Can We Reach You by Telephone?

Fannie Cobben (Statistics Netherlands)

1. Introduction

Blaise supports various data collection modes. Statistics Netherlands has always favoured computer assisted face-to-face interviewing (CAPI) 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 CAPI surveys into computer assisted telephone surveys (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.

The sampling frame for a telephone survey is obtained by linking telephone numbers to the names/addresses from the Municipal Base Admininstration. 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, indviduals 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%.

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'.

2. Survey errors

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).

3. Analysis

3.1. Introduction

The aim of our research is to answer the following question:

What is the effect on the quality of estimates of population characteristics, when going from a face-to-face survey to a telephone survey and, consequently, c*an we adjust for this effect?*

We thereby focus on undercoverage and nonresponse bias. In section 3.2 we introduce the datasets that we use for our analysis. The methods that we apply to adjust for undercoverage- and nonresponse bias are explained in section 3.3. How we apply these methods to our data in order to answer our questions is explicated in section 3.4.

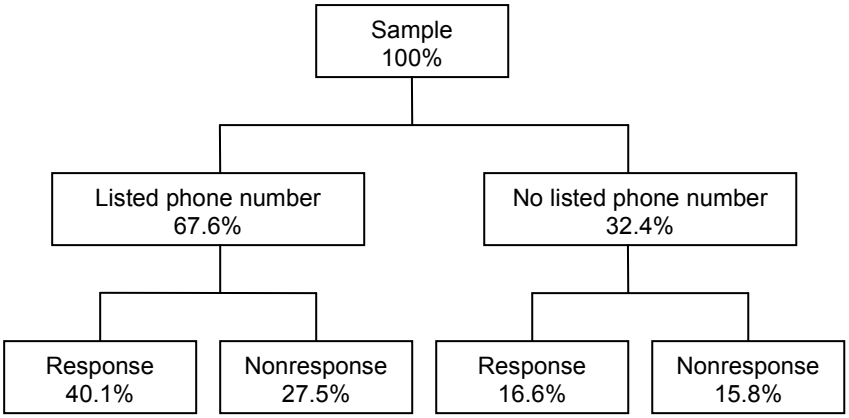
3.2. The data

POLS is a continuous face-to-face 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 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.

It is difficult to distinguish the contributions of both types of non-sampling errors in the total bias. Figure 3.2.1. describes the situation graphically for the POLS 2002 survey.

Figure 3.2.1: Graphical representation of undercoverage 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.

As we already said, 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. We construct the telephone survey data file from the face-to-face survey data file. We can do so by matching the records in the face-to-face data file to the telephone directory provided by the Dutch telephone company KPN. Deleting records without a listed, fixed-line telephone number provides us with the data file that would have been used had the survey been performed by telephone interviewing.

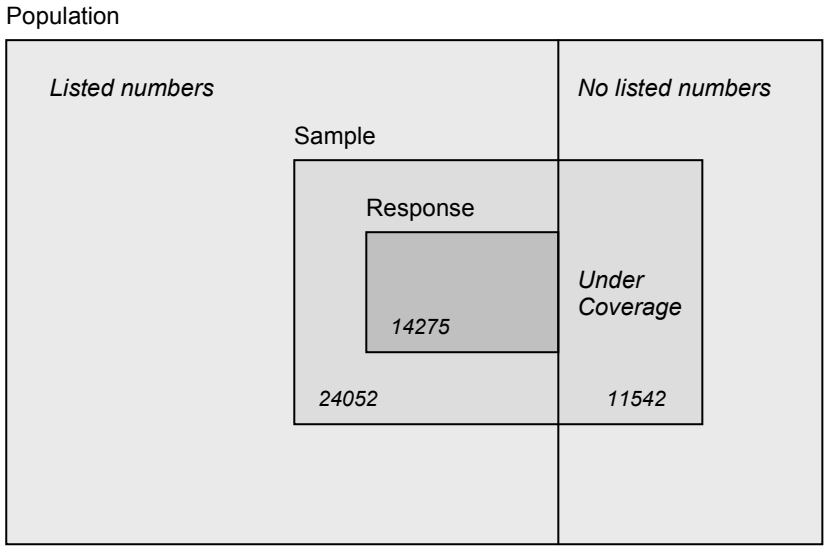
An advantage of this artificial way of generating a telephone survey data file is that possible mode effects caused by differences in face-to-face and telephone interviewing are avoided. The face-to-face sample consists of 35594 individuals, 24052 of which have a listed, fixed-line telephone (67,6%). In Table 3.2.1, the number of available records is displayed.

