample s has been divided over the two subsamples. In our example, 67% of the sampled persons have been approached by CATI and the other 33% by paper/web questionnaires. $\overline{y}_w$ would thus be calculated by $\overline{y}_w = 0.67\overline{y}_w^{(1)} + 0.33\overline{y}_w^{(2)}$. ■

9.5.1.3 Third approach - Sample selection model using paradata

The first two approaches do not use paradata. Furthermore, these approaches calculate weights for the respondents based on the calibration estimator. The respondents receive the same weights, for every survey item Y. In the third approach, we do include paradata. Besides that, the adjustment is made for specific survey items. The relationship between the response behaviour and the survey item Y is modelled with a multivariate probit model for each data collection mode. Therefore, the respondents receive a different weight for every survey item Y.

The most important and informative paradata for nonresponse adjustment, is the type of response. In Chapters 2, 3 and 8 we elaborated on the distinct response types. The third approach explicitly models the response process. By distinguishing between the response types, we can introduce auxiliary variables even more precisely there where they have an influence.

Example 9.5.2 Paradata in the basic-question approach
For the CATI sample $s^{(1)}$ we have information about the number of call attempts and the outcome of each of these attempts. Furthermore, we know which interviewer made the call and we also have information about the age, gender and experience of the interviewers. However, for the paper- and web questionnaires we only observe whether a person returned the paper questionnaire, or filled-in the web questionnaire. For the web questionnaire we can see how much time a respondent needed to finish the questionnaire. The reliability of this data is, however, questionable because respondents may not have spent all the recorded time on completing the survey. For the paper questionnaire we know the elapsed time between the moment the questionnaire was mailed, and the moment the respondent returned it. We do not have this information for the nonrespondents, for obvious reasons. This sort of information, consequently, cannot be used for nonresponse adjustment but may be very helpful in explaining and adjusting for measurement errors. ■

The approach comes down to modelling the response process in each mode separately, like in approach 2 where the linear weighting model was constructed for every mode separately. The distinction between approach 2 and approach 3 is that approach 3 is not based on calibration but models the different response probabilities and survey items using multivariate probit models for each mode. We already defined the response probabilities in Chapter 1. We now define the probabilities for each of the modes in the design. We illustrate approach 3 with the two most common response types, contact and participation. Generalisation to more response types is straightforward, but cumbersome in notation. Furthermore, we assume that the survey item Y is a continuous variable.

Denote by γ_k^* the latent contact probability for mode k, with corresponding indicator $C_k = 1$ if $\gamma_k^* > 0$ and $C_k = 0$ otherwise. Likewise, the latent participation probability for mode k is denoted ϱ_k^* with corresponding indicator $P_k = 1$ if $\varrho_k^* > 0$ and $P_k = 0$ otherwise. In the concurrent mixed mode design, the sample s is divided into k disjunct, exhaustive subsamples $s^{(1)}, s^{(2)}, \ldots, s^{(k)}$ and these subsamples are all approached in a different mode. We illustrate approach three for $k = 2$. For sample elements i with $C_{ki} = 1$ and $P_{ki} = 1$ we observe the survey

item Y_i. For the sample elements for which $C_{ki} = 0$, or $C_{ki} = 1$ and $P_{ki} = 0$, we do not observe Y_i. Again let us introduce a latent survey item Y_i^*, and the latent variables γ_{ki}^* and ϱ_{ki}^* as in

$$\begin{array}{rl} \gamma_{ki}^* = & \mathbf{X}_{ki}^C \boldsymbol{\beta}^C + \epsilon_{ki}^C, \\ \varrho_{ki}^* = & \mathbf{X}_{ki}^P \boldsymbol{\beta}^P + \epsilon_{ki}^P, \\ Y_i^* = & \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \epsilon_i^Y \end{array} \tag{9.3}$$

for $i = 1, \ldots, n$. Furthermore, Y_i can be described as

$$Y_i = \begin{cases} Y_i^*, & \text{if } (P_{ki} = 1; C_{ki} = 1) \\ \text{missing}, & \text{if } (C_{ki} = 0) \text{ or } (C_{ki} = 1; P_{ki} = 0) \end{cases} \tag{9.4}$$

for $i = 1, \ldots, n$ and $k = 1, 2$. The method is similar to the sample selection model described in Chapter 8, with the distinction that now the contact- and the participation process are defined for multiple modes. In addition, we assume that the allocation of sample elements to mode 1 or 2 is known. With probability τ_1, sample elements are approached in mode 1, and with probability $\tau_2 = 1 - \tau_1$ sample elements proceed to mode 2. This approach is graphically represented in figure 9.9.

We illustrate how the model in figure 9.9 can be estimated by constructing the likelihood. First, observe that every sample element contributes to a specific part of the likelihood. The first division is based on the allocation probability τ. With probability τ_1, $M = 1$ and sample elements proceed to mode 1. With probability $\tau_2 = 1 - \tau_1$, $M = 0$ and the sample elements proceed to mode 2. The total sample is now divided into two subsamples, $s^{(1)}$ for mode 1, and $s^{(2)}$ for mode 2. Without loss of generality let us assume that the first $i = 1, \ldots, n^1$ elements are approached in mode 1, and the sample elements $i = n^1 + 1, \ldots, n$ are approached in mode 2. Then, in each mode k with $k = 1, 2$ the sample elements either are not contacted ($C_{ki} = 0$), or they are contacted but do not participate ($C_{ki} = 1; P_{ki} = 0$), or they are both contacted and participate ($C_{ki} = 1; P_{ki} = 1$). In the latter case we observe the survey item, otherwise we do not observe Y_i^*.

The likelihood can be divided into six parts. See table 9.7. Observations for the survey item come from two different modes. The process in the two modes can be different, and thus the relationship between the response behaviour and the survey item may be different as well. To estimate this model, we need to make an assumption about the distributions of the error terms in equations

Table 9.7: *Contributions to the likelihood for different groups in concurrent mixed mode design with two modes*

Group	*Contribution*
Noncontact mode 1	$M \times \tau_1 P(C_1 = 0 \vert \mathbf{X}^{C_1})$
Refusal mode 1	$M \times \tau_1 P(C_1 = 1; P_1 = 0 \vert \mathbf{X}^{C_1}, \mathbf{X}^{P_1})$
Participation mode 1	$M \times \tau_1 P(Y = y; C_1 = 1; P_1 = 1 \vert \mathbf{X}^{C_1}, \mathbf{X}^{P_1}, \mathbf{X}^{Y})$
Noncontact mode 2	$(1 - M) \times \tau_2 P(C_2 = 0 \vert \mathbf{X}^{C_2})$
Refusal mode 2	$(1 - M) \times \tau_2 P(C_2 = 1; P_2 = 0 \vert \mathbf{X}^{C_2}, \mathbf{X}^{P_2})$
Participation mode 2	$(1 - M) \times \tau_2 P(Y = y; C_2 = 1; P_2 = 1 \vert \mathbf{X}^{C_2}, \mathbf{X}^{P_2}, \mathbf{X}^{Y})$

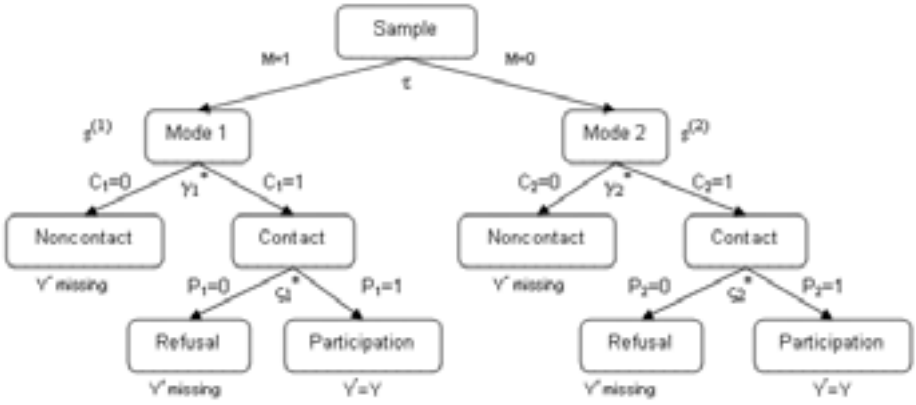

Figure 9.9: *Approach 3 - simultaneous estimation of the response process within modes*

(9.3). There are five error terms: $\epsilon_{1i}^{C}, \epsilon_{2i}^{C}, \epsilon_{1i}^{P}, \epsilon_{2i}^{P}$ and ϵ_{i}^{Y}. Like in Chapter 8, we assume the vector of error terms follows a multivariate normal distribution with expectation zero and covariance matrix $\mathbf{\Sigma}$.

$$\begin{pmatrix} \epsilon_{1i}^{C_1} \\ \epsilon_{2i}^{C_2} \\ \epsilon_{1i}^{P_1} \\ \epsilon_{2i}^{P_2} \\ \epsilon_{i}^{Y} \end{pmatrix} \backsim N\left(\mathbf{0}, \mathbf{\Sigma} = \begin{bmatrix} 1 & 0 & \zeta_{CP_1}\sigma & 0 & \zeta_{CY_1}\sigma \\ 0 & 1 & 0 & \zeta_{CP_2}\sigma & \zeta_{CY_2}\sigma \\ \zeta_{PC_1}\sigma & 0 & 1 & 0 & \zeta_{PY_1}\sigma \\ 0 & \zeta_{PC_2}\sigma & 0 & 1 & \zeta_{PY_2}\sigma \\ \zeta_{YC_1}\sigma & \zeta_{YC_2}\sigma & \zeta_{YP_1}\sigma & \zeta_{YP_2}\sigma & \sigma^2 \end{bmatrix}\right) \quad (9.5)$$

where $\zeta_{CP_k} = \zeta_{PC_k}$ for $k = 1, 2$ is the correlation between contact and participation in mode k. $\zeta_{YC_k} = \zeta_{CY_k}$ different $\zeta_{YP_k} = \zeta_{PY_k}$ is the correlation between the survey item and contact different participation. And σ^2 is the variance of the survey item. The zero correlations follow from the fact that mode 1 and mode 2 are assumed to be independent. This is not exactly true, there is a relation between contact and participation in different modes, however we cannot identify the relationship based on one observation for each sample element.

The risk of non-identification of the model decreases because now there can be mode-specific instrumental variables for which it is clear beforehand that they do not play a role in other equations. This means that the exclusion restriction

is more easily satisfied, see Chapter 8. Because of the division of the sample into subsamples that are approached in different modes, the error term specification is slightly different from the error term specification in the models in Chapter 8.

Example 9.5.3 Approach three in the BQA
For the basic-question approach, there were two modes: CATI and a combination of a web and a mail survey. For the CATI approach, we observed whether contact was made, and for the contacted sample elements, whether they participated. However, for the combination of web and mail, we only observed whether the questionnaire was returned. Graphically, this process can be displayed as in figure 9.10. Instead of contact and participation processes in both modes, we now only have contact- and participation probabilities in the CATI mode, and a response probability for the web/mail survey. So, the model for this situation can be described as

$$\begin{array}{rl} \gamma_i^* = & \mathbf{X}_i^C \boldsymbol{\beta}^C + \epsilon_i^C, \\ \varrho_i^* = & \mathbf{X}_i^P \boldsymbol{\beta}^P + \epsilon_i^P, \\ \rho_i^* = & \mathbf{X}_i^R \boldsymbol{\beta}^R + \epsilon_i^R, \\ Y_i^* = & \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \epsilon_i^Y \end{array} \tag{9.6}$$

for $i = 1, \ldots, n$. Again, we observe Y_i that can be described as

$$Y_i = \begin{cases} Y_i^*, & \text{if } (P_i = 1; C_i = 1) \text{ or } (R_i = 1) \\ \text{missing}, & \text{if } (C_i = 0) \text{ or } (C_i = 1; P_i) = 0 \text{ or } (R_i = 0) \end{cases} \tag{9.7}$$

for $i = 1, \ldots, n$. Sample elements with a listed land-line telephone were allocated to CATI. So, τ_1 is the probability that sample element i has a listed land-line telephone and τ_2 is the probability that sample elements cannot be reached by telephone. The likelihood now consists of five parts, see table 9.8. Likewise, the error term distribution becomes

$$\begin{pmatrix} \epsilon_i^C \\ \epsilon_i^P \\ \epsilon_i^R \\ \epsilon_i^Y \end{pmatrix} \backsim N\left(\mathbf{0}, \begin{bmatrix} 1 & \zeta_{CP}\sigma & 0 & \zeta_{CY}\sigma \\ \zeta_{PC}\sigma & 1 & 0 & \zeta_{PY}\sigma \\ 0 & 0 & 1 & \zeta_{RY}\sigma \\ \zeta_{YC}\sigma & \zeta_{YP}\sigma & \zeta_{YR}\sigma & \sigma^2 \end{bmatrix} \right) \tag{9.8}$$

the correlation between the response types in CATI and the response behaviour in the web/mail survey is zero. The other correlations are defined as described in approach 3, with $\zeta_{YR} = \zeta_{RY}$ the correlation between response behaviour in the web/mail survey and the survey item. ■

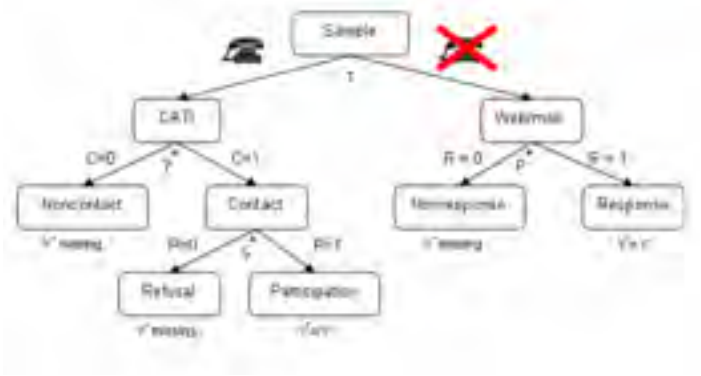

Figure 9.10: *The concurrent mixed mode design in the BQA*

Table 9.8: *Contributions to the likelihood for different groups in the basic-question approach*

Group	*Contribution*
Noncontact CATI	$M \times \tau_1 P(C = 0 \mid \mathbf{X}^C)$
Refusal CATI	$M \times \tau_1 P(C = 1; P = 0 \mid \mathbf{X}^C, \mathbf{X}^P)$
Participation CATI	$M \times \tau_1 P(Y = y; C = 1; P = 1 \mid \mathbf{X}^C, \mathbf{X}^P, \mathbf{X}^Y)$
Nonresponse web/mail	$(1 - M) \times \tau_2 P(R = 0 \mid \mathbf{X}^R)$
Response web/mail	$(1 - M) \times \tau_2 P(Y = y; R = 1 \mid \mathbf{X}^R, \mathbf{X}^Y)$

9.5.2 Sequential mixed mode and nonresponse adjustment

So far, we have focussed on the concurrent mixed mode design with two modes. Now we translate approaches 1 to 3 to a sequential mixed mode design. We regard the sequential mixed mode design with two modes, see figure 9.11. More complex mixed mode designs can be treated in a similar way.

