Adjusting undercoverage and non-response bias in telephone surveys

Discussion paper 05006

Fannie Cobben and Jelke Bethlehem

Statistics Netherlands

Voorburg/Heerlen, September 2005

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2004–2005	= 2004 to 2005 inclusive
2004/2005	= average of 2004 up to and including 2005
2004/'05	= crop year, financial year, school year etc. beginning in 2004 and ending in 2005

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314
Key figure: X-10

Statistics Netherlands

Adjusting undercoverage and nonresponse bias in telephone surveys

Summary: Individuals that are selected in a sample but that do not participate are referred to as nonrespondents. That nonresponse can cause statistics to be biased is a well known fact. In this paper we investigate another non-observation error: the bias due to undercoverage. Undercoverage is the situation in which the sample is not 'representative' of the population of interest because there are certain groups in this population that are not covered by the sampling frame underlying the survey. If the noncovered groups respond systematically different to the questions in the survey, the statistics are biased due to the undercoverage. Undercoverage occurs even before the survey has been allocated to the interviewers and is therefore invisible in most datasets.

In this paper, we analyse the undercoverage bias in telephone surveys. This bias is induced by the restriction of using the telephone as a communication medium. Only that part of the population that has a listed, fixed-line telephone can be interviewed by means of CATI. We use data from a face-to-face survey, the Integrated Survey of Living Conditions, to assess the undercoverage bias when restricting the sample to individuals that own a listed, fixed-line telephone. We compare individuals with and without a listed, fixed-line number. Several methods to adjust for the undercoverage bias are applied to the data from the Integrated Survey of Living Conditions: linear weighting and propensity score weighting and propensity score stratification. Finally, we analyse a strategy to simultaneously adjust for nonresponse and undercoverage bias.

Keywords: undercoverage bias, nonresponse bias, CAPI, CATI, propensity score stratification, linear weighting

1. Introduction

Traditionally, sample surveys used paper questionnaire forms to collect data. The questions on these forms could be completed in different modes. Face-to-face interviewing produced high response rates and high quality data, but it was expensive. A mail survey was much cheaper (no interviewers needed), but the price to be paid was low response rates and low data quality. Somewhere in between (in terms of quality and costs) was telephone interviewing.

The rapid developments of computer technology since the 1970s made it possible to replace paper questionnaire forms by electronic ones. The computer took control of the interviewing process by performing two important tasks: managing the correct route through the questionnaire, and checking the answers to the questions on the spot.

Computer assisted interviewing comes in three forms. It started in the 1970s with Computer Assisted Telephone Interviewing (CATI). More recent is Computer Assisted Personal Interviewing (CAPI). It stands for face-to-face interviewing in which interviewers use a notebook computer to ask the questions. Computer-assisted self-interviewing (CASI) emerged as the electronic form of mail interviewing. Also web surveys can be seen as a form of self-interviewing. A good account of the state of the art in computer assisted interviewing is given by Couper et al. (1998)

Usually, one of the main objectives of a sample survey is to compute estimates of population characteristics. Such estimates will never be exactly equal to the population characteristics; there will always be some error. This error can have many causes; see e.g. Bethlehem (1999). The ultimate result of all these errors is a discrepancy between the survey estimate and the population characteristic to be estimated. This discrepancy is called the *total error*. Two broad categories can be distinguished contributing to this total error: sampling errors and non-sampling errors.

Sampling errors are introduced by the sampling design. They arise due to the fact that estimates are based on a sample and not on a complete enumeration of the population. Sampling errors vanish if the complete population is observed. Since only a sample is available for computing population characteristics, and not the complete data set, one has to rely on estimates.

Non-sampling errors may even occur if the whole population is investigated. They denote errors made during the process of recording answers to the questions. Non-sampling errors can be divided into various types of errors:

- An *overcoverage error* means that elements are included in the survey that do not belong to the target population.
- A *measurement error* occurs if a respondent does not understand a question, or does not want to give the true answer, or if the interviewer makes an error in recording the answer. Also, interview effects, question wording effects, and

memory effects belong to this group of errors. A measurement error causes a difference between the true value and the value processed in the survey.

- *Undercoverage* occurs when elements of the target population do not have a corresponding entry in the sampling frame. So, these elements can never be contacted.
- Another important non-sampling error is *nonresponse*. It is the phenomenon that individuals selected in the sample do not provide the required information because they have not been contacted, are unable to co-operate or refuse to.

In the ideal situation, the individuals participating in a survey are selected by means of a probability sample. If selection probabilities of all elements in the population are known and strictly positive, unbiased estimates can always be computed, see Horvitz and Thompson (1952). However, due to non-sampling errors like undercoverage or nonresponse, actual selection probabilities can be zero (e.g. if an element is not included in the sampling frame, it can never be selected) or unknown (e.g. if selected individuals refuse to co-operate).

Statistics Netherlands faces the challenge to produce high quality statistics under severe budget cuts imposed by the government. One way of realising this is implementing more cost-effective modes of data collection. Statistics Netherlands has always favoured face-to-face interviewing for its social and demographic surveys. Due to the persuasive power and assistance of interviewers visiting selected individuals or households, nonresponse 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. To reduce costs, Statistics Netherlands is now considering changing some of its face-to-face surveys into telephone surveys. 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 face-to-face interviewing to telephone interviewing may substantially reduce the costs of the surveys, there is also a potential drawback: it may reduce the quality of the produced statistics. In this paper we explore the possible effects of changing the mode of data collection in one of the Statistics Netherlands surveys: the Integrated Survey on Living Conditions, denoted by its Dutch acronym POLS; 'Permanent Onderzoek LeefSituatie'.

Up until now, POLS is a face-to-face survey. There is some undercoverage and a substantial amount of nonresponse. A change to a telephone survey will increase the undercoverage error, and in general not reduce the nonresponse. We analyse what the effect is of this change on the bias of estimates of population characteristics. To do so, we artificially construct a telephone survey data file from the face-to-face survey data file. In this way, we avoid possible mode effects caused by differences in face-to-face and telephone interviewing. The records in the face-to-face survey data file are matched to the telephone directory provided by the Dutch telephone company KPN and records without a listed number are deleted to construct the

telephone survey data file. In section 4.3 and 4.4 we investigate whether we can reduce the bias due to undercoverage by using the propensity score method.

Adjustment weighting is usually applied to correct for the negative effect of nonresponse on survey estimates. We explore whether weighting can help to reduce the combined effect of undercoverage and nonresponse in a telephone survey. In section 4.5 and 4.6 we compare traditional weighting methods with the propensity score method.

The outline of this paper is as follows. In section 2 we describe different data collection modes in some more detail. Section 3 presents results from the analysis of mode effects in the 2002 Integrated Survey on Household Living Conditions. Section 4 describes the propensity score method and presents the results of applying this technique to the data. Section 5 discusses the results and proposes topics for further research.

2. Data collection modes

The data collection modes that we compare in this paper are computer assisted telephone interviewing and computer assisted personal interviewing. In this section we discuss differences between these two modes from the perspective of Statistics Netherlands. We restrict ourselves to three aspects: sampling frame, data quality and interviewer-respondent interaction. For a more in-depth overview of data collection technologies, we refer to Nichols *et al.* (1997).

Sampling frame

Statistics Netherlands uses one sampling frame for all its social and demographic surveys: the Municipal Basic Administration (in Dutch: Gemeentelijke Basis Administratie or GBA). Every municipality has its own basic population adminstration. It contains data like date of birth, gender, and home address of all the registered inhabitants of the municipality. For face-to-face surveys, samples are selected from this sampling frame. Coverage is close to 100%. Remaining undercoverage mainly relates to illegal immigrants.