Table 3.2.1 Available data for 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%

Figure 3.2.1 graphically displays the population and the two datasets.

Figure 3.2.1 Graphical representation of the division of the population with respect to telephone ownership



There are two kinds of variables that we use in the analysis: *background variables* and *target variables*. Background variables are available for both respondents and nonrespondents. These variables come from registers like the Municipal Basic Administration (GBA) and the Centre for Work and Income (CWI). The target variables 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 3.2.2.

Table 3.2.2. Background- and target variables used in the analysis

Background 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

[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.

Background variable	*Categories*
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
Target variable	***Categories***
Employment	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

3.3. The methods

Linear weighting

The linear weighting technique is based on the generalized regression estimator; see also Bethlehem (1988).

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

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

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

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

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

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

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

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

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.

The propensity score

Before we come to the actual propensity scores, first some context. Again, it is assumed that the aim of the sample survey is to estimate the population mean of the target variable *Y,* see (3.3.1). 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 $k = 1, 2, \ldots, N$.
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

$$\bar{y}_{HT}' = \frac{1}{N} \sum_{k=1}^{N} \frac{s_k T_k Y_k}{\pi_k \tau_k}. \qquad (3.3.5)$$

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 ($T = 1$):

$$\tau(X) = P(T = 1 \mid X) \qquad (3.3.6)$$

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). The propensity score method thus relies on the same assumption as linear weighting.

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

$$\log\left(\frac{\tau(X_k)}{1-\tau(X_k)}\right) = \alpha_k + \beta_k' X_k + \varepsilon_k \qquad (3.3.7)$$

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

The propensity score can be used in two different ways: propensity score weighting and propensity score stratification. With *propensity score weighting*, the estimated propensity scores $\hat{\tau}(X_k)$ are directly used in the adapted Horvitz-Thompson estimator (3.3.5). *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

$$\bar{y}_{PS} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h \qquad (3.3.8)$$

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

$$B(\bar{y}_{PS}) = \sum_{h=1}^{L} W_h \frac{C_h(\tau, Y)}{\bar{\tau}_h}, \qquad (3.3.9)$$

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.

Linear weighting combined with the propensity score

There are two ways to combine linear weighting and the propensity score method. The first approach is based on the idea of propensity score weighting and comes down to linear weighting with adjusted inclusion probabilities.
See that linear weighting as described by expression (3.3.3) 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 $\hat{\tau}(X_k)$.

The second approach is based on propensity score stratification. The idea is to include an additional variable in the weighting model that accounts for the probability of having a listed number. This additional variable is a categorical variable with each category representing a stratum based on the estimated propensity scores.

3.4. The data and the methods

So, we have a face-to-face survey that, in the nearby future, could become a telephone survey. Besides the missing data due to nonresponse, we would then also have missing data because of the undercoverage of non-listed, fixed line numbers. Like with nonresponse, undercoverage is only a threath to survey statistics if it is selective. Cobben and Bethlehem (2005) show that this is indeed so for POLS 2002. The characteristics of owners of a listed, fixed line number differ from those without such a number, especially with respect to the variables *ethnic group, % of foreigners* and *household type.*

To obtain insight into the consequences of this additional missing data, we apply the methods discussed in section 3.3 to our data sets. First, the response means of bot the face-to-face survey and the telephone survey are compared. The method that is commonly applied to adjust for nonresponse is linear weighting. For the POLS survey, we apply a weighting model that is proposed by Schouten (2004). This model is[2]

$$Age_{15} + House\ value_{14} + Percentage\ foreign_{8} + Ethnic\ group_{7} + Region_{15} + Household\ type_{4} \quad (1)$$

The subscripts denote the number of categories. When applying this model to both surveys, we do not take into account the selectiveness of the telephone survey yet. Therefore, we apply the propensity score method to the telephone survey to adjust for the propbability of telephone ownership. The combination of linear weighting with the propensity score methods is then applied to adjust for nonresponse and undercoverage simultaneaously. The results are discussed in section 4.