In the first approach the data from the different modes is combined into one dataset. For the calculation of the weights w_i approach 1 comes down to comparing the respondents in mode 1 and mode 2 with the final nonrespondents in mode 2. However, respondents to mode 2 first did not respond to mode 1. In the first approach, the sequential nature of the response process is hence not accounted for.

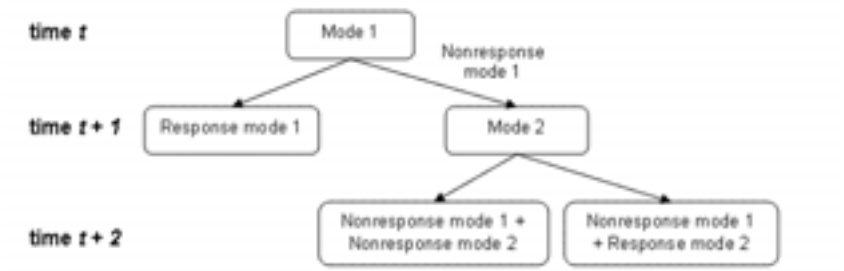

Figure 9.11: *A sequential mixed mode design with two modes*

In approach 2 for the concurrent mixed mode design, the variable $M =$ **mode** is included in the weighting model. For the situation with two concurrent modes, M has two categories; one for each mode. In the sequential design we are not only interested in the mode, but also in the sequence of modes. This sequence can be regarded as the mode response path, so to say the history of sample elements before a response or a final nonresponse is obtained. Let us introduce the variable $M^r =$ **mode response path**, which indicates the sequence of modes a sample element has followed before the final result is obtained. For instance, in figure 9.11 the variable M^r consists of three categories:

$$M_i^r = \begin{cases} 1, & \text{i responds in mode 1} \\ 2, & \text{i does not respond in mode 1 and responds in mode 2} \\ 3, & \text{i does not respond in mode 1 nor in mode 2} \end{cases} \tag{9.9}$$

for $i = 1, \ldots, n$.

The variable M^r as defined in (9.9) cannot be used in a linear weighting model, since $M^r = 1$ does not have a nonresponse counterpart. The adjustment method can only be applied to nonrespondents in mode 1, that either responded or did not respond to mode 2, i.e. only the categories $M^r = 2$ and $M^r = 3$ take part in the weighting adjustment. Hence respondents to mode 2 receive an adjustment weight, whereas respondents to mode 1 would all receive weights equal to 1. Approach 2 is not interesting for a sequential mixed mode design, unless not all nonrespondents to mode 1 are followed-up in another mode. An example of this is provided in section 9.6.

Therefore, we proceed to approach 3. This approach distinguishes between response types and models the relationship with the survey item as well. The response behaviour, the survey item and their relations are modelled with a

multivariate probit model. In the sequential mixed mode design the allocation of sample elements to follow-up modes depends on the outcome of the response process in the preceding modes(s). Let us again consider the situation where we distinguish the two main response types, i.e. contact and participation. Now, both non-contacts and refusals are re-allocated to a follow-up mode. See figure 9.12. It is also possible that only one of the response types is followed-up, for example only non-contacts. This situation can be treated in a similar way. In this situation, we have four possible outcomes for the sample elements: response in mode 1, non-contact in mode 2, refusal in mode 2 and response in mode 2. However, sample elements can proceed to the second mode following two paths: either they refuse participation in mode 1, or they are not contacted in mode 1. Therefore, there are seven possible mode response paths sample elements can follow in this example. The situation can be described as follows

$$\begin{array}{rcl} \gamma_{ki}^{*} = & \mathbf{X}_{ki}^{C}\boldsymbol{\beta}^{C} + \epsilon_{ki}^{C}, \\ \varrho_{ki}^{*} = & \mathbf{X}_{ki}^{P}\boldsymbol{\beta}^{P} + \epsilon_{ki}^{P}, \\ Y_{i}^{*} = & \mathbf{X}_{i}^{Y}\boldsymbol{\beta}^{Y} + \epsilon_{i}^{Y} \end{array} \tag{9.10}$$

for $i = 1, \ldots, n$ and $k = 1, 2$. Furthermore, we observe Y_i

$$Y_i = \begin{cases} Y_i^*, & \text{if } (C_{1i} = 1; P_{1i} = 1) \text{ or} \\ & (C_{1i} = 0; C_{2i} = 1; P_{2i} = 1) \text{ or} \\ & (C_{1i} = 1; P_{1i} = 0; C_{2i} = 1; P_{2i} = 1) \\ \text{missing}, & \text{if } (C_{1i} = 0; C_{2i} = 0) \text{ or} \\ & (C_{1i} = 1; P_{1i} = 0; C_{2i} = 0 \text{ or} \\ & (C_{1i} = 0; C_{2i} = 1; P_{2i} = 0) \text{ or} \\ & (C_{1i} = 1; P_{1i} = 0; C_{2i} = 1; P_{2i} = 0) \end{cases} \tag{9.11}$$

for $i = 1, \ldots, n$. The vector of error terms is assumed to follow a multivariate normal distribution, i.e.

$$\begin{pmatrix} \epsilon_i^C \\ \epsilon_i^P \\ \epsilon_i^Y \end{pmatrix} \backsim N\left(\mathbf{0}, \begin{bmatrix} 1 & \zeta_{CP}\sigma & \zeta_{CY}\sigma \\ \zeta_{PC}\sigma & 1 & \zeta_{PY}\sigma \\ \zeta_{YC}\sigma & \zeta_{YP}\sigma & \sigma^2 \end{bmatrix}\right) \tag{9.12}$$

Hence, the parameters that have to be estimated are $\boldsymbol{\beta}^C$, $\boldsymbol{\beta}^P$, $\boldsymbol{\beta}^Y$, $\boldsymbol{\epsilon}^C$, $\boldsymbol{\epsilon}^P$, $\boldsymbol{\epsilon}^Y$, ζ_{CP}, ζ_{CY}, ζ_{PY}, σ. This can be done by MLE. The likelihood can be split into seven parts as displayed in table 9.9. Already for two modes and two response types the model becomes quite complex.

Table 9.9: *Contributions to the likelihood for different groups in the sequential mixed mode design with two modes and two response types*

Group	*Contribution*
Participation mode 1	$P(Y = y; C_1 = 1; P_1 = 1 \vert \mathbf{X}^{C_1}, \mathbf{X}^{P_1}, \mathbf{X}^{Y})$
Noncontact mode 2, preceded by	
Noncontact mode 1	$P(C_1 = 0; C_2 = 0 \vert \mathbf{X}^{C_1}, \mathbf{X}^{C_2})$
Refusal mode 1	$P(C_1 = 1; P_1 = 0; C_2 = 0 \vert \mathbf{X}^{C_1}, \mathbf{X}^{P_1}, \mathbf{X}^{C_2})$
Refusal mode 2, preceded by	
Noncontact mode 1	$P(C_1 = 0; C_2 = 1; P_2 = 0 \vert \mathbf{X}^{C_1}, \mathbf{X}^{C_2}, \mathbf{X}^{P_2})$
Refusal mode 1	$P(C_1 = 1; P_1 = 0; C_2 = 1; P_2 = 0 \vert \mathbf{X}^{C_1}, \mathbf{X}^{P_1}, \mathbf{X}^{C_2}, \mathbf{X}^{P_2})$
Participation mode 2, preceded by	
Noncontact mode 1	$P(Y = y; C_1 = 0; C_2 = 1; P_2 = 1 \vert \mathbf{X}^{C_1}, \mathbf{X}^{C_2}, \mathbf{X}^{P_2}, \mathbf{X}^{Y})$
Refusal mode 1	$P(Y = y; C_1 = 1; P_1 = 0; C_2 = 1; P_2 = 1 \vert \mathbf{X}^{C_1}, \mathbf{X}^{P_1}, \mathbf{X}^{C_2}, \mathbf{X}^{P_2}, \mathbf{X}^{Y})$

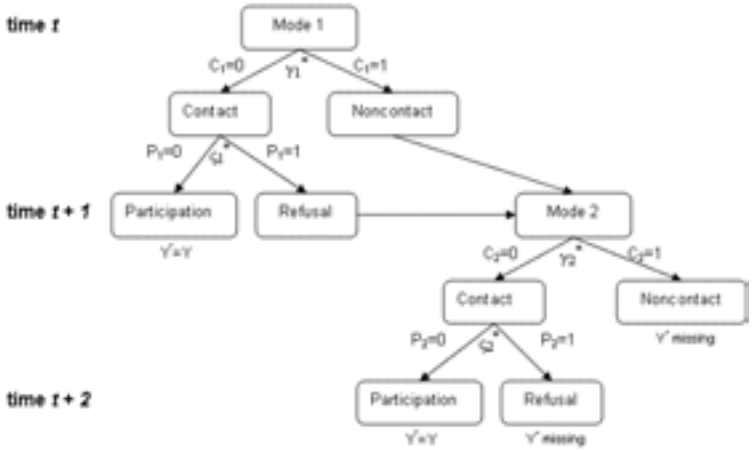

Figure 9.12: *Approach 3 in a sequential mixed mode design with two modes*

Based on these expressions for the different parts of the likelihood the parameters can be estimated. In Chapter 8 we described a number of methods that can be employed for this purpose. With the estimated parameters, an estimator for the survey item can be constructed. In Chapter 8 we showed how the Horvitz-Thompson estimator can be modified for the sample selection model, i.e.

$$\overline{y}_{ht}^{sel} = \frac{1}{N}\sum_{i=1}^{n} d_i \hat{E}\Big[Y_i | R_i = 1, \mathbf{X}_i^R, \mathbf{X}_i^Y\Big] \tag{9.13}$$

where $d_i = 1/\pi_i$. This expression holds for the sample selection model that distinguishes response and nonresponse. For the sequential mixed mode design with two response types, the expression would become

$$\overline{y}_{ht}^{smm} = \frac{1}{N}\sum_{i=1}^{n} d_i \hat{E}\Big[Y_i | R_i = 1, \mathbf{X}_i^{C_1}, \mathbf{X}_i^{C_2}, \mathbf{X}_i^{P_1}, \mathbf{X}_i^{P_2}, \mathbf{X}_i^Y\Big] \tag{9.14}$$

where *smm* stands for *sequential mixed mode*, and $R_i = 1$ if $(C_{1i} = 1; P_{1i} = 1)$ or $(C_{1i} = 0; C_{2i} = 1; P_{2i} = 1)$ or $(C_{1i} = 1; P_{1i} = 0; C_{2i} = 1; P_{2i} = 1)$.

9.6 Application to the pilot Safety Monitor-1

In this section, we apply approaches 1, 2.1, 2.2 and a restricted version of approach 3 to the pilot Safety Monitor-1 (SM-1) that we described in examples 9.2.1, 9.2.3 and 9.3.3. First, we present the pilot SM-1 in detail in section 9.6.1, with an emphasis on the survey items that we used for the evaluation of the approaches. Next, we apply approaches 1, 2.1, 2.2 and 3 to the pilot SM-1 in section 9.6.2. In section 9.6.3 we summarise the models, and present and discuss the results from the approaches.

9.6.1 Survey items in the Safety Monitor

Van den Brakel et al. (2007) analysed the survey items in the different data collection modes in the pilot SM-1. In the analysis, Van den Brakel et al. (2007) could not distinguish between nonresponse bias, coverage- and measurement errors in the different modes because the pilot was not designed for that purpose. Van den Brakel et al. (2007) estimated the values for the survey items, using the generalized regression estimator (see Chapter 5). For the regular Safety Monitor, the weighting model is

$$\begin{array}{l} Age_{11} \times Gender_2 + Age_{11} \times PoliceRegion_{25} + \\ + MaritalStatus_4 + DegreeUrbanization_5 \times Region_{16} + \\ + Telephone_2 \times Age_6 + DegreeUrbanization_5 \times Income_6 + \\ + Region_{16} \times Ethnicity_3 + HouseholdSize_5 \end{array} \tag{9.15}$$

The subscripts denote the number of categories. '$\times$' implies that the variables are crossed, whereas '+' indicates that the variables are additive. The sample for the mixed mode pilot is smaller than for the regular Safety Monitor. Therefore, a reduced weighting model is applied to the data for the pilot SM-1. This model is

$$\begin{array}{l} MaritalStatus_4 + DegreeUrbanization_5 + Region_{13} + \\ + Householdsize_5 + Age_{11} \times Gender_2 + Telephone_2 \times Age_6 \end{array} \tag{9.16}$$

One set of items in the analysis concerned scale variables based on a set of attitudinal items in the questionnaire. These items are: the opinion on the availability of the police, inconvenience caused by traffic, and degradation of the neighborhood. These items are based on a number of attitudinal questions, where respondents have to give their opinion on a number of statements. In the CATI and the CAPI surveys, the interviewers had a hidden answer category for 'no opinion' and 'refuse' (except for the opinion on the availability of the police, where the option 'no opinion' was presented directly to the respondents). The interviewers could choose one of these options if the respondent appeared not to

have an opinion or did not want to answer the question. In the web survey, these hidden answer categories were first not shown to the respondents. Respondents with no opinion could choose the neutral answer category, or skip the question. When they skipped it, a second screen with the same question appeared, this time with 'no opinion' and 'refuse' as additional options.