The sampling frame for telephone surveys is obtained by linking telephone numbers to the names/addresses from the Municipal Base Administration. This is achieved by handing over the names/addresses to the Dutch telephone company KPN. However, such links will only be established for individuals with a listed, fixed-line number. As a consequence, individuals with an unlisted fixed-line number, individuals with only a mobile phone, and individuals without a phone, will never be selected for a telephone survey. Currently, the percentage of individuals with a listed, fixed-line number is estimated to be between 60% and 70%. This means that there is a substantial undercoverage of 30% to 40%.

An alternative way of selecting a sample for a telephone survey is to use Random Digit Dialling (RDD). This method randomly generates telephone numbers. Lepkowski (1988) provides a review of RDD and other telephone survey sampling methods. An advantage of RDD is that both listed and unlisted numbers (and possibly mobile numbers as well) can be selected. However, RDD also has a number of disadvantages. It suffers from overcoverage, because it is possible that generated numbers do not exist or not belong to households (but companies). Furthermore, households may have a higher selection probability than anticipated, because they have more than one phone number. Sometimes, it is not clear whether a generated number is not connected (in which case it should be treated as overcoverage) or whether there is no answer due to nonresponse. A final severe disadvantage is that no information at all becomes available about the nonrespondents. Therefore, no effective weighting adjustment procedure can be carried out. Because of all these problems with RDD, Statistics Netherlands has decided not to use this selection technique.

Data quality

De Leeuw (1992) performed a meta-data 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. 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, as is the case at Statistics Netherlands), 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.

Interviewer-respondent interaction

The main difference between a face-to-face and a telephone survey is the way of communication between interviewers and respondents. 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.

3. Empirical research

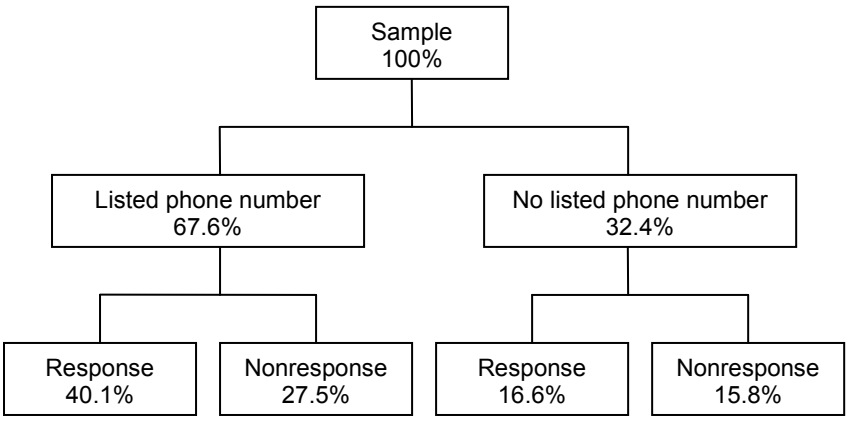
3.1 Introduction

The aim of our research is to investigate a possibly increased bias of estimates when going from a face-to-face survey to a telephone survey. This bias can be caused by both undercoverage and nonresponse. The first step involves an exploratory analysis of the undercoverage. In section 3.2, we compare demographic and social background characteristics of individuals with and without a listed, fixed-line telephone. Literature suggests that the characteristics indeed differ for individuals with and without a telephone, see e.g. Pickery and Carton (2005), Rispens and Van Goor (2000), Kuusela (2000), Ellis and Krosnick (1999) and Smith (1990).

Our main interest lies in the bias in statistics that are ultimately generated from the survey data. This bias is caused both by undercoverage and nonresponse. It is difficult to distinguish the contributions of both types of non-sampling errors in the

total bias. Figure 3.1.1 describes the situation graphically for the POLS 2002 Survey.

Figure 3.1.1: Graphical representation of noncoverage and nonresponse



In the ideal situation where the sampling frame exactly covers the population, the bias will only be caused by nonresponse of both individuals with and without a listed phone number. Face-to-face interviewing is close to this situation. The percentage response is 40.1% + 16.6% = 56.7%.

In case of a telephone 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 phone numbers is much higher (59.4%) than for no listed numbers (51.1%). Apparently, individuals without a listed phone number behave differently from individuals with a listed phone number.

To analyse the implications of the undercoverage bias, the same nonresponse adjustment treatment is applied to the face-to-face survey data and the telephone survey data. For the nonresponse adjustment two different weighting models are used. The models are described in section 3.3, and the results of the analysis are discussed in section 3.4.

Analysis is based on data from the 2002 Integrated Survey on Living Conditions (POLS). See e.g. Bethlehem and Schouten (2004) and Schouten (2004). POLS is a continuous face-to-face survey. Every month a sample of 3000 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 base module is answered by everyone.

The target population is not the same for every module. However, all target populations at least consist of persons of 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 paper, 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.

POLS is a face-to-face survey. For our research we need both a face-to-face survey data file and a telephone survey data file. The telephone survey data file is constructed from the face-to-face survey data file by matching the records in the former data file to the telephone directory provided by the Dutch telephone company KPN and deleting records without a listed number. An advantage of this artificial way of generating a telephone survey data file is that possible mode effects caused by differences in face-to-face and telephone interviewing are avoided.

The face-to-face sample consists of 35594 individuals, 24052 of which have a listed, fixed-line telephone (67,6%). In Table 3.1.1, the number of available records is displayed.

Table 3.1.1 Available data for the both surveys

	Face-to-face survey	Telephone survey
Sample size	35 594	24 052
Number of respondents	20 168	14 275
Percentage of respondents	56.7%	59.4%

3.2 An exploratory analysis of the individuals with a telephone

This section explores differences between individuals with and without a listed phone number. Our primary interest is the possible effect on estimates for the important survey questions, but due to nonresponse it is not possible to determine such effects. Therefore, analysis is restricted to auxiliary variables for which values for both respondents and nonrespondents are known. These variables are listed in Table 3.2.1.

Table 3.2.1: Auxiliary variables used for analysis of relationship with listed phone number

Variable	Categories
Gender	Male, Female
Age in 3 classes	0 – 34, 35 – 54, 55 +
Age in 15 classes	12 – 14, 15 – 17, …, 70 – 74, 74 +
Marital status in 2 classes	Married, Not married
Marital status in 4 classes	Married, Not married, Divorced, Widowed
Ethnic group in 8 classes	Native; Moroccan; Turkish; Surinam; Netherlands Antilles/Aruba; other non-Western non-native; other Western non-native
Province of residence and three largest cities in 15 classes[1]	
Region in 4 classes	North, East, South, West
Degree of urbanization in 5 classes	Very low, Low, Average, High, Very high
Household size in 5 classes	1, 2, 3, 4, >4
Household type in 5 classes	Single, Couple, Couple with children, Single parent, Other
Interviewer month in 12 classes	January, February, …, December
Listed fixed-line telephone in 2 classes	Yes, No
Average house value in 14 classes	Missing; 0; 0 – 50000; 50000 – 75000; …; 275000 – 300000; 300000 – 350000; > 350000 (euro)
% non-natives in 6-digit postcode area in 8 classes	0 – 5%, 5 – 10%, …, 40 – 50%, 50% and more

The total sample consists of 35594 individuals, of which 24052 individuals have a listed phone number and 11542 do not have one. At first sight, it seems obvious to use the Chi-square test to compare the two groups. However, performing a Chi-square test on such a large data set is hardly informative. Every relationship, how small it may be, will turn out to be significant due to the large number of observations. We are not interested in the existence of differences, but in the magnitude of these differences. Therefore, we use the *average relative difference* between the two groups.