4. Results

To analyse the implications of changing a face-to-face survey into a telephone survey, in Table 4.1 we compare the response means and the adjusted estimates when applying the nonresponse adjustment; weighting model (1). 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 an additional bias.

The second and third column in Table 4.1 display the unweighted response means for both surveys. There are quite some differences. Respondents that own a listed, fixed-line telephone tend to work less, have a higher education and are more often non-religious than respondents that do not own a listed, fixed-line telephone. However, it maybe that adjusting the for nonresponse handles the selectivity of telephone ownership and reduces these differences. Therefore, we apply weighting model (1) to both surveys. The results can be found Table 4.1 in columns four and five.

We regard the results from weighting the face-to-face survey to be the most unbiased, i.e. the most close to the true population characteristics. When we compare the estimates from the weighted telephone survey to these final estimates, we see that some bias still remains. The weighted estimates tend to adjust in the right direction, but then over- or underestimate the characteristic. Take for instance the percentage of persons that work 12 hours or more. The response mean for the

[2] Originally, the variable *telephone* was included in the model. However, when applying the model 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 the model to the face-to-face survey.

telephone survey is 54.5%. The adjusted face-to-face survey estimate is 56.1%. The adjusted telephone estimate indeed shows an increase but overestimates the face-to-face survey estimate by 0.6%. The differences may seem minor, but 0.6% of 16 million persons still are approximately 100.000 persons.

Table 4.1. Comparison of unweighted and weighted response means for the face-to-face survey and the telephone survey (in %)

Variable	Response mean telephone survey	Response mean face-to-face survey	Weighted estimates telephone survey	Weighted estimates face-to-face survey
Employment				
12 hours or more	54.5	55.3	56.7	56.1
Less than 12 hours	7.4	7.1	6.9	6.7
Unemployed	38.1	37.6	36.4	37.3
Educational level				
Primary	6.7	7.2	5.8	6.1
Junior general secondary	12.2	12.0	11.9	11.9
Pre-vocational	19.5	19.7	19.3	19.9
Senior general secondary	6.9	7.2	7.2	7.4
Secondary vocational	31.0	30.7	31.4	31.1
Higher professional	17.2	16.6	17.6	16.9
University	6.5	6.4	6.8	6.6
Other	0.2	0.2	0.2	0.2
Religion				
None	36.2	37.7	37.8	38.4
Roman-Catholic	35.4	35.6	32.8	32.5
Protestant	22.7	21.4	21.0	20.5
Islamic	1.2	2.6	3.2	3.3
Other	4.5	5.2	5.3	5.4

These differences are remarkable since model (1) 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.

These results show that we can not just ignore the change of data collection mode from CAPI to CATI. Therefore, we tried to adjust for the selectiveness of the persons that own a listed, fixed-line number by using the propensity score method discussed in section 3.3.

We model the propensity score by means of a logit model:

$$\log\left(\frac{\tau(X_k)}{1-\tau(X_k)}\right) = \alpha_k + \beta_k' X_k + \varepsilon_k \qquad (4.1)$$

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

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

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 (3.3.5) 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. Cochran (1968) suggests that as much as five strata may be sufficient to remove a large part of the bias.

The results from these two techniques are displayed in Table 4.2.

Table 4.2. Estimates for the telephone survey data based on propensity score weighting and –stratification.