For the items that determine the opinion on the availability of the police, Van den Brakel et al. (2007) found that in the web survey more persons tended to choose the neutral answering category than in the CAPI and CATI survey. However, the increase in neutral answers was at the expense of both agree and disagree, so that in the final scale variable there was no effect. Hence, in the underlying items the mode effects are visible, but in the final construct the effects cancelled out. The number of persons that chose the 'no opinion' did not appear to be different in CAPI/CATI compared to the web survey.

The 'no opinion' category for the scale variables traffic inconvenience and neighbourhood degradation did not differ much between CAPI/CATI and the web survey either. However, the number of persons that chose the neutral category in the web survey largely increased. Furthermore, this increase was at the expense of the negative answering category. As a consequence, the web respondents had a significantly higher score on the variables neighborhood degradation and traffic inconvenience.

We also consider one other survey item, the level of satisfaction with the police performance in the neighborhood. Respondents could choose between one of the answering categories: very satisfied, satisfied, neutral, dissatisfied, very dissatisfied and don't know. The last category was explicitly presented to respondents, both in the CAPI/CATI survey as in the web survey. Van den Brakel et al. (2007) found that in the web survey, respondents more often answered 'neutral' at the expense of 'don't know'.

In table 9.10 the estimates for the survey items satisfied or very satisfied with police performance in the neighborhood (satisfaction), opinion on the availability of the police (availability), inconvenience caused by traffic (traffic inconvenience) and degradation of the neighborhood (neighborhood degradation) are shown. In parentheses are the corresponding standard errors of the estimated values. Van den Brakel et al. (2007) also tested for significance. Only for the variables traffic inconvenience and neighborhood degradation the differences between the estimates for the regular Safety Monitor and the mixed mode pilot are statistically significant. The design of the regular Safety Monitor does not include a web survey. The more neutral and less negative answers of the web respondents in the mixed mode pilot are reflected in the estimates for neighborhood degradation and traffic inconvenience in the mixed mode pilot. The

Table 9.10: *Estimates for survey items in the regular Safety Monitor and the mixed mode pilot SM-1. In parentheses are standard errors of the estimated values*

Variable	*Regular*	*Pilot*
Satisfaction	42.5% (0.37)	40.9% (1.13)
Availability	4.80 (0.02)	4.76 (0.06)
Traffic inconvenience	3.60 (0.02)	3.97 (0.05)
Neighborhood degradation	2.94 (0.01)	3.19 (0.03)

increase in the values for these two variables is significant.

In this section, we apply the mixed mode approaches to nonresponse adjustment presented in the previous section. Therefore we assume that there are no measurement errors. However, Van den Brakel et al. (2007) show that measurement errors are likely to be present in the answers from the web survey respondents.

9.6.2 Mixed mode approaches applied to the pilot SM-1

The software package Bascula is used for the application of the mixed mode approaches 1 to 3 described in the previous section. Bascula offers various weighting methods: post stratification, ratio estimation, linear weighting based on the generalized regression estimator and multiplicative weighting. For more information on Bascula, see Nieuwenbroek and Boonstra (2001) and Bethlehem (1998). We first describe the available auxiliary variables for the analysis. Then, we outline how the different approaches described in section 9.5 are applied to the pilot SM-1. In section 9.6.3 we present the resulting estimated survey items for the pilot based on the different approaches, and compare the results with the results from the regular Safety Monitor.

Auxiliary variables

The three approaches make use of auxiliary information. We linked the survey sample to some external registers. The information from these registers is available for both respondents and nonrespondents. Furthermore, we also have information that becomes available from the data collection process, i.e. the paradata (see section 9.4.2).

For the web survey no information about the data collection process is recorded. For the CAPI and CATI surveys we have information about the in-

terviewers (CAPI) and the number of call attempts (CATI). The paradata is hence not available for every person in the sample, but only for the persons that were approached in the corresponding data collection mode. Table 9.11 displays the available information.

Table 9.11: *Available auxiliary information for the pilot SM-1*

Variable	*Labels*
Administrative data	
Listed telephone	Yes, no
Degree of urbanization	Very strong, strong, moderate, low, not
Region	12 provinces, large cities: Utrecht, The Hague, Rotterdam and Amsterdam
Gender	Male, female
Age_6	$\leq$ 25, 26 - 35, ... , 56 - 65, 66+
Age_{11}	15 - 19, 20 - 24 , ... , 60 - 64, 65+
Ethnicity	Native, Western non-native, non-Western non-native
Position in household	Partner no children, partner with children, single parent, child, single, other
Size of household	1, 2, ... , 5+
Marital status	Not married, married/partner, widowed, divorced
Paradata	
Mode	Web, CATI, CAPI
Experience CAPI-interviewer	Missing, $<$ 1 year, 2 - 3, 3 - 4, $>$ 4
Gender CAPI-interviewer	Missing, male, female
Age CAPI-interviewer	Missing, $\leq$ 45, 46 - 50, 51 - 55, 56 - 60, $>$ 60
N^o of attempts (CATI)	1, 2, ... , 7, 8+

First approach

The first approach combines all the survey data into one single dataset. The respondents receive weights according to the reduced weighting model of Van den Brakel et al. (2007), as described in (9.16). The variable telephone is included in the weighting model. To some extent, the approach distinguishes between CATI and CAPI. However, the approach does not fully acknowledge the mixed mode design of the survey. The respondents are compared to the nonrespondents, no distinction is made with respect to mode. Since the most profound mode effect is to be expected with regard to the web survey respondents, mode specific effects will only partly be reflected in the adjustment weights. If the relationship between auxiliary variables and the response behaviour is not the same in each mode, this approach will not fully adjust for that difference.

Second approach - option one

Approach 2.1 also uses the combined dataset. One additional variable is added to the reduced weighting model of Van den Brakel et al. (2007), namely M^r = mode response path, i.e.

$$M_i^r = \begin{cases} 1, & \text{if person i responded to web survey} \\ 2, & \text{if person i did not respond to web, but did respond to CAPI} \\ 3, & \text{if person i did not respond to web, but did respond to CATI} \end{cases} \quad (9.17)$$

In the pilot SM-1 a total of 246 persons were not eligible for a follow-up. These are 85 persons that explicitly stated to the help desk at Statistics Netherlands that they did not want to cooperate in the web survey, plus a total of 161 early web survey responses that could not be used due to technical problems. This group of persons enables inclusion of variable M^r in the weighting model, because now the web response has a nonresponse counterpart. If there would not have been ineligible persons for the follow-up with CATI or CAPI, then M^r would reduce to two categories (only 2 and 3) and the adjustment would be made only for persons that were followed-up.

Including variable M^r in the weighting model hence ensures that persons that follow the same mode response path receive a path-specific correction. Because variable M^r is included in the weighting model additively, it only allows for a path-specific adjustment. If there is a different relationship between one of the auxiliary variables and M^r, this difference will only be reflected in the adjustment weights if M^r is crossed with that particular auxiliary variable. Including variable M^r additively in the weighting model only allows for a general adjustment for the mode response path a person followed.

One major drawback of this approach in a sequential mixed mode design, is that web survey respondents are not included in the prediction of survey items for nonrespondents. In this example, web respondents are only used because there is a small group of ineligible persons for the follow-up. If it weren't for this group, web respondents would not be included in the weighting adjustment approach.

Let us regard the situation without ineligible persons. For the calculation of adjustment weights, respondents in a specific path are compared to nonrespondents in that path. CATI respondents receive weights so that they represent all persons that did not respond to the web survey and have listed land-line telephones. The CAPI respondents receive weights to also represent web nonrespondents without a listed land-line telephone. Consequently, web survey respondents all receive weights equal to 1. This can be illustrated by looking at figure 9.5. Every person in the sample finally ends up in one of six categories: web respondent, web nonrespondent not eligible for follow-up, CATI respon-

dent, CATI nonrespondent, CAPI respondent or CAPI nonrespondent. There is no final category for web nonrespondents, since all web nonrespondents are followed-up in either CATI or CAPI.

Now, in the pilot SM-1 there is a category for web nonrespondents due to ineligible persons. However these persons are a very selective group that is not representative of all web nonrespondents. Furthermore, the small size of this group combined with a small number of variables that significantly explains the response behaviour can lead to an increased variance of the estimates.

Second approach - option two

Approach 2.2 models the response process in each of the response paths separately, thus allowing for path-specific information to be included in the adjustment models. This enables information to be inserted in the models only there where it adds explanatory power. Similar to approach 2.1, web respondents also receive weights equal to one. Furthermore, because the follow-up modes are modelled separately, the number of observations is small. Consequently, we find a small number of variables that significantly explains the response behaviour in the two follow-up paths (web to CATI and web to CAPI) since only a few sample elements followed those paths.

Weighting model (9.16) by Van de Brakel et al. (2007) was constructed for the pilot SM-1. We constructed separate weighting models for the mode response paths using logistic regression models. We used the software package Stata to apply the logistic models to our data. The auxiliary variables from table 9.11 are used for the analysis. However we excluded the variable telephone, because persons were assigned to either CATI or CAPI based on telephone ownership.

After fitting the models, we tested all variables for joint significance of its categories. The auxiliary variable with the highest p-value on this test is excluded from the model. This process is repeated untill only significant variables remain, at a level of 10% significance[4]. In tables 9.12 and 9.13 both full- and final models are given. We report χ^2 value and the corresponding p-values. The final models for the web to CATI different CAPI path become

$$\begin{aligned} CATI: \quad & DegreeUrbanization_5 + Ethnicity_3 + HouseholdSize_5 \\ & + PlaceFamily_5 \\ CAPI: \quad & DegreeUrbanization_5 + PlaceFamily_5 \end{aligned}$$

For both follow-up paths we now have a different weighting model. The next step in approach 2.2 comprehends the calculation of adjustment weights in Bascula. The respondents in each group received weights according to the

[4]Due to the small sample size, we chose a higher level of significance than in the analyses in Chapters 3 and 4.

Table 9.12: *Approach 2.2 for the web to CATI path; χ^2-values and corresponding p-values (in parentheses) for full- and final model*

Variable	*Full model*	*Final model*
Degree of urbanization	4.35 (0.3611)	11.73 (0.0194)
Region	7.50 (0.8229)	-
Gender	1.39 (0.2383)	-
Age	11.12 (0.3481)	-
Ethnicity	15.15 (0.0005)	16.96 (0.0007)
Place in family	5.18 (0.2695)	11.32 (0.0232)
Size of household	9.59 (0.0878)	15.42 (0.0039)
Marital status	4.90 (0.1789)	-

Table 9.13: *Approach 2.2 for the web to CAPI path; χ^2-values and corresponding p-values (in parentheses) for full- and final model*

Variable	*Full model*	*Final model*
Degree of urbanization	6.78 (0.1479)	8.14 (0.0867)
Region	16.86 (0.1549)	-
Gender	0.00 (0.9988)	-
Age	6.04 (0.8116)	-
Ethnicity	3.88 (0.1435)	-
Place in family	11.88 (0.0183)	18.03 (0.0012)
Size of household	7.88 (0.1628)	-
Marital status	3.63 (0.3045)	-

corresponding weighting model. In figure 9.5 the CATI- and CAPI-respondents are now assigned weights. One level up, we find the group of persons that did not respond to the web survey and that was not eligible for a follow-up. This group consists of persons that either responded to the web survey very quickly, or explicitly refused to cooperate. Even though it is a small group, it is interesting to include these persons in the analysis. They seem to have a different response behaviour than the other groups. If an analysis using auxiliary information shows that this group is different on some of these characteristics as well, we may obtain different estimates for the survey items. However, because the group is small, it may also lead to a large variance.

The last step of approach 2.2 comprehends the analysis of the weighted CATI- and CAPI respondents and the unweighted web respondents to the ineligibles. We thus apply the adjustment weights that are produced by the earlier steps of approach 2.2 to the CATI- and the CAPI respondents. Next, the

weighted respondents are combined with the web respondents for the analysis of the ineligible persons. This is the last step of approach 2.2. The model is displayed in table 9.14. Hence, there are three variables on which the group of ineligibles is different from the other sample elements: gender, age and size of household. The final model for the web respondents and the weighted CATI- and CAPI respondents becomes

$$Gender_2 + Age_{11} + HouseholdSize_5 \tag{9.18}$$

This model is used in Bascula to calculate the final adjustment weights for approach 2.2.

Third approach

Approach 3 makes use of paradata and estimates the survey items simultaneously in all modes. We, however, have no information on the distinct response types in the pilot SM-1. Furthermore, the simultaneous estimation as described in section 9.5 requires additional research. The methods have yet to be programmed. Therefore, we restrict the analysis to an exploration of the effect of mode specific paradata.

In table 9.11 we presented the available paradata. This information is not available for every person in the sample, but only for the persons that are approached in the corresponding data collection mode. Therefore, we could not use it in approach 1 and 2. Approach 3, however, does allow for the inclusion of mode specific data. Similar to approach 2.2, the only reason why the web respondents are included in the weighting adjustment is the small group of ineligible persons.