The number of individuals with a listed telephone number (24052) exceeds the number of individuals without a listed telephone number (11542). Furthermore, individuals without a listed telephone number cannot participate in a telephone survey. Therefore, we consider not having a listed telephone number as deviant behaviour and take the group of individuals with a listed number as the reference

[1] The Netherlands is divided in 12 provinces. The three largest cities are Rotterdam, The Hague and Utrecht city. Amsterdam is excluded from the analysis because no listed numbers were matched.

group. We compare sample fractions of individuals without a listed telephone number with sample fractions of individuals with a listed telephone number.

Suppose an auxiliary variable has p categories. We introduce p dummy variables. The values for sample element i are denoted by $x_{i,1}$, *(...,)* $x_{i,p}$, where

$$x_{i,j} = \begin{cases} 1, \textit{ if individual i belongs to category j} \\ 0, \textit{ otherwise} \end{cases}$$

Let the sample s of size n consist of a subsample s_L (of size n_L) of elements with a listed telephone number and a sub-sample s_U (of size n_U) of elements without a listed number. Then

$$\bar{x}_j^L = \frac{1}{n_L} \sum_{i \in s_L} x_{i,j} \quad (for\ j = 1,..., p) \tag{3.2.1}$$

are the sampling fractions for persons with a listed number, and likewise

$$\bar{x}_j^U = \frac{1}{n_U} \sum_{i \in s_U} x_{i,j} \quad (for\ j = 1,..., p) \tag{3.2.2}$$

denote the sampling fractions for persons without a listed number. The relative difference for category j of the auxiliary variable is defined by

$$d_j = \frac{\bar{x}_j^U - \bar{x}_j^L}{\bar{x}_j^L}. \tag{3.2.3}$$

The *average relative difference* for the auxiliary variable is now defined by

$$\bar{d} = \frac{1}{p} \sum_{j=1}^{p} \left| d_j \right|, \tag{3.2.4}$$

where any value between 0 and ∞ can be assumed.

As an illustration, consider the auxiliary variable *Gender.* There are two categories, *Male* and *Female,* so $p = 2$. Then, the relative difference for category 1 (*Male) is*:

$$d_1 = \frac{\bar{x}_1^U - \bar{x}_1^L}{\bar{x}_1^L} \tag{3.2.5}$$

A positive value of d_1 would indicate *males* to be underrepresented in the telephone sample. And $d_1 < 0$ implies an overrepresentation of *males* in this survey. Likewise, d_2 measures the over- or underrepresentation of females.

The values of the average relative difference for a number of POLS data are displayed in Table 3.2.2.

Table 3.2.2 Average relative differences between individuals with and without a listed phone number

Variable	Average relative difference
Ethnic group	2,70
Percentage foreign	0,63
Province	0,50
Average house value	0,48
Household type	0,44
Marital status	0,43
Urbanization	0,32
Age	0,19
Household size	0,16
Region	0,15
Interview month	0,04
Gender	0,01

The variable *Ethnic group* shows the largest deviation, followed by *Percentage foreign* and *Household type*. Within *Ethnic group*, the differences are displayed in Table 3.2.3.

Table 3.2.3. Relative difference within Ethnic group

Variable	Relative difference
Ethnic group	*2,70*
Native	0,1
Moroccan	6,9
Turkish	3,2
Surinam	3,5
Netherlands Antilles/Aruba	2,8
Other non-Western non-native	1,9
Other Western non-native	0,3

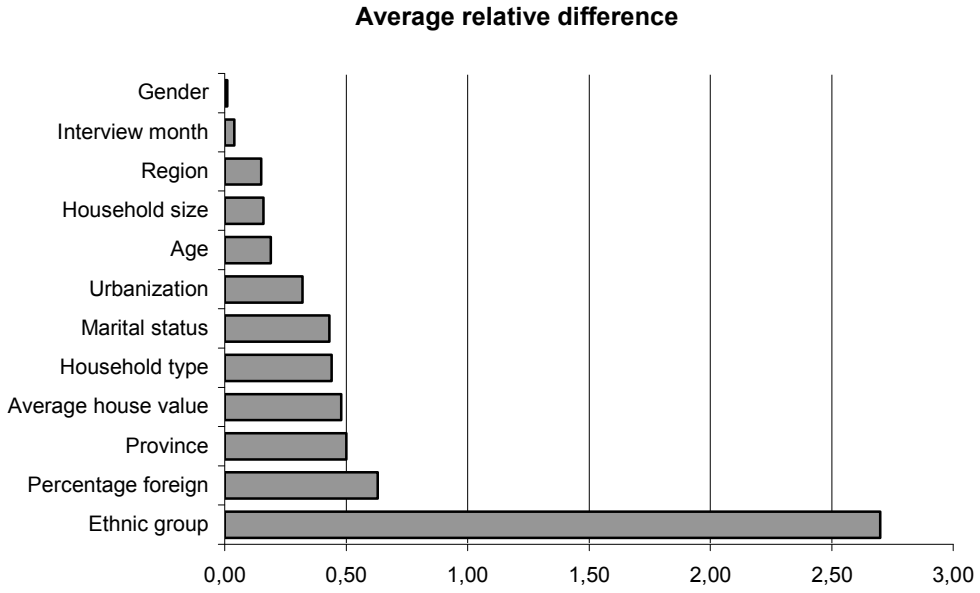
This indicates that, with respect to these variables, the telephone sample is not a 'representative' sample from the population. Heavily underrepresented are:

- Non-Western, non-native individuals (especially Moroccan, Surinam and Turkish people),
- Regions where the percentage of non-natives is higher than 20%.

Some other findings: Individuals of age 55 and older are slightly over-represented, as are persons from the North and the South (also visible in the variable *Province*). Furthermore, divorced persons are underrepresented, as well as persons who live in the city or areas with a high degree of urbanization.

Figure 3.2.1 shows the average relative difference in a graphical way. It becomes once more clear that the differences between individuals with a listed and an unlisted number is for a large part determined by variables related to ethnic background.

Figure 3.2.1. Graphical representation of the average relative difference



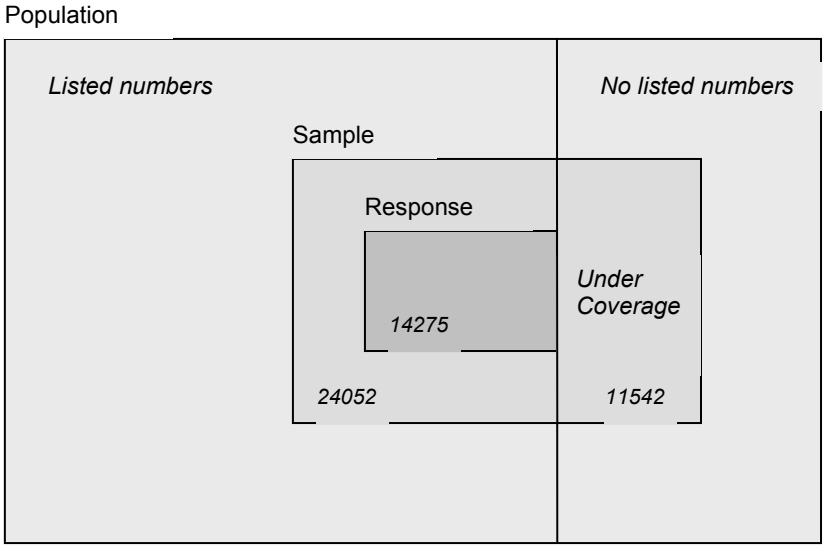
These results are consistent with the results in similar analyses for other countries. Pickery and Carton (2005) analyse the representativeness of telephone surveys in Dutch-speaking Belgium. They find the same differences in ethnic group, household type and characteristics of housing, and they report differences related to the educational level, the employment situation and ownership of a house. Rispens and Van Goor (2000) 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 in the degree of urbanization and in the size and type of the household. For the USA, Smith (1990) developed a multivariate model to predict household telephone ownership based on the General Social Survey conducted by the University of Chicago. The analysis indicated that the socio-economic factor in general (occupational prestige, educational level, employment situation) and income in particular has the largest impact. Ellis and Krosnick (1999) report differences with respect to education, income and ethnicity.

The most consistent finding in these studies is the difference in employment situation and educational level. These variables are the most important questions in our household surveys. For those survey questions that are related to income and education, we expect to find large biases in the statistics based on telephone surveys.

3.3 Linear weighting

Estimates computed from the telephone survey data may be biased due to undercoverage and nonresponse. Because they are both non-observational errors, it is hard to distinguish between them. In this section, we apply a weighting adjustment approach to correct for nonresponse bias and compare the estimates computed from the telephone sample with those computed from the full sample. This is done to see what the influence is of not taking into account the undercoverage bias. The situation is as described in Figure 3.3.1.

Figure 3.3.1. Weighting the telephone survey



The POLS sample (of size 35594) is selected from the total population using the population register. Of the sampled individuals 24052 has a listed phone number. Of those only 14275 respond. A weighting technique is carried out in which the response is weighted to the complete sample, i.e. the weighted distribution of the auxiliary variables in the response is made equal to their unweighted sample distribution.

In line with earlier research performed by Schouten (2004) on POLS 1998, six important target variables are selected from the POLS 2002 base module:

- Employment in three classes: employed for 12 hours or more per week, employed less than 12 hours per week, or unemployed.
- Owner of a house: yes or no.
- At least one activity per month in a club (sports, music, etc.): yes or no.
- Owner of a PC or laptop: yes or no.

- Highest educational level successfully completed: primary, junior general secondary, pre-vocational, senior general secondary / pre-university, secondary vocational, higher professional, university, or other.
- Religion: none, Roman-Catholic, Protestant, Islamic, and other.

The weighting technique used is based on the generalized regression estimator; see also Bethlehem (1987).

Let the target population of a sample survey consist of N individuals 1, 2,…, N. Let Y denote a target variable of the survey. Associated with each individual k is a value Y_k of this target variable. Assume that the aim of the sample survey is to estimate the population mean of the target variable

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k \ . \tag{3.3.1}$$

Furthermore, let X be a vector of auxiliary variables or covariates, with values X_k, for $k = 1, 2, \ldots, N$.

A sample selected without replacement from the population can be represented by an N-vector $s = (s_1, s_2, \ldots, s_N)'$ of indicators, where $s_k = 1$ if individual k is selected in the sample, and where $s_k = 0$ otherwise. The expected value of s is equal to $E(s) = \pi$, where $\pi = (\pi_1, \pi_2, \ldots, \pi_N)'$ is the N-vector of first order inclusion probabilities. The sample size is denoted by n. Horvitz and Thompson (1952) show that

$$\bar{y}_{HT} = \frac{1}{N}\sum_{k=1}^{N} \frac{s_k Y_k}{\pi_k} \tag{3.3.2}$$

is an unbiased estimator of the population mean (3.3.1).

Linear weighting amounts to applying the generalized regression estimator. This estimator is defined by

$$\bar{y}_{GR} = \bar{y}_{HT} + (\bar{X} - \bar{x}_{HT})'b, \tag{3.3.3}$$

where $\bar{X}$ is the vector of population means of a set of auxiliary variables, $\bar{x}_{HT}$ is the vector of Horvitz-Thompson estimators for the auxiliary variables, and b is a vector of regression coefficients defined by

$$b = \left(\sum_{k=1}^{N} \frac{s_k X_k X_k'}{\pi_k}\right)^{-1} \left(\sum_{k=1}^{N} \frac{s_k X_k Y_k'}{\pi_k}\right). \tag{3.3.4}$$

In the case of nonresponse, the Horvitz-Thompson estimators $\bar{y}_{HT}$ and $\bar{x}_{HT}$ cannot be used. Furthermore, estimation of b will have to be based on available observations only. We then use the modified generalized regression estimator, see Bethlehem (1988). The assumption underlying this estimator is that the data are *Missing At Random (MAR)*. This is the case if target variable and response behaviour are not independent, but they are independent given the values of the

auxiliary variables. Without use of auxiliary information, estimates of the population characteristics would be biased. For qualitative auxiliary variables, MAR means that observed values of the target variable not necessarily constitute a random subsample of the sampled values, but they are a random subsample of the sampled values within subclasses defined by the values of the auxiliary variables. This implies that there are auxiliary variables conditionally on which the nonresponse is not selective for the target variable.

Two weighting models are tested. The first one is the model currently used in the POLS surveys. It can be denoted by

$$\begin{aligned} &(Gender_2 \times Age_3 \times Marital\ status_2) + (Gender_2 \times Age_{15}) + \\ &+ (Region_4 \times Age_3) + Urbanization_5 + Region_{15} + Household\ size_5 + \\ &+ Marital\ status_4 + Interview\ month \end{aligned} \tag{1}$$

Subscripts denote the number of categories of the variable, see Table 3.2.1. The model is selected on the basis of the availability of auxiliary variables with a known population distribution. Effectively, the model consists of seven different variables: age, gender, marital status, degree of urbanization, region, household size and interview month. There are three interaction-terms, all including the variable *Age*. A different number of categories is used in different terms to avoid the risk of empty strata.

The second weighting model is proposed by Schouten (2004). This model is the result of a selection strategy of auxiliary variables that simultaneously accounts for the relation with response behavior and the relation with survey questions. A remarkable finding of this research is that estimates for target variables based on weighting models including interaction terms of auxiliary variables differ at most 0,1% from estimates based on weighting models that contain only main effects. This means that the number of parameters in the weighting model can be substantially reduced without affecting the outcomes and hence more auxiliary variables can be included. The resulting weighting model is

$$\begin{aligned} &Age_{15} + House\ value_{14} + Telephone_2 + Percentage\ foreign_8 + \\ &+ Ethnic\ group_7 + Region_{15} + Household\ type_4 \end{aligned} \tag{2}$$

Again, subscripts denote the number of categories. Note that only the variables Age_{15} and $Region_{15}$ are shared with weighting model (1).

3.4 Results of linear weighting

The software package Bascula is used to compute weights for both weighting models. Weighting is applied to both the telephone- and the face-to-face survey data file. The Bascula package is developed by Statistics Netherlands. It offers various weighting methods: poststratification, ratio estimation, linear weighting based on the general regression estimator and multiplicative weighting based on iterative proportional fitting. For more information, see Nieuwenbroek and Boonstra (2001) and Bethlehem (1998).

The results for the face-to-face survey can be found in Table 3.4.1 and for the telephone survey in Table 3.4.2. For variables with two categories (yes/no) only the mean for the answer 'yes' is given. Note that, when applying model (2) to the telephone survey, the variable *telephone* is redundant and thus removed. For reasons of comparison, we decide to also remove this variable when applying model (2) to the face-to-face survey.

Table 3.4.1. Unweighted and weighted estimates of sample distributions (percentages) based on face-to-face survey data.