We first model the response behaviour in the web to CATI response path, the results are given in table 9.15. There is only one paradata variable for this

Table 9.14: *Approach 2.2 for the ineligibles; χ^2-values and corresponding p-values (in parentheses) for full- and final model*

Variable	*Full model*	*Final model*
Degree of urbanization	1.41 (0.2275)	-
Region	0.86 (0.5876)	-
Gender	4.01 (0.0454)	3.32 (0.0684)
Age	1.51 (0.1299)	2.19 (0.0159)
Ethnicity	1.69 (0.1857)	-
Place in family	0.74 (0.5650)	-
Size of household	1.74 (0.1225)	2.82 (0.0152)
Marital status	0.41 (0.7459)	-

path, namely the number of call attempts in 8 classes. It is not surprising that this variable turns out to be significant in explaining response behaviour, since a large number of call attempts will more frequently occur for persons that cannot be contacted. The final model for the web to CATI response path becomes

$$DegreeUrbanisation_5 + Ethnicity_3 + HouseholdSize_5 + PlaceFamily_5 + N^oCallAttempts_8 \quad (9.19)$$

For the web to CAPI response path there is more paradata. We linked interviewer information to every person in the sample. The gender, age and years of experience of the interviewers are known for almost all interviewers. We did not, however, allow for interviewer variance in the models and therefore the standard errors for this model will be underestimated. We had to discard 36 persons (all nonrespondents) from the CAPI follow-up due to the missing interviewer information. The total for the web to CAPI group is then 577. The sample total becomes $3,579$. In table 9.16 we give the full- and final model. Of the included paradata, only the gender of the interviewer appears to be related to the response behaviour. Furthermore, again place in family turns up in the response model. The final model becomes

$$PlaceFamily_5 + GenderInterviewer_2 \quad (9.20)$$

We used the response models in (9.19) and (9.20) to calculate the adjustment weights in Bascula.

The last step of approach 3 comprehends the comparison of these weighted groups and the unweighted web respondents to the total sample, thus including the ineligible persons. The results of this analysis are given in table 9.17. The model for the last step becomes

$$Age_{11} + HouseholdSize_5 \quad (9.21)$$

This model is used in Bascula to calculate the final adjustment weights for approach 3.

9.6.3 Summary of the models and the results

In table 9.18 we summarise all the adjustment approaches. The resulting estimates for the survey items are given in table 9.19. The last column of table 9.19 gives the values of the estimated items in the regular Safety Monitor. These values can be used as a reference.

The resulting estimates from the three approaches that we applied to the mixed mode pilot SM-1 do not seem to differ much. The only approach that shows slightly different values for the items, is approach 2.1. With this approach,

Table 9.15: *Approach 3 for the web to CATI path; χ^2-values and corresponding p-values (in parentheses) for full- and final model*

Variable	*Full model*	*Final model*
Degree of urbanization	3.89 (0.4214)	10.45 (0.0335)
Region	6.26 (0.9021)	-
Gender	1.55 (0.2134)	-
Age	10.64 (0.3863)	-
Ethnicity	18.60 (0.0003)	17.99 (0.0004)
Place in family	4.51 (0.3419)	10.01 (0.0402)
Size of household	11.50 (0.0215)	13.98 (0.0073)
Marital status	2.33 (0.5062)	-
N^o of call attempts	121.88 (0.0000)	125.23 (0.0000)

the estimated level of satisfaction is higher than in the other approaches, as is the opinion on the availability of the police. For the other two variables, traffic inconvenience and neighborhood degradation, the estimates are lower than in the other approaches. In general, it seems that in approach 2.1 the answers are more positive towards police performance and safety aspects.

When we compare the mixed mode approaches with the regular Safety Monitor, we see that in the regular Safety Monitor the estimated values are more positive than in the mixed mode pilot. In the adjustment approaches, we ignored the fact that there are measurement errors. Hence, the estimates based on the pilot SM-1 may differ from the regular Safety Monitor. As we discussed in section 9.6.1, it appears that the more neutral answers of the web survey respondents are reflected in the lower estimates for traffic inconvenience and neighborhood degradation. None of the approaches was able to fully adjust for this effect. The estimates based on approach 2.1 were closest to the regular Safety Monitor. A possible explanation for this result, is that in approach 2.1 a general adjustment for the mode is added to the other nonresponse adjustments. The web respondents seemed to have in general a more negative attitude towards safety and police performance, and approach 2.1 adjusted for that effect for the web respondents. Approach 1 did not perform a mode specific adjustment. In approaches 2.2 and 3, the web respondents hardly received any adjustment weights, only with respect to the difference with the ineligibles.

This result would argue for the use of approach 2.1, extending the regular nonresponse adjustment with an additional variable to account for the mixed mode design. This approach is also suggested by De Leeuw (1998). However, if we look at the weighting models for approaches 2.2 and 3, we see that in each of the

Table 9.16: *Approach 3 for the web to CAPI path; χ^2-values and corresponding p-values (in parentheses) for full- and final model*

Variable	*Full model*	*Final model*
Degree of urbanization	5.79 (0.2153)	-
Region	12.89 (0.3771)	-
Gender	0.01 (0.9149)	-
Age	10.64 (0.3863)	-
Ethnicity	3.95 (0.1387)	-
Place in family	13.21 (0.0103)	21.77 (0.0002)
Size of household	8.04 (0.1538)	-
Marital status	4.38 (0.2236)	-
Gender interviewer	6.12 (0.0133)	7.10 (0.0077)
Age interviewer	2.55 (0.6350)	-
Experience interviewer	0.87 (0.8325)	-

Table 9.17: *Approach 3 for the ineligibles; χ^2-values and corresponding p-values (in parentheses) for full- and final model*

Variable	*Full model*	*Final model*
Degree of urbanization	1.84 (0.1182)	-
Region	0.96 (0.4885)	-
Gender	collinear	-
Age	1.49 (0.1358)	2.22 (0.0144)
Ethnicity	1.53 (0.2177)	-
Place in family	0.69 (0.5954)	-
Size of household	1.88 (0.0947)	3.04 (0.0097)
Marital status	0.40 (0.7550)	-

follow-up modes different variables entered the response models. Furthermore, in approach 3 the paradata entered the response models for both the web to CATI and the web to CAPI group. It appears that the relationship between response behaviour and auxiliary variables is different in different modes.

We observed that in approaches 2.2 and 3 the web respondents hardly played a role in the adjustment for nonresponse. Because the nonrespondents to the web survey were all (except for the group of ineligibles) followed-up in either CATI or CAPI, there were no nonrespondents left to provide the web respondents with an adjustment weight. Furthermore, the size of the CATI and especially CAPI group was small. This caused the response models to be small too. We did not consider the variance of the weights and the resulting estimates. We suspect,

Table 9.18: *Summary of the models for mixed mode approaches 1, 2.1, 2.2 and 3*

Approach	*Path*	*Model*
1		$MaritalStatus_4 + DegreeUrbanization_5 + Region_{13}$ $+Householdsize_5 + Age_{11} \times Gender_2 +$ $+Telephone_2 \times Age_6$
2.1		$MaritalStatus_4 + DegreeUrbanization_5 + Region_{13}$ $+Householdsize_5 + Age_{11} \times Gender_2 +$ $+Telephone_2 \times Age_6 + M_3^r$
2.2	web to CATI	$DegreeUrbanization_5 + Ethnicity_3 + HouseholdSize_5$ $+PlaceFamily_5$
2.2	web to CAPI	$DegreeUrbanization_5 + PlaceFamily_5$
2.2	ineligibles	$Gender_2 + Age_{11} + HouseholdSize_5$
3	web to CATI	$DegreeUrbanization_5 + Ethnicity_3 + HouseholdSize_5$ $+PlaceFamily_5 + N^oCallAttempts_8$
3	web to CAPI	$PlaceFamily_5 + GenderInterviewer_2$
3	ineligibles	$Age_{11} + HouseholdSize_5$

Table 9.19: *Resulting estimates for the survey items*

	Approach				
Survey item	*1*	*2.1*	*2.2*	*3*	*Regular*
Satisfaction	41.0%	41.5%	41.2%	41.0%	42.5%
Availability	4.80	4.85	4.80	4.81	4.80
Traffic inconvenience	3.89	3.79	3.88	3.88	3.60
Neighborhood degradation	3.13	3.06	3.11	3.17	2.94

however, that due to the nested process in the sequential mixed mode design the variance in approaches 2.2 and 3 will be larger than in approaches 1 and 2.1.

Approach 3 as described in section 9.5 has not been applied to the SM-1 pilot. The procedures for estimation in approach 3 are complex and need further research. This approach can possibly overcome the drawbacks of approach 2.2 because the adjustment for nonresponse bias is done simultaneously for all modes. The main motivation is efficiency; by combining all data the relation between survey items and auxiliary variables can be modelled based on larger sets of respondents. Consequently, the accuracy of estimates will improve, however, at the expense of complexity and computation time.

All approaches assume implicitly that measurement error is not present in

the survey response. If some or all modes do suffer from measurement error, then the adjustment techniques will in general also lead to biased estimates. The approaches all balance the measurement error in the different modes, but each to a different extent. Approaches 1 and 2.1 simply increase the amount of measurement error where the response is high. Approaches 2.2 and 3 mix the measurement error according to the proportions in which the modes are mixed in the sample.

If we assume that the regular Safety Monitor produces the best survey estimates, then based on the results in table 9.19 our recommendation would be to use the more simple approach 2.1. The main drawback of approaches 2.2 and 3 is caused by the nesting of the data collection modes. This nesting is inherent to a sequential mixed mode design, where nonrespondents are re-approached in other modes than the original data collection mode. However, in a concurrent mixed mode design this will not happen. Further research therefore involves application of the mixed mode approaches to a concurrent mixed mode survey and further development of the methodology for simultaneous estimation of the survey items including the relation between survey item and response behaviour (approach 3).

9.7 Concluding remarks

Approaches 1, 2.1 and 2.2 produce weights for respondents that can be used for all survey items. These approaches are based on the calibration estimator, implemented as a linear weighting adjustment. Approach 3 models the relationship between survey item and response behaviour in the different modes using a multivariate probit model. This relationship can be different for each survey item and this approach hence produces a set of different weights for every survey item. Furthermore, the multivariate probit model in approach 3 is described in terms of a continuous survey item. The model in approach 3 can be extended to handle ordered and unordered categorical survey items, as we showed in Chapter 8.

If we use the survey estimates from the regular Safety Monitor as a benchmark, the results from the pilot SM-1 suggest that including a variable to account for the mode strategy in the linear weighting model is the best way to adjust for nonresponse in mixed mode data collection strategies. In pilot SM-1, the web respondents in general gave answers to the survey questions that were more neutral than the respondents to the other modes, i.e. satisficing in the self-administered mode. Inclusion of a variable that distinguishes between dif-

ferent modes allows for a general shift of survey estimates, which in this pilot was exactly the adjustment that was needed.

However, we do find evidence that paradata may have an influence on the adjustment methods. In approach 3, some paradata variables were included in the adjustment models. However, we did not allow for interviewer variance that is associated with the inclusion of paradata on the interviewer. Further refinements can be made to the approaches described by allowing for the inclusion of interviewer variables, i.e. the refinement described for the multilevel model in Chapter 8. Furthermore, the survey estimates based on approach 3 were not closer to the estimates based on the reference survey. This, however, could be caused by the fact that they were not simultaneously estimated. We cannot check this hypothesis, as further development of software for the application of the methodology described in approach 3 is needed. However, despite the issues concerning paradata (randomness, causal relation to response behaviour), these results seem to indicate that inclusion of paradata is valuable in nonresponse adjustment and therefore the tailored approach should be preferred over the uniform approach.

For now, we would advise survey researchers that work with mixed mode surveys to include at least an additional variable for the mode into their nonresponse adjustment models. This will capture the general mode effect as described in the application to the pilot SM-1.

Mixed mode data collection has become common practice to conduct a survey. There are a number of issues concerning mixed mode. In this chapter, we have developed methods to tackle one issue: combined nonresponse adjustment in mixed mode surveys. The methods that we propose combine data from different modes and adjust for nonresponse bias at the same time. We have, however, not dealt with measurement errors. Instead, we assumed that measurement errors are not present in the data. This is not a realistic assumption. The administration is different per mode and can be either self-administered or interviewer-assisted. Furthermore, per mode the data can be collected either on a laptop, by telephone, over the mail or web. This can lead to answers that differ per mode. Indeed, a lot of research is aimed at minimising measurement error. However, it is likely that measurement errors are present in the data after all, and therefore it is important that future research is directed at the development of methods that explicitly account for these mode-effects.

Chapter 10

Summary and Conclusions

In this thesis, we have developed methods for the treatment of unit nonresponse in sample surveys of households. Sample elements either respond to the survey request, or they do not. However, the reasons for nonresponse vary. This leads to different types of respondents. For example, persons that cannot be contacted during the fieldwork period often have different reasons not to respond than persons that refuse participation. To compare surveys, over time or across countries, it is important to distinguish between different types of respondents. When the composition of the nonresponse varies over surveys, and in the treatment of nonresponse no distinction is made between different types of respondents, the conclusions based on the surveys will reflect changes in the composition of the response instead of real changes in society. Furthermore, the causes, correlates and the effect on survey estimates are known to be different for different response types. The methods for nonresponse analysis and adjustment for nonresponse bias hence fit the data better when different (non)respondents are not combined into one group, but distinguished by cause. Throughout this thesis, we therefore distinguish between different response types.

Traditionally, treatment of nonresponse starts by reducing nonresponse in the field. Despite methods to prevent nonresponse, the response will never be 100%, nor will it be completely non-selective. Hence after the data have been collected, an estimation technique is applied to the data to adjust for nonresponse bias. Recently, nonresponse reducers have shifted focus from increasing response rates to obtaining a more balanced composition of the response. Amongst others, the research into representativeness indicators described in Chapter 5 of this thesis has contributed to this shift in paradigm. The tendency of research

on nonresponse reduction to balance the composition of the response resembles nonresponse adjustment, where the difference is in the timing since nonresponse adjustment balances the composition after the data have been collected.

On the interface of nonresponse reduction and adjustment for nonresponse bias lies analysis of response behaviour. Nonresponse analysis reveals how respondents differ from nonrespondents. This information is useful for the construction of weighting models to adjust for nonresponse bias, as well as for the construction of models to predict response behaviour. At the same time, analysis of nonresponse points at characteristics of the sample that have become unbalanced due to nonresponse. The nonresponse can be efficiently reduced by concentrating nonresponse reduction efforts on these unbalanced groups. In Chapter 5 of this thesis, we have developed an indicator for the quality of survey response that can be seen as a tool for the analysis of response behaviour that bridges the gap between adjustment and reduction of nonresponse. This so-called R-indicator is an indicator for the representativeness of survey response. It describes how balanced the composition of the response is compared to the composition of the sample, with respect to a set of predefined characteristics. The R-indicator can be used for multiple purposes. It can serve as an indicator for survey quality, supplementary to the response rate. But it can also be used during the fieldwork, to efficiently enhance response by concentrating fieldwork efforts on underrepresented groups. By using the R-indicator in an early stage of the data collection process, different strategies can be followed based on the course of the data collection. This will lead to a responsive survey design (Groves and Heeringa, 2005), or dynamic data collection strategies (Bethlehem and Schouten, 2008). At a time when survey costs and respondents' burden have to be reduced while preserving survey quality, dynamic survey design has great potential.

The analysis of response behaviour becomes increasingly important in dynamic survey design, since response behaviour will serve as the input for constructing different strategies. The R-indicator serves as a tool for analysing response behaviour; it summarizes multivariate information on response behaviour in an easy to interpret, univariate figure. The R-indicator is based on estimated response probabilities, or response propensities. Several methods can be used to compute these. As we have outlined before, different response types should be distinguished, leading to a sequential representation of the response process. For example, sample elements have to be contacted first, before we can observe whether or not they participate. In Chapter 8 of this thesis, we have developed methods that account for the sequential nature of the response process. These methods consist of a number of equations; one for every response type. Like

this, the response process is closely followed and it becomes possible to use different variables in each equation, as well as to allow for different relations between variables and the type of respondent. Another issue that we have dealt with, is the correlation between different types of response. A correlation arises when there are unobserved characteristics that influence more than one type of response, or when sample elements are misclassified.

For researchers that are interested in the quality of the obtained response, the R-indicator can be computed to supplement the response rate. If, however, the aim of analysing response behaviour is to build a weighting model for nonresponse adjustment, the described methods in Chapter 8 can be applied. Depending on the amount of auxiliary information, both socio-economic and demographic as well as information about the data collection process or paradata, different models can and should be used. For instance, when only little information is available it is not possible to closely describe the response process. However, when more detailed information is available, we have shown in Chapters 3 and 8 that by using more complex methods we are better able to understand the causes and correlates of (non)response. It is therefore important to collect good auxiliary information. Preferably, this information is obtained by linking the survey sample to one or more registers. Like this, individual data about both respondents and nonrespondents is obtained. Other ways to collect data on nonrespondents are: using the population characteristics published by the national statistical office; collecting interviewer observations and other information from the data collection process, i.e. paradata; or re-approaching nonrespondents.

A topic for future research is the integration of methods to compute response propensities with different types of respondents and the R-indicator. In addition, a different R-indicator can be constructed for each type of respondent. This information could prove useful in dynamic survey design, for example when the fieldwork organisation employs measures to reduce the non-contact rate but does not attempt a refusal conversion. The R-indicator can be employed to compute the representativeness of the subsample of respondents, if auxiliary information is available for respondents and nonrespondents. Organisations that do not have this type of information cannot compute the indicator. However, national statistical offices (NSO's) publish population distributions for specific domains. This information is publicly available. Hence, another topic for future research is the development of an indicator for representativeness that can be computed with aggregate population information. Within the 7th Framework of the European Union, in 2008 a project has started that aims at the development of these indicators (Bethlehem and Schouten, 2008). This RISQ project (Rep-

resentativity Indicators for Survey Quality) investigates how we can measure representativeness and how we can use the R-indicator to compare surveys or to construct dynamic survey designs.

Even with a dynamic survey design, some nonresponse will always remain. Therefore, the need for nonresponse bias adjustment once the data have been collected also remains. In Chapter 6 of this thesis, we have presented an overview of nonresponse adjustment methods. Depending on the type of auxiliary information, different estimation techniques can be applied. For instance, when the researcher knows individual values for the respondents only and in addition has the population distributions published by the NSO, then calibration estimators (under which the generalised regression estimator and post-stratification) can be applied. However, when the researcher can obtain information on the nonrespondents by linking the survey sample to a register, besides the calibration estimators also the propensity score methods can be used. Another way of obtaining information about nonrespondents, is to re-approach nonrespondents. In Chapter 4 of this thesis we have described the re-approach of nonrespondents to the Dutch LFS with the call-back approach and the basic-question approach. To validate the results from both re-approaches we used linked data from several registers. It turned out that the responding households in the call-back approach were different from the regular LFS-respondents. Furthermore, these households resembled the remaining nonrespondents. Hence, the composition of the response improved by adding the call-back respondents. The results for the basic-question approach were less satisfactory, but this was caused by the design of the approach which led to a confounding effect of telephone ownership. The additional information that is obtained with the re-approach can also be used for nonresponse bias adjustment. How this should be done is described in, for example, Bethlehem et al. (2006) or Laaksonen and Chambers (2006).

The methods that we describe in Chapter 6 do not account for the different causes of nonresponse. Therefore, in Chapter 8 we have developed nonresponse adjustment methods that take into account the different types of respondents. These methods consists of separate equations for the response types, and an additional equation for the survey item. They are in fact an extension of the methods for nonresponse analysis in Chapter 8. The nonresponse bias in the survey item is modelled by means of correlations between the different response types and the survey item.

The reason to distinguish different types of response comes from the intuition that the fieldwork process encompasses information that is useful in explaining response behaviour and, ultimately, in adjustment for nonresponse bias. However, the amount of information about the data collection process, or paradata,

is different for each response type. For instance, interviewer observations are available for contacted sample elements only. The methods that we have developed in Chapter 8 facilitate the inclusion of paradata because each response type is described by its own equation. However, again the applicability of these methods depends on the available information. In the best situation, there is detailed socio-economic and demographic information for all sample elements, as well as a large amount of paradata. This is the situation at NSO's in the The Netherlands, the Scandinavian countries and some other North-European countries. However, persons working in commercial market research and academics are usually less fortunate when it comes to auxiliary information. The methods described in Chapter 8 can only be applied when the information is available for all elements in the sample. When there is not much available information it is not possible to closely follow the response process and the methods described in Chapter 6 will be more appropriate. However, when there is detailed information about the sample elements, but no paradata, it is still possible to apply the methods described in Chapter 8.

Methods for nonresponse adjustment have improved by the use of auxiliary information, and hence profit from the availability of registers. However, nonresponse adjustment is also challenged by developments in survey design. We already briefly mentioned dynamic survey design. The first steps towards such designs have already been taken in the development and implementation of mixed mode surveys. Mixed mode data collection has become common practice to conduct a survey. By using a mixture of modes, data collection can be done cheaper and faster at the same quality, if performed in a clever way. From the perspective of nonresponse adjustment, the question arises how to combine data collected in different modes into one survey estimate? In Chapter 9 of this thesis, we have developed mixed mode-methods that combine data from different modes, while describing the fieldwork process in each mode separately. The methods closely follow the data collection process, by distinguishing different types of response as well as the sequence and combination of modes. However, if detailed information is not available to closely follow the fieldwork process and the response process, we recommend survey researchers that work with mixed mode surveys to at least include an additional variable for the mode into their nonresponse adjustment models. This will capture the general mode effect.

By explicitly modelling response behaviour (Chapter 8) and by distinguishing the different modes in the data collection process (Chapter 9), it becomes possible to include paradata besides the usual demographic and socio-economic information. However, this can also result in a large number of equations and, consequently, a large number of underlying assumptions. Estimation of both

methods, the nonresponse adjustment methods with different response types and the mixed mode-methods, requires evaluations of multiple integrals of the multivariate normal distribution. We have identified some important directions for future research. First, the estimation of these methods can be simplified by using Markov Chain Monte Carlo methods or Bayesian estimation. Secondly, since the methods depend on the distributional assumptions, future research should be directed at assessing the robustness to the assumptions and the development of alternative models. Finally, the described methods can only be applied to one survey item at the time. If the survey consists of a large number of items, it is unpractical to perform a different adjustment for each survey item, or for each group of survey items. To improve the practical value of these methods, future research should focus on how to combine these methods for different survey items, both for the nonresponse adjustment methods described in Chapter 8 and the mixed mode-methods in Chapter 9.

Besides mixed mode data collection, another challenge is the increased use of registers. Especially for business surveys, register-based surveys are common use nowadays. In register-based statistics, there is no sampling error. Furthermore, nonresponse occurs less frequently or is even completely absent. The main cause of error is related to the framework and results in a coverage error. Like nonresponse, undercoverage is caused by non-observation. The R-indicator can be used as a quality indicator for the undercoverage of a register. Furthermore, in Chapter 7 of this thesis we have shown that adjustment methods for nonresponse can be applied to adjust for errors caused by undercoverage of telephone households. It is likely that in register-based surveys similar problems exist and, hence, the same methods can be applied.

The methods that we have developed in this thesis can be applied to other situations, for example nonresponse in business surveys. The available auxiliary information for business surveys is less rich than for household surveys. However, information about the size of the business, the number of employees, or the branch to which a business belongs is generally available. The representativeness of the response to a business survey can also be computed with the R-indicator. For example to follow the representativeness of the response over time, to see whether at some point in time the subsample of responding businesses is representative enough for the production of so-called flash statistics. In addition, the use of registers is more prevailing in business surveys than it is in households surveys, since personal information can less frequently be found in a register as opposed to the economic information from businesses. Therefore, the application of adjustment methods to errors caused by undercoverage in business surveys seems promising.

In this thesis we have focussed on unit nonresponse. Unit nonresponse causes missing data on survey items, but in addition some observations of survey items may be missing due to item nonresponse. Treatment of nonresponse therefore also requires a decision on how to treat item nonresponse. Sensitive survey items are more susceptible to item nonresponse than others. For instance, the income question is known to suffer severely from item nonresponse (see, for example, Frick and Grabka 2005). The R-indicator can be used to compute the representativeness of individual items. Item nonresponse is commonly dealt with by imputation, see for instance Little and Rubin (2002). This approach replaces the missing values for the survey items by proxy values. These proxy values are estimated by some imputation procedure. A combined approach of imputation and weighting can be used to handle both item- and unit nonresponse, see for example Särndal and Lundström (2005). The models described in Chapter 8 can be applied to adjust for item nonresponse. The auxiliary information that is used in these models for a description of the response process can be extended with answers to other, related survey questions.

Bibliography

AAPOR (2006): "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 4th edition.," Discussion Paper Kansas: AAPOR, The American Association For Public Opinion Research.

Agresti, A. (2002): *Categorical Data Analysis*. John Wiley & Sons, Inc.

Albert, J., and S. Chib (1993): "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88(422), 669 – 679.

Anderson, D., and M. Aitkin (1985): "Variance Component Models with Binary Response: Interviewer Variability," *Journal of the Royal Statistical Society, Series B*, 47(2), 203 – 210.

Ariel, A., D. Giessen, F. Kerssemakers, and R. Vis-Visschers (2008): "Literature Review on Mixed-Mode Studies," Technical paper, Statistics Netherlands.

Arts, C., and E. Hoogteijling (2002): "Het Sociaal Statistisch Bestand 1998 en 1999 (in Dutch)," *Sociaal Economische Maandstatistiek*, 12, 13–21.

Beaumont, J.-F. (2005): "On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse through Weight Adjustment," *Survey Methodology*, 31(2), 227 – 231.

Bentler, P. (1990): "Comparative Fit Indexes in Structural Models," *Psychological Bulletin*, 107(2), 238–246.

Bertino, S. (2006): "A Measure of Representativeness of a Sample for Inferential Purposes," *International Statistical Review*, 74, 149–159.

Bethlehem, J. (1988): "Reduction of Nonresponse Bias through Regression Estimation," *Journal of Official Statistics*, 4(3), 251–260.

——— (1998): "Bascula 4.0 for Adjustment Weighting: Tutorial," Discussion paper, Statistics Netherlands.

——— (1999): "Cross-sectional Research.," in *Research Methodology in the Social, Behavioural & Life Science*, ed. by H. Adèr, and G. Mellenbergh, pp. 110–142. Sage Publications, London.

——— (2002): "Weighting Nonresponse Adjustments Based on Auxiliary Information," in *Survey Nonresponse*, ed. by R. Groves, D. Dillman, J. Eltinge, and R. Little, pp. 275 – 288. Wiley: New York.

——— (2007): "Reducing the Bias of Web Survey Based Estimates," Discussion paper 07001, CBS Voorburg, Available at www.cbs.nl.

Bethlehem, J., F. Cobben, and B. Schouten (2006): "Nonresponse in Household Surveys," CBS Voorburg, Course book for European Statistical Training Program.

Bethlehem, J., and W. Keller (1987): "Linear Weighting of Sample Survey Data," *Journal of Official Statistics*, 3(2), 141–153.

Bethlehem, J., and H. Kersten (1986): "Werken met Nonrespons (in Dutch)," Ph.D. thesis, Statistische Onderzoekingen M30.

Bethlehem, J., and F. V. d. Pol (1998): "The Future of Data Editing," in *Computer assisted survey information collection*, ed. by M. Couper, R. Baker, J. Bethlehem, C. Clark, J. Martin, W. Nicholls II, and J. O'Reilly, pp. 201 – 222. John Wiley & Sons, Inc.

Bethlehem, J., and B. Schouten (2004): "Nonresponse Analysis of the Integrated Survey on Living Conditions (POLS)," Discussion paper 04004, Statistics Netherlands, Available at www.cbs.nl.

——— (2008): "Representativity Indicators for Survey Quality (RISQ), Description of Work," FP7 project 216036, Statistics Netherlands.

Biemer, P., and M. Link (2006): "A Latent Call-Back Model for Nonresponse," in *17th International Workshop on Household Survey Nonresponse, Omaha, Nebraska, USA*.

Biemer, P., and L. Lyberg (2003): *Introduction to Survey Quality*. John Wiley & Sons, Inc.