Variable	Response mean	Weight model 1	Weight model 2
Employment			
12 hours or more	55.3	56.7	56.1
Less than 12 hours	7.1	6.6	6.7
Unemployed	37.6	36.6	37.3
Owner of a house	66.3	64.8	63.6
Active in club	46.1	45.3	44.8
Owner of a PC	75.8	75.4	74.8
Educational level			
Primary	7.2	6.0	6.1
Junior general secondary	12.0	11.8	11.9
Pre-vocational	19.7	19.6	19.9
Senior general secondary	7.2	7.4	7.4
Secondary vocational	30.7	31.1	31.1
Higher professional	16.6	17.2	16.9
University	6.4	6.8	6.6
Other	0.2	0.2	0.2
Religion			
None	37.7	38.8	38.4
Roman-Catholic	35.6	32.8	32.5
Protestant	21.4	20.6	20.5
Islamic	2.6	2.6	3.3
Other	5.2	5.3	5.4

To see how the various adjustment techniques perform, we use a simple measure: the average relative difference. Its definition is given in (3.2.4), but here the average is taken over the average relative difference for all the variables in the table. The reference group is the result of linear weighting of the full sample with weight model (2). The average relative difference per variable is given in parentheses.

Table 3.4.2. Unweighted and weighted estimates of sample distributions (percentages) based on telephone survey data.

Variable	Response mean	Weight model 1	Weight model 2
Employment	*(0.051)*	*(0.047)*	*(0.022)*
12 hours or more	54.5	58.0	56.7
Less than 12 hours	7.4	7.0	6.9
Unemployed	38.1	35.0	36.4
Owner of a house	*(0.113)*	*(0.074)*	*(0.019)*
Yes	70.8	68.3	64.8
Active in club	*(0.080)*	*(0.049)*	*(0.022)*
Yes	48.4	47.0	45.8
Owner of a PC	*(0.028)*	*(0.035)*	*(0.013)*
Yes	76.9	77.4	75.8
Educational level	*(0.031)*	*(0.048)*	*(0.023)*
Primary	6.7	5.6	5.8
Junior general secondary	12.2	11.6	11.9
Pre-vocational	19.5	18.5	19.3
Senior general secondary	6.9	7.2	7.2
Secondary vocational	31.0	31.4	31.4
Higher professional	17.2	18.3	17.6
University	6.5	7.2	6.8
Other	0.2	0.2	0.2
Religion	*(0.211)*	*(0.166)*	*(0.020)*
None	36.2	39.0	37.8
Roman-Catholic	35.4	33.2	32.8
Protestant	22.7	21.7	21.0
Islamic	1.2	1.3	3.2
Other	4.5	4.7	5.3
Average relative difference	*0.086*	*0.070*	*0.020*

A great degree of similarity of telephone estimates and face-to-face survey estimates means that going from a face-to-face survey to a telephone survey does not introduce a bias.

In general, the pattern observed for the estimates is quite similar for both surveys. Regarding the two-category variables two conclusions can be drawn: weighting models (1) and (2) both adjust in the same direction and the adjustments by model (2) are larger than those by model (1) (except for the variable *PC*). Considering variables with more categories, the adjustments show a different pattern. The adjustments by model (1) and model (2) follow the same direction, but this time model (2) adjusts less than model (1). The largest reduction in relative difference is made for the variable *Religion*. Model (1) reduces the difference from 0.211 to 0.166; model (2) brings the difference down to 0.020. For two variables, *Educational level* and *Owner of a PC*, the telephone response mean displays a smaller difference than the estimates from weight model (1).

By applying weighting model (1), the average relative difference reduces only from 0.086 to 0.070, but for weighting model (2) the reduction is larger: from 0.086 to 0.020. Both weighting models seem to be able to adjust for nonresponse bias in the face-to-face survey. However, estimates based on telephone survey data remain

biased. We believe that this bias is caused by undercoverage. Apparently, these weighting models are not able to reduce an undercoverage bias.

Weighting model (2) seems to adjust better than model (1). However some bias still remains. This is remarkable since model (2) incorporates the variables that cause the largest selectivity (*ethnic group*$_7$, *percentage foreign*$_8$ and *region*$_{15}$). The results suggest that answers to the survey questions for persons with a telephone on average are different than those for persons without a telephone, even after taking into account the variables that we dispose of that are correlated with ownership of a telephone. In the next section we explore an alternative adjustment technique.

4. Propensity score method

4.1 Theoretical framework

Again, it is assumed that the aim of the sample survey is to estimate the population mean of the target variable (see also section 3.3)

$$\overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k \ . \tag{4.1.1}$$

In case of a telephone survey, only people with a listed number can be contacted. We assume that whether or not an individual has a listed number is the result of some random process, where each individual k has a certain, unknown probability τ_k of having a listed number, for k = 1, 2, …, N. Let T denote an indicator variable, where T_k = 1 if individual k has a listed, fixed-line telephone number, and where T_k = 0 otherwise. Then $P(T_k = 1) = \tau_k$.

For a telephone survey, only those values Y_k become available for which individual k is selected in the sample (s_k = 1) and has a listed phone number (T_k = 1). Therefore, the adapted Horvitz-Thompson estimator becomes

$$\overline{y}'_{HT} = \frac{1}{N}\sum_{k=1}^{N} \frac{s_k T_k Y_k}{\pi_k \tau_k} . \tag{4.1.2}$$

Unfortunately, the listed phone probabilities are unknown. Therefore, we estimate them using the available auxiliary information. We use the method of propensity scores to achieve this. See Rosenbaum and Rubin (1983).

Translated in the current context, the propensity score $\tau(X)$ is the conditional probability that an individual with observed characteristics X has a listed telephone number (T = 1):

$$\tau(X) = P(T = 1 \mid X) \tag{4.1.3}$$

It is assumed that within subpopulations defined by values of the observed characteristics X, all individuals have the same probability of having a listed number. In the context of nonresponse in survey sampling, this assumption is also

referred to as Missing At Random (MAR). Both linear weighting and the propensity score method rely on this assumption.

4.2 The propensity score model

Often, the propensity score is modelled by means of a logit model:

$$\log\left(\frac{\tau(X_k)}{1-\tau(X_k)}\right) = \alpha_k + \beta_k' X_k + \varepsilon_k \qquad (4.2.1)$$

Other models can be used too, but e.g. Dehija and Wahba (1999) conclude that different models often produce similar results.

Using the complete POLS 2002 data set, the propensity scores are modelled with the software package Stata. By stepwise excluding insignificant variables, the final model turns out to be

$$\begin{aligned}
&\alpha_k + \beta_1 * Percentage\ foreigners_{8,k} + \beta_2 * \text{Region}_{15,k} + \beta_3 * Ethnic\ group_{7,k} + \\
&+ \beta_4 * Urbanization_{5,k} + \beta_5 * Marital\ \ status_{2,k} + \beta_6 * Household\ \ type_{4,k} + \\
&+ \beta_7 * House\,value_{14,k} + \beta_8 * Age_{3,k} + \beta_9 * Disability\,insurance_{2,\mathrm{k}} + \\
&+ \beta_{10} * \text{Social security}_{2,\mathrm{k}} + \varepsilon_k
\end{aligned}$$

The first subscript of each variable denotes the number of categories. We estimate the model parameters by Maximum Likelihood Estimation. The value of the pseudo R^2 for this model turns out to be 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 propensity scores can be predicted. These predicted scores can be used in various ways. The following techniques are explored:

- *Propensity score weighting*: the listed number probabilities τ_k in the adapted Horvitz-Thompson estimator (4.1.2) are replaced by their estimates $\tau(X_k)$ from the logit model.
- *Propensity score stratification*: This is a form of post-stratification where strata are being formed on the basis of the propensity scores.
- *Linear weighting with adjusted inclusion probabilities*: This is a form of linear weighting in which the inclusion probabilities π_k in (3.3.4) are replaced by $\pi_k \tau(X_k)$.
- *Linear weighting with propensity score strata*. This is a form of linear weighting in which one of the auxiliary variables denotes strata which are formed by grouping the propensity scores into a number of categories.