Binson, D., J. Canchola, and J. Catania (2000): "Random Selection in a National Telephone Survey: A Comparison of the Kish, Next-Birthday, and Last-Birthday Methods," *Journal of Official Statistics*, 16(1), 53–59.

Blau, P. (1964): *Exchange and Power in Social Life*. New York: John Wiley & Sons.

Brakel, J. V. d., K. V. Berkel, and B. Janssen (2007): "Mixed-Mode Experiment bij de Veiligheidsmonitor Rijk (in Dutch)," Technical paper DMH-2007-01-23-JBRL, Statistics Netherlands.

Brakel, J. V. d., and R. Renssen (1998): "A Field Experiment to Test Effects of an Incentive and a Condensed Questionnaire on Response Rates in the Netherlands Fertility and Family Survey," *Research in Official Statistics*, 3(1), 55–63.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1998): *Classification and Regression Trees*. Chapman & Hall, Florida, USA.

Callegaro, M., and T. Poggio (2004): "Where Can I Call You? The 'Mobile' Revolution and its Impact on Survey Research and Coverage Error: Discussing the Italian Case.," in *RC 33 Sixth International Conference on Social Science Methodology*.

CBS (2006): "Methoden en Definities. Enquête Beroepsbevolking 2005 (in Dutch)," Publicatie cbs-website, Statistics Netherlands, Available at www.cbs.nl.

CBS-wet (2004): "Wet op het Centraal Bureau voor de Statistiek," *Staatsblad*, pp. 695 – 710, Available at www.cbs.nl.

Christian, L., D. Dillman, and J. Smyth (2006): "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys," in *Second conference on Telephone Survey Methodology, Florida, USA*.

Cialdini, R. (1988): *Influence: Science and Practice*. Glenview, IL: Scott, Foresman and Co.

Clark, R., and D. Steel (2007): "Sampling within Households in Household Surveys," *Journal of the Royal Statistical Society - Series A*, 170(1), 63–82.

Cobben, F. (2007): "Mode effects in a basic question approach for the Dutch LFS," Discussion paper 03007, Statistics Netherlands, Available at www.cbs.nl.

Cobben, F., and J. Bethlehem (2005): "Adjusting Undercoverage and Nonresponse Bias in Telephone Surveys.," Discussion paper 05006, Statistics Netherlands, Available at www.cbs.nl.

Cobben, F., B. Janssen, K. V. Berkel, and J. V. d. Brakel (2007): "Statistical Inference in a Mixed-Mode Data Collection Setting," in *57th session of the ISI conference, Lisbon, Portugal*.

Cobben, F., and B. Schouten (2007): "Are you the next to have your birthday? Congratulations! You may answer some questions," Nota DMV-2007-02-15-FCBN, Statistics Netherlands.

——— (2008): "An Empirical Validation of R-Indicators," Discussion paper 08006, Statistics Netherlands, Available at www.cbs.nl.

Cochran, W. (1968): "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies.," *Biometrics*, 24, 205–213.

——— (1977): *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Couper, M. (1998): "Measuring Survey Quality in a CASIC Environment," in *Joint Statistical Meetings of the American Statistical Association, Dallas, TX*.

——— (2000): "Web Surveys: a Review of Issues and Approaches," *The Public Opinion Quarterly*, 64(4), 464–494.

——— (2005): "Technology Trends in Survey Data Collection," *Social Science Computer Review*, 23(4), 486 – 501.

Couper, M., and L. Lyberg (2005): "The Use of Paradata in Survey Research," in *56th session of the ISI conference, Sydney, Australia*.

Couper, M., and W. Nichols II (1998): "The History and Development of Computer Assisted Survey Information Collection Methods," in *Computer assisted survey information collection*, ed. by M. Couper, R. Baker, J. Bethlehem, C. Clark, J. Martin, W. Nicholls II, and J. O'Reilly, pp. 1 – 22. John Wiley & Sons, Inc.

Curtin, R., S. Presser, and E. Singer (2000): "The Effects of Response Rate Changes on the Index of Consumer Sentiment," *The Public Opinion Quarterly*, (64), 413–428.

Daalmans, J., A. Israëls, A. Loeve, and B. Schouten (2006): "Tussenrapportage Responscijfers EBB 2002-2004 (in Dutch)," nota TMO-R&D-2006-03-10-JDAS, Statistics Netherlands.

Das, M., A. Kapteijn, and A. Van Soest (2006): "An Advanced Multi-Disciplinary Facility for Measurement and Experimentation in the Social Sciences (MESS)," Research funded by nwo, Available at www.centerdata.nl.

Das, M., W. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33 – 58.

De Leeuw, E. (1992): "Data Quality in Mail, Telephone, and Face to Face Surveys.," Ph.D. thesis.

——— (2005): "To Mix or not to Mix Data Collection Modes in Surveys.," *Journal of Official Statistics*, 21(2), 233–255.

De Luca, G., and F. Peracchi (2007): "A Sample Selection Model for Unit and Item Nonresponse in Cross-Sectional Surveys," Research Paper Series 99, Centre for International Studies on Economic Growth.

De Ree, S. (1989): "Cluster Effect in the Labour Force Survey," *Netherlands Official Statistics*, 4(1), 32.

Dehija, R., and S. Wahba (1999): "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs.," *Journal of the American Statistical Association*, 94(448), 1053–1062.

Deville, J., and C. Särndal (1992): "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87(418), 376–382.

Deville, J., C. Särndal, and O. Sautory (1993): "Generalized Raking Procedures in Survey Sampling," *Journal of the American Statistical Association*, 88(423), 1013–1020.

Dillman, D. (1978): *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley-Interscience.

——— (2000): *Mail and Internet Surveys: The Total Design Method*. New York: John Wiley & Sons, Inc.

Dillman, D., and L. Christian (2005): "Survey Mode as a Source of Instability in Response Across Surveys," *Field Methods*, 17(1), 30 – 52.

Dillman, D., R. Sangster, J. Tarnai, and T. Rockwood (1996): "Understanding Differences in People's Answers to Telephone and Mail Surveys," in *Current issues in survey research: New directions for program evaluation series*, ed. by M. Braverman, and J. Slater, pp. 45 – 62. Jossey-Bass, San Francisco.

Dillman, D., and J. Tarnai (1991): "Mode Effects of Cognitive Designed Recall Questions: A Comparison of Answers to Telephone and Mail Surveys.," in *Measurement Errors in Surveys*, ed. by P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. J. Wiley & Sons, New Jersey.

Dubin, J., and D. Rivers (1989): "Selection Bias in Linear Regression, Logit and Probit Models," *Sociological methods and research*, 18(2), 360 – 390.

Duffy, B., K. Smith, G. Terhanian, and J. Bremer (2005): "Comparing Data from Online and Face-to-Face Surveys," *International Journal of Market Research*, 47(6), 615–639.

Durrant, G., and F. Steele (2007): "Multilevel Modelling of Refusal and Noncontact Nonresponse in Household Surveys: Evidence from Six UK Government Surveys," s3ri working paper m07-11, University of Southampton.

Efron, B., and R. Tibshirani (1993): *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Ellis, C., and J. Krosnick (1999): "Comparing Telephone and Face-to-Face Surveys in Terms of Sample Representativeness: A Meta-Analysis of Demographic Characteristics.," Technical Report Series nes010871, NES.

Ellis, R., C. Endo, and M. Armer (1970): "The Use of Potential Nonrespondents for Studying Nonresponse Bias," *The Pacific Sociological Review*, 13(2), 103–109.

Erwich, B., and J. Van Maarseveen (1999): *Een Eeuw Statistieken (in Dutch)*. CBS Voorburg/Stichting beheer IISG, Amsterdam.

Eurostat (2003): "The European Union labour force survey. Methods and definitions 2001," Discussion paper, Luxembourg: Office for Official Publications of the European Communities, Available at ec.europa.eu.

Feskens, R., J. Hox, G. Lensvelt-Mulders, and H. Schmeets (2006): "Collecting Data among Ethnic Minorities in an International Perspective," *Field Methods*, 18(3), 284–304.

Filion, F. (1976): "Exploring and Correcting for Nonresponse Bias Using Follow-Ups of Nonrespondents," *The Pacific Sociological Review*, 19(3), 401–408.

Fitzgerald, R., and L. Fuller (1982): "I Hear you Knocking but you can't Come in. The Effects of Reluctant Respondents and Refusers on Sample Survey Estimates," *Sociological Methods & Research*, 11(1), 3–32.

Fouwels, S., B. Janssen, and W. Wetzels (2006): "Experiment Mixed-Mode Waarneming bij de VMR (in Dutch)," Technical paper SOO-2007-H53, Statistics Netherlands.

Frick, J., and M. Grabka (2005): "Item Nonresponse on Income Questions in Panel Surveys: Incidence, Imputation and the Impact on Inequality and Mobility," *Journal Allgemeines Statistisches Archiv*, 89(1), 49–61.

Gallant, A., and D. Nychka (1987): "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 55(2), 363–390.

Gaziano, C. (2005): "Comparative Analysis of Within-Household Respondent Selection Techniques," *The Public Opinion Quarterly*, 69(1), 124–157.

Goodman, L., and W. Kruskal (1979): *Measures of Association for Cross-Classifications*. Berlin: Springer-Verlag.

Goor, A. v., and H. v. Goor (2003): "De Bruikbaarheid van de Methode van de Centrale Vraag voor het Vaststellen van Non-Responsvertekeningen: Vier Empirische Tests (in Dutch)," *Sociologische Gids*, 50, 378 – 391.

Goor, H. v., and S. Rispens (2004): "A Middle Class Image of Society," *Quality and Quantity*, 38(1), 35 – 49.

Gouweleeuw, J., and H. Eding (2006): "De Informele Economie: Analyse en Vergelijking van de Mixed-Mode en de Face-to-Face Respons (in Dutch)," Nota SOO-202797-03, Statistics Netherlands.

Goyder, J. (1987): *The Silent Minority: Nonrespondents on Sample Surveys*. Polity Press.

Greene, W. (1981): "Sample Selection Bias as a Specification Error: A Comment," *Econometrica*, 39(3), 795–798.

——— (2003): *Econometric Analysis, Fifth Edition*. Prentice-Hall.

Griffin, D., F. D.P., and M. Morgan (2001): "Testing an Internet Response Option for the American Community Survey," in *Annual conference of the AAPOR*.

Grijn, F. V. d., B. Schouten, and F. Cobben (2006): "Balancing Representativity, Costs and Response Rates in a Call Scheduling Strategy," in *17th International Workshop on Household Survey Nonresponse, Omaha, Nebraska, USA*.

Gronau, R. (1974): "Wage Comparisons - A Selectivity Bias," *Journal of Political Economy*, 82(6), 1119–1144.

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

——— (2006): "Nonresponse Rates and Nonresponse Bias in Household Surveys," *The Public Opinion Quarterly*, 70(5), 646–675.

Groves, R., R. Cialdini, and M. Couper (1992): "Understanding the Decision to Participate in a Survey," *The Public Opinion Quarterly*, 56(4), 475 – 495.

Groves, R., and M. Couper (1998): *Nonresponse in Household Interview Surveys*. Wiley series in probability and statistics. Survey methodology section.

Groves, R., D. Dillman, J. Eltinge, and R. Little (2002): *Survey Nonresponse*. New York: Wiley Series in Probability and Statistics.

Groves, R., and S. Heeringa (2006): "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs," *Journal of the Royal Statistical Society - Series A*, 169(3), 439 – 457.

Groves, R., and E. Peytcheva (2006): "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis," in *17th International Workshop on Household Survey Nonresponse, Omaha, Nebraska, USA*.

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

Groves, R., E. Singer, and A. Corning (2000): "A Leverage-Salience Theory of Survey Participation," *The Public Opinion Quarterly*, 64, 299 – 308.

Hajék, J. (1981): *Sampling from Finite Populations*. Marcel Dekker, New York, USA.

Hansen, M., and W. Hurwitz (1946): "The Problem of Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, 41, 517–529.

Heckman, J. (1974a): "Effects of Child-Care Programs on Women's Work Effort," *Journal of Political Economy*, 82(2, supplement), 136 – 163.

Heckman, J. (1974b): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42(4), 679–694.

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153 – 161.

Heer, d. W. (1999): "International Response Trends: Results of an International Survey," *Journal of Official Statistics*, 15(2), 129 – 142.

Heerwegh, D. (2003): "Explaining Response Latencies and Changing Answers using Client-Side Paradata from a Web Survey," *Social Science Computer Review*, 21(3), 360 – 373.

Heerwegh, D., K. Abts, and G. Loosveldt (2007): "Minimizing Survey Refusal and Noncontact Rates: Do our Efforts Pay Off?," *Survey Research Methods*, 1(1), 3–10.

Henly, M., and N. Bates (2005): "Using Call Records to Understand Response in Panel Surveys," in *Annual AAPOR meeting - ASA section on Survey Research Methods*.

Holbrook, A., J. Krosnick, D. Moore, and R. Tourangeau (2007): "Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes," *Public Opinion Quarterly*, 71, 325 – 348.

Horvitz, D., and D. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Hox, J. (1994): "Hierarchical Regression Models for Interviewer and Respondent Effects," *Sociological methods & research*, 22(3), 300 – 318.

——— (1995): *Applied Multilevel Analysis - Second Edition*. Amsterdam: TT-Publikaties.

Iannacchione, V. (2003): "Sequential Weight Adjustments for Location and Cooperation Propensity for the 1995 National Survey of Family Growth," *Journal of Official Statistics*, 19(1), 31 – 43.

Imai, K., and D. Van Dyk (2005): "A Bayesian Analysis of the Multinomial Probit Model using Marginal Data Augmentation," *Journal of Econometrics*, 124(2), 311 – 334.

Janssen, B. (2006): "Web Data Collection in a Mixed Mode Approach: An Experiment," in *Proceedings of the European Conference on Quality in Survey Statistics, Q2006, Cardiff, UK*.

Janssen, B., M. Schrooten, and W. Wetzels (2007): "Mixed-Mode Enquêtering bij Personen en Huishoudens 2005-2007: Een Overzicht (in Dutch)," Technical paper SOO-2007-H231, Statistics Netherlands, SOO Heerlen.