The first two approaches are used to adjust the telephone sample for undercoverage whereas the last two approaches are employed to simultaneously adjust for undercoverage and nonresponse.

All approaches are explored in the following subsections. To see how the techniques perform, the results from propensity score weighting and –stratification are compared to the response mean from the full sample. The results from linear weighting with adjusted inclusion probabilities resp. propensity score strata are compared to the results of linear weighting of the full sample with weight model (2) from section 3. Since the full sample data are used in combination with the best possible weight model, these estimates should be closest to the true value.

4.3 Propensity score weighting

Propensity score weighting means that the listed number probabilities τ_k in the adapted Horvitz-Thompson estimator (4.1.2) are replaced by their estimates $\tau(X_k)$ from a logit model.

Table 4.3.1 contains the results of applying this adjustment technique. The second column contains the unadjusted response mean of the telephone survey data. The third column contains the unadjusted response mean of the full sample. These estimates are used as a reference in determining the effect of the applied adjustment technique. In parentheses, the average difference is given for each target variable separately.

Propensity score weighting does not seem to perform very well. Many values of estimates shift in opposite directions when compared to the response mean of the full sample. The results for the variable *Religion* are remarkable, the average relative difference increases from 0.156 to 0.232. Over all, the average relative difference increases compared to the unadjusted response mean from the telephone survey (from 0.056 to 0.083). This might be caused by the fact that estimates become highly dependent on the model used for the propensity scores.

Table 4.3.1. Estimates for the telephone survey data based on propensity score weighting

Variable	Response mean telephone survey	Response mean full sample	Propensity score weighting
Employment	*(0.023)*		*(0.034)*
12 hours or more	54.5	55.3	53.7
Less than 12 hours	7.4	7.1	7.4
Unemployed	38.1	37.6	38.8
Owner of a house	*(0.068)*		*(0.107)*
Yes	70.8	66.3	73.4
Active in club	*(0.050)*		*(0.082)*
Yes	48.4	46.1	49.9
Owner of a PC	*(0.015)*		*(0.016)*
Yes	76.9	75.8	77.0
Educational level	*(0.025)*		*(0.028)*
Primary	6.7	7.2	6.5
Junior general secondary	12.2	12.0	12.2
Pre-vocational	19.5	19.7	19.8
Senior general secondary	6.9	7.2	6.8
Secondary vocational	31.0	30.7	31.1
Higher professional	17.2	16.6	17.2
University	6.5	6.4	6.4
Other	0.2	0.2	0.2
Religion	*(0.156)*		*(0.232)*
None	36.2	37.7	35.0
Roman-Catholic	35.4	35.6	36.6
Protestant	22.7	21.4	23.5
Islamic	1.2	2.6	0.6
Other	4.5	5.2	4.2
Average relative difference	*0.056*	*0.000*	*0.083*

.

4.4 Propensity score stratification

Propensity score stratification is a form of post-stratification where strata are being formed on the basis of the propensity scores.

Suppose the sample is stratified into L strata by means of the estimated propensity score. The poststratification estimator is defined by

$$\bar{y}_{PS} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h \tag{4.2.3}$$

where N_h is the number of elements in stratum h and $\bar{y}_h$ is the mean of the available observations in stratum h. Bethlehem (1988) shows that the bias of the post-stratified Horvitz-Thompson estimator can be written as

$$B(\bar{y}_{PS}) = \sum_{h=1}^{L} W_h \frac{C_h(\tau, Y)}{\bar{\tau}_h}, \tag{4.2.4}$$

where summation takes place of the L strata. The quantity $W_h = N_h / N$ is the relative size of stratum h, $C_h(\tau, Y)$ is the covariance between the values of target variable and

listed number probabilities within stratum h, and $\bar{\tau}_h$ is the average of the listed number probabilities in stratum h.

This bias is small if the covariance is small in each stratum and the covariance is small when the variation in listed number probabilities is small. So it makes sense to construct strata in such a way that most variation of these probabilities is between strata and not within strata. Cochran (1968) suggests that as much as five strata may be sufficient to remove a large part of the bias. We apply propensity score stratification to our dataset using 5 and 25 propensity score strata.

Following Cochran, two categorical variables are constructed that divide the propensity scores in 5 resp. 25 categories. Categories correspond to equal width intervals of the propensity scores. The distributions of these two variables in the full sample are displayed in Figure 4.4.1.

Figure 4.4.1 Propensity score in 5 resp. 25 categories

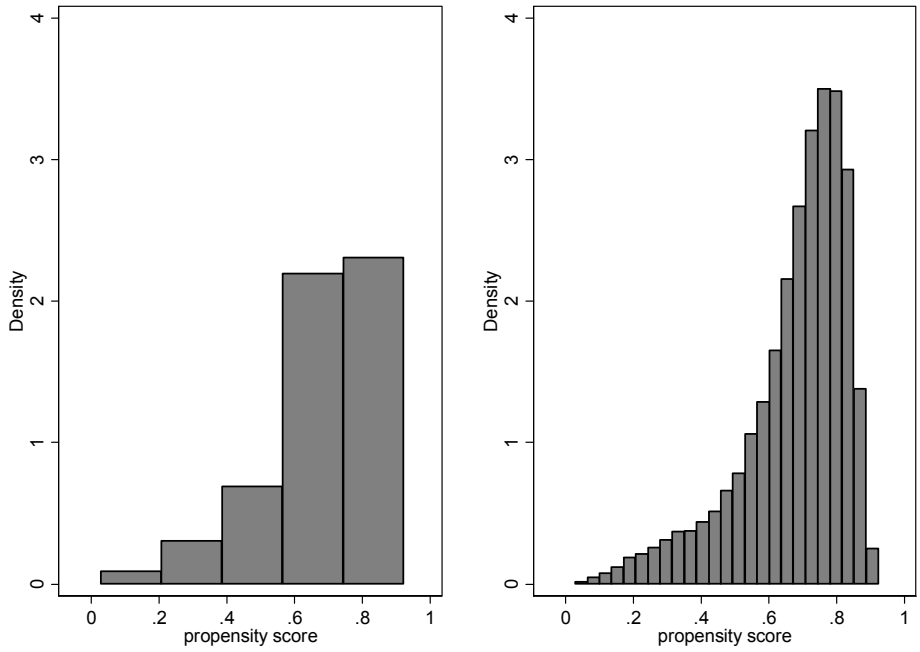


Table 4.4.1 contains the results obtained by just using one of these variables for weighting. This comes down to poststratification with either 5 or 25 strata.

Table 4.4.1. Estimates for the telephone survey data based on propensity score stratification

Variable	Response mean telephone survey	Response mean full sample	Propensity score stratification (5 strata)	Propensity score stratification (25 strata)
Employment	*(0.023)*		*(0.017)*	*(0.017)*
12 hours or more	54.5	55.3	54.9	54.9
Less than 12 hours	7.4	7.1	7.4	7.4
Unemployed	38.1	37.6	37.7	37.7
Owner of a house	*(0.068)*		*(0.017)*	*(0.009)*
Yes	70.8	66.3	67.4	66.9
Active in club	*(0.050)*		*(0.020)*	*(0.017)*
Yes	48.4	46.1	47.0	46.9
Owner of a PC	*(0.015)*		*(0.008)*	*(0.008)*
Yes	76.9	75.8	76.4	76.4
Educational level	*(0.025)*		*(0.019)*	*(0.022)*
Primary	6.7	7.2	7.1	7.0
Junior general secondary	12.2	12.0	12.2	12.3
Pre-vocational	19.5	19.7	19.1	19.1
Senior general secondary	6.9	7.2	6.9	7.0
Secondary vocational	31.0	30.7	30.9	30.8
Higher professional	17.2	16.6	17.1	17.1
University	6.5	6.4	6.5	6.6
Other	0.2	0.2	0.2	0.2
Religion	*(0.156)*		*(0.029)*	*(0.027)*
None	36.2	37.7	36.9	37.2
Roman-Catholic	35.4	35.6	33.8	33.6
Protestant	22.7	21.4	21.7	21.6
Islamic	1.2	2.6	2.5	2.5
Other	4.5	5.2	5.1	5.1
Average relative difference	*0.056*	*0.000*	*0.018*	*0.017*

By comparing the results after stratification with those for the unadjusted response mean in the full sample, it can be observed that this approach leads to estimates that are shifted in the right direction. The largest deviation after adjustment is found for the variable *Religion*, which displays an average relative difference of 0.156 based on the telephone response mean and a difference of 0.027 after adjustment (based on 25 strata). However, estimates are often still far away from the full sample response mean. So, stratification based on just propensity scores is not able to completely correct for the undercoverage bias.