Jenkins, S., L. Cappellari, P. Lynn, A. Jäckle, and E. Sala (2006): "Patterns of Consent: Evidence from a General Household Survey," *Journal of the Royal Statistical Society - Series A*, 169(4), 701 – 722.

Johnston, J., and J. Dinardo (1997): *Econometric Methods*. McGraw-Hill.

Kalton, G., and I. Flores-Cervantes (2003): "Weighting Methods," *Journal of Official Statistics*, 19(2), 81–97.

Kass, G. (1980): "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Applied Statistics*, 29(2), 119–127.

Keeter, S., C. Miller, A. Kohut, R. Groves, and S. Presser (2000): "Consequences of Reducing Nonresponse in a National Telephone Survey," *The Public Opinion Quarterly*, 64, 125–148.

Kersten, H., and J. Bethlehem (1984): "Exploring and Reducing the Nonresponse Bias by Asking the Basic Question," *Statistical Journal of the United Nations*, ECE 2, 369–380.

Kish, L. (1949): "A Procedure for Objective Respondent Selection within the Household," *Journal of the American Statistical Association*, 44(247), 380–387.

——— (1965): *Survey Sampling*. New York: John Wiley & Sons.

Klaauw, B. V. d., and R. Koning (1996): "Some Applications of Semi - Nonparametric Maximum Likelihood Estimation," Discussion paper 96-161/7, Tinbergen Institute.

——— (2000): "Testing the Normality Assumption in Sample Selection Models with an Application to Travel Demand," Research Report 281, SOM.

Kohler, U. (2007): "Surveys from Inside: An Assessment of Unit Nonresponse Bias with Internal Criteria," *Survey Research Methods*, 1(2), 55 – 67.

Kott, P. (2006): "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors," *Survey Methodology*, 32(2), 133 – 142.

Kreuter, F. (2006): "Paradata - Introduction to the Special Interest Section International Workshop on HH NR 2007," in *17th International Workshop on Household Survey Nonresponse, Omaha, Nebraska, USA*.

Krosnick, J. (1991): "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5, 213 – 236.

Kruskal, W., and F. Mosteller (1979a): "Representative Sampling I: Non-Scientific Literature," *International Statistical Review*, 47, 13–24.

——— (1979b): "Representative Sampling II: Scientific Literature Excluding Statistics," *International Statistical Review*, 47, 111–123.

——— (1979c): "Representative Sampling III: Current Statistical Literature," *International Statistical Review*, 47, 245–265.

Kuusela, V. (1998): "A Survey on Telephone Coverage in Finland," *Zuma Nachrichten Special*, August, 113 – 119.

Kwak, N., and B. Radler (2002): "A Comparison between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality," *Journal of Official Statistics*, 18, 257 – 273.

Laaksonen, S. (2006): "Does the Choice of Link Function Matter in Response Propensity Modelling?," *Model Assisted Statistics and Applications I*, pp. 95–100.

Laaksonen, S., and R. Chambers (2006): "Survey Estimation Under Informative Nonresponse with Follow-up," *Journal of Official Statistics*, 22(1), 81–95.

Lavrakas, P., E. Stasny, and B. Harpuder (2000): "A Further Investigation of the Last-Birthday Respondent Selection Method and Within-Unit Coverage Error," in *American Statistical Association, proceedings of the Survey Research Methods Section*.

Lee, S. (2006): "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys," *Journal of Official Statistics*, 22(2), 329–349.

Lepkowski, J. (1988): "Telephone Sampling Methods in the United States," in *Telephone Survey Methodology*, ed. by R. Groves, L. Biemer, L. Lyberg, J. Massey, W. Nichols II, and J. Waksberg, pp. 73 – 98. Wiley, New York.

Lepkowski, J., and M. Couper (2002): "Nonresponse in the Second Wave of Longitudinal Household Surveys," in *Survey nonresponse*, ed. by R. Groves, G. Kalton, J. Rao, N. Schwarz, and C. Skinner, pp. 259 – 272. John Wiley & Sons, Inc.

Lin, I., and N. Schaeffer (1995): "Using Survey Participants to Estimate the Impact of Nonparticipation," *The Public Opinion Quarterly*, 49(2), 236–258.

Lind, K., M. Link, and R. Oldendick (2000): "A Comparison of the Accuracy of the Last Birthday Versus the Next Birthday Methods for Random Selection of Household Respondents," in *American Statistical Association, proceedings of the Survey Research Methods Section*.

Little, R. (1986): "Survey Nonresponse Adjustments for Estimates of Means," *International Statistical Review*, 54, 139–157.

Little, R., and D. Rubin (2002): *Statistical Analysis with Missing Data*. Hoboken (NJ): John Wiley & Sons, Inc.

Little, R., and S. Vartivarian (2005): "Does Weighting for Nonresponse Increase the Variance of Survey Means?," *Survey Methodology*, 31, 161 – 168.

Luiten, A. (2008): "Responsberekening Sociale Statistieken (in Dutch)," Methodenreeks Thema: Veldwerkorganisatie, Statistics Netherlands.

Lundström, S., and C. Särndal (1999): "Calibration as Standard Method for Treatment of nonresponse," *Journal of Official Statistics*, 15(2), 305–327.

Lynn, P. (2003): "PEDAKSI: Methodology for Collecting Data about Survey Non-Respondents," *Quality & Quantity*, 37, 239–261.

Lynn, P., R. Beerten, J. Laiho, and J. Martin (2002a): "Towards Standardisation of Survey Outcome Categories and Response Rate Calculations," *Research in Official Statistics*, 2002(1), 63–86.

Lynn, P., and P. Clarke (2002): "Separating Refusal Bias and Non-Contact bias: Evidence from UK National Surveys," *Journal of the Royal Statistical Society, Series D*, 51(3), 319 – 333.

Lynn, P., P. Clarke, J. Martin, and P. Sturgis (2002b): "The Effects of Extended Interviewer Efforts on Nonresponse Bias," in *Survey Nonresponse*, ed. by R. Groves, D. Dillman, J. Eltinge, and R. Little, pp. 135 – 148. New York: John Wiley & Sons, Inc.

Maddala, G. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

——— (1991): "A Perspective on the Use of Limited-Dependent and Qualitative Variables Models in Accounting Research," *The Accounting Review*, 66(4), 788 – 807.

Marsh, H., J. Balla, and R. McDonald (1988): "Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size," *Psychological Bulletin*, 103(3), 391–410.

Melenberg, B., and A. V. Soest (1993): "Semi-Parametric Estimation of the Sample Selection Model," Paper 9334, University of Tilburg - Center for Economic Research.

Merkle, D., and M. Edelman (2002): "Nonresponse in Exit Polls: A Comprehensive Analysis," in *Survey Nonresponse*, ed. by R. Groves, D. Dillman, J. Eltinge, and R. Little, pp. 243–258. Wiley: New York.

Nagelkerke, N. (1991): "Miscellanea. A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78(3), 691 – 692.

Nicoletti, C., and F. Peracchi (2005): "Survey Response and Survey Characteristics: Microlevel Evidence from the European Community Household Panel," *Journal of the Royal Statistical Society - Series A*, 168(4), 763 – 781.

Nieuwenbroek, N., and H. Boonstra (2001): "Bascula 4.0 Reference Manual," Technical paper 3554-99-RSM, Statistics Netherlands.

Oh, H., and F. Scheuren (1983): "Weighting Adjustment for Unit Nonresponse," in *Incomplete Data in Sample Surveys*, ed. by W. Madow, I. Olkin, and D. Rubin, vol. 2, pp. 143–184.

Oldendick, R., G. Bishop, S. Sorenson, and A. Tuchfaber (1988): "A Comparison of the Kish and Last Birthday Methods of Respondent Selection in Telephone Surveys," *Journal of Official Statistics*, 4(4), 307–318.

O'Neill, M. (1979): "Estimating the Nonresponse Bias Due to Refusals in Telephone Surveys," *The Public Opinion Quarterly*, 43(2), 218–232.

Pickery, J., and A. Carton (2005): "Hoe Representatief zijn Telefonische Surveys in Vlaanderen? (in Dutch)," Nota 4, Administratie Planning en Statistiek, Ministerie van de Vlaamse Gemeenschap.

Pierzchala, M. (2006): "Disparate Modes and their Effect on Instrument Design," in *10th International Blaise Users Conference, Papendal, The Netherlands*.

Pierzchala, M., D. Wright, C. Wilson, and P. Guerino (2004): "Instrument Design for a Blaise Multimode Web, CATI, and Paper Survey," in *9th International Blaise Users Conference, Québec, Canada*.

Poirier, D. (1980): "Partial Observability in Bivariate Probit Models," *Journal of Econometrics*, 12(2), 209–217.

Roberts, C. (2007): "Mixing Modes of Data Collection in Surveys: A Methodological Review," NCRM review papers 008, ESRC National Centre for Research Methods.

Roose, H., J. Lievens, and H. Waege (2007): "The Joint Effect of Topic Interest and Follow-Up Procedures on the Response in a Mail Questionnaire: An Empirical Test of the Leverage-Salience Theory in Audience Research," *Sociological Methods & Research*, 35(3), 410–428.

Rosenbaum, P., and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

——— (1984): "Reducing Bias in Observational Studies using Subclassification on the Propensity Score.," *Journal of the American Statistical Association*, 79(387), 516–524.

Rubin, D. (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592.

Salmon, C., and J. Nichols (1983): "The Next-Birthday Method of Respondent Selection," *The Public Opinion Quarterly*, 47(2), 270–276.

Särndal, C. (1980): "A Two-Way Classification of Regression Estimation Strategies in Probability Sampling," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 8(2), 165–177.

——— (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, 494 – 513.

Särndal, C., and S. Lundström (2005): *Estimation in Surveys with Nonresponse*. Chichester: John Wiley & Sons, Inc.

——— (2008): "Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator," *Journal of Official Statistics*, 24(2), 167 – 191.

Särndal, C., B. Swensson, and J. Wretman (1992): *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Schafer, J., and J. Graham (2002): "Missing Data: Our View of the State of the Art," *Psychological Methods*, 7(2), 147–177.

Schonlau, M., A. v. Soest, and A. Kapteyn (2007): "Are Webographics or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?," *Survey Research Methods*, 1(3), 155 – 163.

Schonlau, M., K. Zapert, L. Simon, K. Sanstad, S. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner, and S. Berry (2004): "A Comparison between Responses from a Propensity-weighted Web Survey and an Identical RDD Survey," *Social Science Computer Review*, 22(1), 128–138.

Schouten, B. (2004a): "Adjustment Bias in the Integrated Survey on Living Conditions (POLS) 1998.," Discussion paper 04001, Statistics Netherlands, Available at www.cbs.nl.

——— (2004b): "Adjustment for Bias in the Integrated Survey on Household Living Conditions (POLS) 1998," Discussion paper 04001, Statistics Netherlands, Available at www.cbs.nl.

——— (2007): "A Selection Strategy for Weighting Variables under a Not-Missing-At-Random Assumption," *Journal of Official Statistics*, 23(1), 51–68.

Schouten, B., and J. Bethlehem (2002): "Analyse van de Non-Respons uit het Aanvullend Voorzieningengebruik Onderzoek 1999 (in Dutch)," Discussion paper 781-02-TMO, Statistics Netherlands, Available at www.cbs.nl.

——— (2007): "Increasing Response Rates and Representativeness Using Follow-Up Surveys with Basic Questions," in *57th session of the ISI conference, Lisbon, Portugal*.

Schouten, B., and F. Cobben (2007): "R-Indicators for the Comparison of Different Fieldwork Strategies and Data Collection Modes," Discussion paper 07002, Statistics Netherlands, Available at www.cbs.nl.

Schouten, B., and G. De Nooij (2005): "Nonresponse Adjustment using Classification Trees," Discussion paper 05001, Statistics Netherlands, Available at www.cbs.nl.

Singer, E. (2002): "The Use of Incentives to Reduce Non Response in Household Surveys," in *Survey Nonresponse*, ed. by R. Groves, D. Dillman, J. Eltinge, and R. Little, pp. 163–177. Wiley: New York.

Skinner, C., D. Holt, and T. Smith (1989): *Analysis of Complex Surveys*. New York: Wiley and Sons.

Smith, J., M. Rhoads, and J. Shepherd (1998): "Getting from Here to There: Electronic Data Communications in Field Surveys," in *Computer assisted survey information collection*, ed. by M. Couper, R. Baker, J. Bethlehem, C. Clark, J. Martin, W. Nicholls II, and J. O'Reilly, pp. 307 – 330. John Wiley & Sons, Inc.

Smith, T. W. (1990): "Phone Home? An Analysis of Household Telephone Ownership.," *International Journal of Public Opinion Research*, 2(4), 369–390.

Snijkers, G. (2002): "Cognitive Laboratory Experiences: On Pre-Testing Computerised Questionnaires and Data Quality," Ph.D. thesis, PhD thesis Utrecht University.

Stinchcombe, A., C. Jones, and P. Sheatsley (1981): "Nonresponse Bias for Attitude Questions," *The Public Opinion Quarterly*, 45(3), 359–375.

Stoop, I. (2001): "Early, Late, Cooperative, Reluctant, Persuadable and Follow-Up Respondents. A Study of Nonresponse using Fieldwork Records, Frame Data, Wave Data and a Follow-Up Survey," in *12th International Workshop on Household Survey Nonresponse, Oslo, Sweden*.

——— (2004): "Surveying Nonrespondents," *Field Methods*, 16, 23–54.

——— (2005): "The Hunt for the Last Respondent. Nonresponse in Sample Surveys.," Ph.D. thesis, University of Utrecht.

Taylor, H., J. Bremer, C. Overmeyer, J. Siegel, and G. Terhanian (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(2), 127 – 136.

Tourangeau, R., and K. Rasinski (1988): "Cognitive Processes Underlying Context Effects in Attitude Measurement," *Psychological Bulletin*, 103, 299 – 314.

Vehovar, V., E. Belak, Z. Batagelj, and S. Cikic (2004): "Mobile Phone Surveys: The Slovenian Case Study," *Methodološki sveski. Advances in Methodology and Statistics*, 1(1), 1–19.

Vella, F. (1998): "Estimating Models with Sample Selection Bias: A Survey," *The journal of human resources*, 33(1), 127 – 169.

Ven, W. V. d., and B. V. Praag (1981): "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection," *Journal of Econometrics*, 17(2), 229 – 252.

Voogt, R. (2004): "I am not Interested: Nonresponse Bias, Response Bias and Stimulus Effects in Election Research," Ph.D. thesis, University of Amsterdam.

Wallgren, A., and B. Wallgren (2007): *Register-Based Statistics. Administrative Data for Statistical Purposes*. John Wiley & Sons, Inc.

Wilcox, J. (1977): "The Interaction of Refusal and Not-at-Home Sources of Nonresponse Bias," *Journal of Marketing Research*, 14(4), 592 – 597.

Samenvatting (Summary in Dutch)

In dit proefschrift bespreken we methoden voor het omgaan met non-respons in survey onderzoek voor personen en huishoudens. Wanneer een persoon wordt geselecteerd voor deelname aan een survey, kunnen er twee situaties optreden: ofwel er wordt medewerking verkregen en de persoon respondeert, ofwel er wordt niet meegewerkt aan het onderzoek wat leidt tot non-respons. Er zijn echter verschillende redenen voor het niet responderen. Personen die niet bereikt kunnen worden tijdens de veldwerkperiode hebben vaak andere redenen om niet mee te doen dan personen die deelname aan het onderzoek weigeren. Dit leidt tot verschillende typen non-respondenten. De belangrijkste typen non-respons zijn geen contact en weigering. In hoofdstuk 2 bespreken we een aantal theorieën voor deze typen non-respons. Daarnaast presenteren we een overzicht van de eigenschappen die samenhangen met geen contact en weigering, zoals de samenstelling van het huishouden of leeftijd. Als gevolg van non-respons kan deze samenhang leiden tot een vertekening van de survey variabelen.

Voor het vergelijken van surveys onderling, over de tijd of tussen verschillende landen, is het belangrijk een onderscheid te maken tussen de diverse typen non-respondenten. Wanneer de samenstelling van de respons verschilt per survey, en hiermee wordt geen rekening gehouden bij de correctie voor non-responsvertekening, dan zal een vergelijking van surveys behalve de verandering in het gemeten verschijnsel ook verschillen in de samenstelling van de respons weergeven. Daarnaast is het een bekend verschijnsel dat de typen respons een verschillende oorzaak en ook een verschillend effect hebben op de uitkomsten van survey onderzoek. De traditionele methoden voor het analyseren van non-respons en de correctie voor non-responsvertekening houden hiermee geen rekening. Daarom hebben we in dit proefschrift methoden ontwikkeld die de

verschillende typen respondenten onderscheiden.

Een manier om met non-respons om te gaan, is door te proberen meer informatie te verkrijgen over de non-respondenten. Dit kan worden gedaan door de steekproef te koppelen aan registers, voor zowel de respondenten als de non-respondenten. Een voorbeeld hiervan wordt gegeven in hoofdstuk 3, waar we laten zien dat we door het onderscheiden van de diverse typen respondenten beter in staat zijn om responsgedrag te verklaren. Een andere manier om meer informatie over non-respondenten te verkrijgen, is het opnieuw benaderen van non-respondenten. In hoofdstuk 4 hebben we de herbenadering van non-respondenten op de Enquête BeroepsBevolking (EBB) beschreven. In juli tot en met oktober 2005 hebben we eenmalig een groot aantal non-respondenten op de EBB opnieuw benaderd met twee methoden: een intensieve herbenadering en de centrale-vraag methode. De resultaten van de herbenadering hebben we bovendien kunnen valideren met behulp van gekoppelde gegevens uit meerdere registers. Dit is ook interessant voor organisaties die zelf geen toegang hebben tot deze registers, zoals marktonderzoeksbureau's en academici. De overgehaalde huishoudens in de intensieve herbenadering bleken te verschillen van de reguliere EBB respondenten, maar vertoonden overeenkomsten met de niet-overgehaalde non-respondenten. Het toevoegen van de overgehaalde huishoudens aan de reguliere EBB respons heeft zodoende de samenstelling van de respons verbeterd. Voor de centrale-vraag methode was dit niet het geval. Dit werd echter grotendeels veroorzaakt door het ontwerp van deze methode, wat heeft geleid tot een verstrengeling van het effect van telefoonbezit en werk-gerelateerde kenmerken.

Een andere manier om non-respons te behandelen is het toepassen van een schattingsmethode voor de correctie van vertekening ten gevolge van non-respons nadat de gegevens zijn verzameld. In hoofdstuk 6 hebben we een overzicht gegeven van de traditionele correctiemethoden gebaseerd op calibratie, evenals meer recente methoden die gebruik maken van geschatte responskansen, zogeheten 'response propensities' of 'propensity scores'. Afhankelijk van het type achtergrondinformatie kunnen verschillende schattingsmethoden worden toegepast. Wanneer de onderzoeker de beschikking heeft over microdata voor alleen de respondenten en daarnaast ook beschikt over de verdeling in de populatie, dan kunnen calibratieschatters zoals de gegeneraliseerde regressieschatter of post-stratificatie toegepast worden. Wanneer de onderzoeker daarnaast ook beschikt over microdata voor de non-respondenten, dan kunnen naast de calibratieschatters ook andere methoden, zoals de propensity score methode, worden toegepast. De besproken methoden in hoofdstuk 6 maken echter geen onderscheid tussen de diverse typen respondenten.

In hoofdstuk 7 bespreken we een toepassing van deze methoden op een tele-

fonische survey om te corrigeren voor non-responsvertekening en vertekening die optreedt als gevolg van de onderdekking van personen met een vaste geregistreerde telefoon. Uit de analyse blijkt dat het weglaten van personen zonder vaste geregistreerde telefoon leidt tot een vertekening van bepaalde survey variabelen, zoals inkomen en opleidingsniveau, waarvoor niet voldoende gecorrigeerd kan worden. Het verschil tussen de methoden is echter niet groot. Daaruit kunnen we concluderen dat niet zozeer de methode, maar de informatie die gebruikt wordt in de methode belangrijk is voor het succes van de correctie.

Het succes van de correctiemethoden is dus sterk afhankelijk van de beschikbare achtergrondinformatie. Om te bepalen welke informatie het beste gebruikt kan worden voor de correctie van non-responsvertekening, is het nuttig het responsgedrag te analyseren. In hoofdstuk 8 hebben we een analyse methode voor responsgedrag ontwikkeld die rekening houdt met de verschillende typen respons. Deze methode bestaat uit afzonderlijke vergelijkingen voor de diverse respons typen. Verder is het ook mogelijk een correlatie tussen de verschillende typen respons te introduceren. Op deze manier kunnen we het responsproces nauwkeurig beschrijven doordat per type respons verschillende verbanden tussen variabelen worden toegestaan. In aanvulling op de responsvergelijkingen, hebben we een extra vergelijking toegevoegd voor het modelleren van de survey variabele. Op deze manier kan de methode ook worden toegepast om te corrigeren voor non-responsvertekening.

De reden voor het onderscheiden van verschillende typen respons komt voort uit de intuïtie dat het veldwerk proces belangrijke informatie bevat voor het verklaren van responsgedrag en uiteindelijk ook voor het corrigeren voor non-responsvertekening. De hoeveelheid informatie die beschikbaar is over het veldwerk verschilt echter per type respons. Waarnemingen van de interviewer zijn bijvoorbeeld alleen beschikbaar voor personen waar contact mee is gemaakt. De methode die we in hoofdstuk 8 ontwikkeld hebben, faciliteert de opname van veldwerkinformatie omdat elk type respons wordt gemodelleerd in een aparte vergelijking. Het kunnen toepassen van deze gedetailleerde methode staat of valt met de beschikbaarheid van achtergrondinformatie. In het beste geval is er gedetailleerde sociaal-economische en demografische informatie beschikbaar voor zowel respondenten als non-respondenten, en daarnaast uitgebreide informatie over het veldwerk proces. Dit is de situatie in Nederland, Scandinavië en sommige andere Noord-Europese landen. Marktonderzoekers en academici beschikken meestal niet over dezelfde gedetailleerde informatie. De methoden die we in hoofdstuk 8 beschrijven, kunnen alleen worden toegepast wanneer de achtergrondinformatie beschikbaar is voor alle personen in de steekproef. In alle andere gevallen moeten de methoden die we in hoofdstuk 6 beschrijven worden

toegepast.

Hoewel de analyse van het responsgedrag en correctie voor vertekening ten gevolge van non-respons pas worden toegepast nadat de gegevens zijn verzameld, wordt hiermee een belangrijk inzicht in non-respons verkregen. Bijvoorbeeld het identificeren van groepen in de respons die onder- of oververtegenwoordigd zijn ten opzichte van de doelpopulatie. Deze informatie kan gebruikt worden in een eerdere fase van het statistisch proces, bijvoorbeeld voor het balanceren van de samenstelling van de respons tijdens het veldwerk. Om te faciliteren dat de informatie op deze manier gebruikt kan worden, hebben we in hoofdstuk 5 een indicator voor de kwaliteit van de respons ontwikkeld: de Representativiteitsindicator of R-indicator. Deze indicator geeft weer hoe evenwichtig de samenstelling van de respons is vergeleken met de steekproef met betrekking tot een verzameling kenmerken die vooraf is vastgesteld. De R-indicator is gebaseerd op geschatte responskansen en vertaalt multivariate informatie over het responsgedrag naar een makkelijk te interpreteren, univarate maat. De R-indicator kan voor verscheidene doeleinden gebruikt worden. Allereerst geeft het een indicatie van de kwaliteit van de respons. Daarnaast kan de R-indicator ook ingezet worden om tijdens het veldwerk de samenstelling van de respons te optimaliseren. Door het concentreren van het veldwerk op groepen die ondervertegenwoordigd zijn in de respons, kan de respons efficiënt verhoogd worden.

De R-indicator kan alleen worden berekend wanneer er achtergrondinformatie beschikbaar is voor zowel respondenten als non-respondenten. Organisaties die niet over deze informatie beschikken, kunnen de R-indicator niet berekenen. Nationaal statistische bureau's publiceren populatie verdelingen voor specifieke domeinen. Deze informatie is vrij beschikbaar. Daarom zou er ook onderzoek gedaan moeten worden naar de constructie van een R-indicator op basis van geaggregeerde informatie over de populatie.

In hoofdstuk 7 hebben we laten zien dat het succes van non-respons correctiemethoden sterk bepaald wordt door de beschikbaarheid van achtergrondinformatie. Deze methoden hebben veel profijt van de toegenomen hoeveelheid beschikbare registers. Aan de andere kant wordt de correctie voor nonresponsvertekening ook uitgedaagd door ontwikkelingen in het ontwerp van surveys, zoals dynamisch survey design. De eerste stap in het dynamisch survey design is de ontwikkeling en implementatie van mixed-mode surveys. Het is tegenwoordig geen uitzondering meer om gegevens te verzamelen met meerdere onderzoeksmodes. Door een efficiënte inzet van verschillende modes kan de dataverzameling goedkoper en sneller worden uitgevoerd, met behoud van kwaliteit. Vanuit het perspectief van de correctie voor non-responsvertekening, rijst de vraag hoe de gegevens verzameld in verschillende onderzoeksmodes gecombineerd kun-

nen worden tot één schatting voor de survey variabele. In hoofdstuk 9 van dit proefschrift hebben we een mixed-mode methode ontwikkeld die de gegevens van verschillende onderzoeksmodes combineert, terwijl het veldwerkproces in elke mode afzonderlijk wordt beschreven. Deze methode volgt het veldwerkproces nauwgezet. Niet alleen door onderscheid te maken tussen verschillende typen respons, maar ook door het beschrijven van de aaneenschakeling en de combinatie van verschillende onderzoeksmodes. Bij een gebrek aan gedetailleerde informatie over het veldwerk, raden we aan tenminste een extra variabele voor de onderzoeksmode op te nemen bij de correctie voor non-responsvertekening. Dit zorgt ervoor dat in ieder geval het algemene mode effect wordt ondervangen.

Door het expliciet modelleren van responsgedrag (hoofdstuk 8) en door het onderscheiden van verschillende onderzoeksmodes in het dataverzamelings proces (hoofdstuk 9) wordt het mogelijk naast de gebruikelijke demografische en sociaal-economische informatie ook informatie over het veldwerkproces op te nemen. De consequentie hiervan is dat er een groot aantal vergelijkingen ontstaan en, als gevolg daarvan, een groot aantal onderliggende aannames wordt gemaakt. Het schatten van beide methoden, correctie voor non-responsvertekening met meerdere responstypen en de mixed-mode methode, vereist de evaluatie van verscheidene integralen van de multivariaat normale verdeling. Toekomstig onderzoek zou gericht moeten zijn op de praktische inzetbaarheid van deze methoden. Allereerst kan onderzocht worden hoe de schattingsprocedure voor deze methoden vereenvoudigd kan worden door de toepassing van bijvoorbeeld Markov Chain Monte Carlo-methoden of Bayesiaanse schattingsmethoden. In de tweede plaats zou toekomstig onderzoek zich moeten richten op de evaluatie van de onderliggende aannames, aangezien de besproken methoden sterk afhankelijk zijn van de gemaakte aannames. Ook het ontwikkelen van alternatieve modellen verdient de aandacht. En als laatste kunnen deze methoden alleen toegepast worden op één survey variabele tegelijk. Als de survey bestaat uit een groot aantal variabelen dan is het niet pragmatisch om voor elke variabele afzonderlijk een correctie uit te voeren. Voor het vergroten van de praktische waarde van de beschreven modellen, zou toekomstig onderzoek moeten uitwijzen hoe we de methoden kunnen uitbreiden voor verschillende survey variabelen.