Note also that there are no big differences between the estimates for 5 and 25 strata. Those for 25 strata are slightly better. This is not surprising, because the strata will be more homogeneous with respect to the values of the propensity scores. In the next sections, the two techniques to simultaneously adjust for nonresponse and undercoverage bias are explored.

4.5 Linear weighting with adjusted inclusion probabilities

Linear weighting as described by expression (3.3.3) in section 3.3 only produces consistent estimates if the proper inclusion probabilities are used. Availability of

data in a telephone survey is determined by both the sampling mechanism and the probability of having a listed number. Therefore, the in π_k in (3.3.4) should be replaced by $\pi_k \tau_k$. Unfortunately, the τ_k are unknown, so they are replaced by their estimates $\tau(X_k)$.

Table 4.5.1 contains the results of this approach when applied to the telephone survey data, where weight model (2) is used.

Table 4.5.1. Estimates for the telephone survey data based on linear weighting with adjusted inclusion probabilities

Variable	Response mean telephone survey	Weight model 2 (full sample)	Linear weighting with adjusted inclusion probabilities
Employment	*(0.051)*		*(0.020)*
12 hours or more	54.5	56.1	56.2
Less than 12 hours	7.4	6.7	7.0
Unemployed	38.1	37.3	36.8
Owner of a house	*(0.113)*		*(0.017)*
Yes	70.8	63.6	64.7
Active in club	*(0.080)*		*(0.018)*
Yes	48.4	44.8	45.6
Owner of a PC	*(0.028)*		*(0.011)*
Yes	76.9	74.8	75.6
Educational level	*(0.031)*		*(0.023)*
Primary	6.7	6.1	5.8
Junior general secondary	12.2	11.9	11.9
Pre-vocational	19.5	19.9	19.2
Senior general secondary	6.9	7.4	7.2
Secondary vocational	31.0	31.1	31.2
Higher professional	17.2	16.9	17.5
University	6.5	6.6	6.8
Other	0.2	0.2	0.2
Religion	*(0.211)*		*(0.009)*
None	36.2	38.4	37.8
Roman-Catholic	35.4	32.5	32.7
Protestant	22.7	20.5	21.0
Islamic	1.2	3.3	3.3
Other	4.5	5.4	5.4
Average relative difference	*0.086*	*0.000*	*0.016*

It is clear from Table 4.5.1 that the estimates are closer to their reference values. Compared to the propensity score methods (weighting and stratification) to adjust for the undercoverage, the largest gain seems to be made for the variable *Religion*. With propensity score weighting, the average relative difference increases compared to the telephone response mean. With propensity score stratification, the average relative difference at best reduces to 0.027. For linear weighting with adjusted inclusion probabilities, the average relative difference is further reduced to 0.009. The variables that display the largest deviation are *Employment* and *Educational*

level. We already expected this from the literature mentioned above; since these are the indicators that best predict telephone ownership.

Note that this approach produces estimates that are also better than those of the linear weighting model (2) in Table 3.4.2. This could be expected, because in section 3.4 the wrong inclusion probabilities have been used.

4.6 Linear weighting including propensity score strata

A final approach explored is to take linear weighting model (2) as a starting model and not adjust the inclusion probabilities, but instead include a categorical propensity score variable in the weighting model. So, the weighting model becomes

$$Age_{15} + House\ value_{14} + Telephone_{2} + Percentage\ foreign_{8} +$$

$$+ Ethnic\ group_{7} + Region_{15} + Household\ type_{4} + Propensity\ score_{5}.$$

Not only a propensity score variable with 5 categories is tested, but also one with 25 categories. The results can be found in Table 4.6.1.

Table 4.6.1. Estimates for the telephone survey data based on linear weighting including propensity score strata.

Variable	Response mean telephone survey	Weight model 2 (full sample)	Linear weighting with 5 propensity score strata	Linear weighting with 25 propensity score strata
Employment	*(0.051)*		*(0.021)*	*(0.020)*
12 hours or more	54.5	56.1	56.3	56.2
Less than 12 hours	7.4	6.7	7.0	7.0
Unemployed	38.1	37.3	36.7	36.8
Owner of a house	*(0.113)*		*(0.016)*	*(0.017)*
Yes	70.8	63.6	64.6	64.7
Active in club	*(0.080)*		*(0.020)*	*(0.018)*
Yes	48.4	44.8	45.7	45.6
Owner of a PC	*(0.028)*		*(0.012)*	*(0.011)*
Yes	76.9	74.8	75.7	75.6
Educational level	*(0.031)*		*(0.024)*	*(0.023)*
Primary	6.7	6.1	5.8	5.8
Junior general secondary	12.2	11.9	11.9	11.9
Pre-vocational	19.5	19.9	19.2	19.1
Senior general secondary	6.9	7.4	7.2	7.2
Secondary vocational	31.0	31.1	31.3	31.2
Higher professional	17.2	16.9	17.6	17.5
University	6.5	6.6	6.8	6.8
Other	0.2	0.2	0.2	0.2
Religion	*(0.211)*		*(0.018)*	*(0.009)*
None	36.2	38.4	37.8	37.8
Roman-Catholic	35.4	32.5	32.7	32.7
Protestant	22.7	20.5	20.9	21.0
Islamic	1.2	3.3	3.2	3.3
Other	4.5	5.4	5.3	5.4
Average relative difference	*0.086*	*0.000*	*0.019*	*0.016*

The weight model with 25 strata performs better than the model with 5 strata. Again, this is no surprise. When we look at the average relative difference per variable, we see the same development as with linear weighting with adjusted inclusion probabilities. The variable with the largest reduction in average difference is *Religion*. The difference reduces from 0.211 to 0.018 for the model with 5 strata; from 0.211 to 0.009 for the model with 25 strata. The variables that display the largest deviation after the adjustment again are *Educational level* and *Employment*.

If we compare the results of these two weighting models with weighting model (2) from section 3.4, then it becomes clear that including a categorical propensity score variable in the model pays. For the 5-category variable, the average relative difference reduces from 0.020 to 0.019. And for the 25-category variable, this quantity reduces from 0.020 to 0.016.

5. Discussion

In this paper 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 linear weighting, and partly on using propensity scores.

Two data collection modes are analysed: face-to-face interviewing and telephone interviewing. It turns out that there are substantial differences between persons with and without a listed phone number. Particularly, these two groups differ on variables *Ethnic group*, *Percentage of non-natives* and *Region*. Especially non-Western, non-native individuals and regions where the percentage of non-natives is higher than 20% are heavily underrepresented in telephone surveys.

It is explored to what extent adjustment techniques can reduce the bias caused by telephone interviewing. First, the telephone sample is adjusted for undercoverage of persons without a listed number. Two methods are used: Propensity score weighting and propensity score stratification. No nonresponse bias is considered yet and the results are compared to the unadjusted response mean of the full sample. Second, the nonresponse bias is taken into account as well. Two combinations of linear weighting and the propensity score method are considered. The results are compared to the estimates from applying weight model (2) to the full sample. Since the full sample data are used in combination with the best possible weight model, these estimates should be closest to the true value.

Table 5.1.1 and Table 5.1.2 summarise the results of the various correction approaches that are explored in sections 3 and 4.

Table 5.1.1. Overview of results to adjust for undercoverage; reference group is the unadjusted response mean from the full sample

Adjustment technique	Average relative difference
Response mean telephone survey	0.056
Propensity score weighting	0.083
Propensity score stratification, 5 strata	0.018
Propensity score stratification, 25 strata	0.017

Table 5.1.2. Overview of results to adjust the telephone survey; reference group is the result from applying weight model (2) to the full sample

Adjustment technique	Average relative difference
Response mean telephone survey	0.086
Weight model (1)	0.070
Weight model (2)	0.020
Linear weighting with adjusted inclusion probabilities	0.016
Linear weighting including 5 propensity score strata	0.019
Linear weighting including 25 propensity score strata	0.016

As can be seen from Table 5.1.1, the best adjustment technique to adjust for undercoverage is propensity score stratification with 25 strata. But an average relative difference of 0.017 still remains. The variable that contributes most to this difference is *Religion.*

The best adjustment technique to reduce the bias caused by telephone interviewing for this case, appears to be a combination of linear weighting in which the true inclusion probabilities are estimated by means of logit model for listed number propensity scores. Then the average relative difference reduces to 0.016. One can also choose to incorporate a categorical version of the propensity score in the model for linear weighting. If this variable has 25 categories, the average relative difference also reduces to 0.016. The variables that display the largest deviation after adjustment are *Educational level* and *Employment*. We already expected this, since the literature mentions these variables to be different for persons with and without a telephone.

There are several issues that need further research. The assumptions that are made with respect to the independence between telephone ownership and survey answers given the covariates should be explored more profoundly. Both the generalised regression estimator and the propensity score method rely heavily on this assumption. Also, we use the average relative difference to judge the models. Other measures can be used as well. Furthermore, the logit model that is used to predict the propensity scores still has a proportion of unexplained variance of over 90%. There seems to be a dependency between the covariates and the ownership of a telephone that is unexplained for after controlling for auxiliary variables. This might be the reason that we are not fully able to adjust for the undercoverage bias. There is

a relationship between telephone ownership and the questions in the survey that we cannot explain with the auxiliary variables that we dispose of. 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. It is rather ironic; the variables that we try to estimate are actually the variables that we would like to use in our model. Therefore, it is no surprise that the ultimate estimates are still biased.

References

Bethlehem, J. (1988) 'Reduction of nonresponse bias through regression estimation' *Journal of Official Statistics*, Vol. 4, No. 3, pp. 251 – 260

Bethlehem, J. (1998) 'Bascula 4.0 for adjustment weighting: Tutorial' Statistics Netherlands, Voorburg

Bethlehem, J. (2002) 'Weighting nonresponse adjustments based on auxiliary information' In: R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little (eds.), *Survey Nonresponse*, Wiley, New York

Bethlehem, J. and Schouten, B. (2003) 'Nonresponse analysis of the integrated survey on living conditions (POLS)', Discussion paper 04004, Statistics Netherlands, Voorburg

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1983) 'Some uses of statistical models in connection with the nonresponse problem' In: Madow, W.G. and Olkin, I., (eds.): *Incomplete Data in Sample Surveys*, Vol. 3: Proceedings of the symposium, Academic Press, New York

De Leeuw, E.D. (1992) 'Data Quality in Mail, Telephone, and Face to Face Surveys', Amsterdam: T.T.-publikaties

Dehjia, R. and Wahba, S. (1999) 'Causal effects in non-experimental studies: re-evaluating the evaluation of training programs' *Journal of the American Statistical Association*, 94, pp. 1053 – 1062

Duncan, K.B and Stasny, E.A. (2001) 'Using propensity scores to control noncoverage bias in telephone surveys' *Survey Methodology* 27, pp. 212 – 130

Ellis, C.H. and Krosnick, J.A. (1999) 'Comparing telephone and face-to-face surveys in terms of sample representativeness: a meta-analysis of demographic characteristics' *NES Technical Report Series*, No. nes010871

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

Horvitz, D.G and Thompson, D.J. (1952) 'A generalization of sampling without replacement from a finite universe' *Journal of the American Statistical Association*, 47, p. 663 – 685

Kalsbeek, W.D. (1980) 'A conceptual review of survey error due to nonresponse' American Statistical Association, proceedings of the section on survey research methods, pp. 131 – 136

Kalton, G. and Flores-Cervantes, I. (2003) 'Weighting models' *Journal of Official Statistics*, Vol. 19, No. 2, pp. 81 – 97

Keeter, S. (1995) 'Estimating telephone noncoverage bias with a telephone survey' *Public Opinion Quarterly*, Vol. 59, pp. 196 – 217

Kuusela V. (2000) 'Telephone coverage in Finland' Statistics Finland, International Blaise Users Group. Available at:

http://www.blaiseusers.org/newsletters/bug11/fintel.htm [Accessed on 15 April 2005]

Lepkowski, J.M. (1988) 'Telephone sampling methods in the United States' In: R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, J. Waksberg (eds.), *Telephone Survey Methodology*, Wiley, New York

Massey, J.T. (1988) 'An overview of telephone coverage' In: R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, J. Waksberg (eds.), *Telephone Survey Methodology*, Wiley, New York

Nieuwenbroek, N. and Boonstra, H.J. (2001) 'Bascula 4.0 reference manual' Statistics Netherlands, BPA-nr 3554-99-RSM

Pickery, J. and Carton, A. (2005) 'Hoe representatief zijn telefonische surveys in Vlaanderen?' Administratie Planning en Statistiek, Ministerie van de Vlaamse Gemeenschap, Nota 4

Rispens, S. and Van Goor, H. (2000) 'Een vertekend beeld van de maatschappij?Een onderzoek naar selectieve uitval in een telefonische enquête ten gevolge van onderdekking en nonrespons' *Sociale Gids* 2000/06, jaargang XLVII, pp. 448 – 474

Rocco, E., Salvati, N. and Pratesi, M. (2004) 'Participation in CATI-surveys: traditional nonresponse adjustment versus propensity score matching in reducing the nonresponse bias' E-proceedings of the European Conference in Quality and Methodology in Official Statistics (Q2004), Mainz, Germany

Rosenbaum, P.R. and Rubin, D.B. (1983) 'The central role of the propensity score in observational studies for causal effects' *Biometrika*, 70, pp. 41 – 50

Rosenbaum, P.R. and Rubin, D.B. (1984) 'Reducing bias in observational studies using subclassification on the propensity score' *Journal of the American Statistical Association*, 79, pp. 516 – 524

Schonlau, M. et al. (2004) 'A comparison between responses from a propensity-weighted web survey and an identical RDD survey' *Social science computer review*, Vol. 22, no. 1, pp. 128 – 138

Schouten, B. (2004) 'Adjustment bias in the integrated survey on living conditions (POLS) 1998' Discussion paper 04001, Statistics Netherlands, Voorburg

Smith, T. W. (1990) 'Phone home? An analysis of household telephone ownership' *International Journal of Public Opinion Research*, Vol. 2 No. 4, pp. 369 – 390

Wun, L.M. et al. 'Using propensity scores to adjust weights to compensate for dwelling unit level nonresponse in the medical expenditure panel survey' Agency for Healthcare Research and Quality