Variable	Response mean telephone survey	Response mean full sample	Propensity score weighting	Propensity score stratification
Employment				
12 hours or more	54.5	55.3	53.7	54.9
Less than 12 hours	7.4	7.1	7.4	7.4
Unemployed	38.1	37.6	38.8	37.7
Educational level				
Primary	6.7	7.2	6.5	7.1
Junior general secondary	12.2	12.0	12.2	12.2
Pre-vocational	19.5	19.7	19.8	19.1
Senior general secondary	6.9	7.2	6.8	6.9
Secondary vocational	31.0	30.7	31.1	30.9
Higher professional	17.2	16.6	17.2	17.1
University	6.5	6.4	6.4	6.5
Other	0.2	0.2	0.2	0.2
Religion				
None	36.2	37.7	35.0	36.9
Roman-Catholic	35.4	35.6	36.6	33.8
Protestant	22.7	21.4	23.5	21.7
Islamic	1.2	2.6	0.6	2.5
Other	4.5	5.2	4.2	5.1

To see how the techniques perform, the results from propensity score weighting and –stratification are compared to the response mean from the full sample. This technique is, after all, meant to adjust for the undercoverage and not for nonresponse. Column two and three display the response means for the telephone resp. face-to-face survey. In column four and five, the results from propensity score weighting resp. –stratifciation are shown.

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 remarkably bad, adjusting most categories in the wrong direction. For instance, the percentage of non-religious persons in the response to the full sample is 37.7%, compared to a

response mean of 36.2% for the telephone survey. In stead of adjusting the underestimation, propensity score weighting increases the difference between the two response means with an additional 1.2%. This might be caused by the fact that estimates become highly dependent on the model used for the propensity scores.

Propensity score stratification leads to estimates that are shifted in the right direction. 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.

So far, we established that propensity score stratification succeeds best in adjusting the response mean for the telephone survey to the response means for the full sample. Now we can proceed with adjusting the final estimates, i.e. including an adjustment for the nonresponse bias as well. We do so by a mixture of linear weighting (nonresponse adjustment) and the propensity method (adjustment for undercoverage). In section 3.3 we discuss two ways to combine these two methods. The results are displayed in Table 4.3.

Table 4.3. Results for the adjustment of the final estimates for both undercoverage and nonresponse.

Variable	Linear weighting the face-to-face sample	Linear weighting with 5 propensity score strata	Linear weighting with adjusted inclusion probabilities
Employment			
12 hours or more	56.1	56.3	56.2
Less than 12 hours	6.7	7.0	7.0
Unemployed	37.3	36.7	36.8
Educational level			
Primary	6.1	5.8	5.8
Junior general secondary	11.9	11.9	11.9
Pre-vocational	19.9	19.2	19.2
Senior general secondary	7.4	7.2	7.2
Secondary vocational	31.1	31.3	31.2
Higher professional	16.9	17.6	17.5
University	6.6	6.8	6.8
Other	0.2	0.2	0.2
Religion			
None	38.4	37.8	37.8
Roman-Catholic	32.5	32.7	32.7
Protestant	20.5	20.9	21.0
Islamic	3.3	3.2	3.3
Other	5.4	5.3	5.4

The results are compared to the estimates from linear weighting the full sample, in the second column, since these estimates should be closest to the true value.
The third and fourth column display the results from adjusting the telephone sample for nonresponse and undercoverage simultaneously with resp. linear weighting with an additional propensity score stratification and linear weighting with adjusted inclusion probabilities.

Including a categorical propensity score variable in the model seems to pay. 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.

5. Conclusion

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.

The aim of our research is to answer the following question:

What is the effect on the quality of estimates of population characteristics, when going from a face-to-face survey to a telephone survey and, consequently, c*an we adjust for this effect?*

We explore 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 unweighted 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 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. However, the ultimate estimates are still biased. There seem tp be 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.

6. References

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

Bethlehem, J. (1999) '*Cross-sectional Research'* In: H.J. Adèr and G.J. Mellenbergh, Research Methodology in the Social, Behaviuoural & Life Science, Sage Publications, London, p.110-142.

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

.

Cobben, F. and Bethlehem, J. (2005) '*Adjusting undercoverage and nonresponse bias in telephone surveys'* Discussion paper 05006, Statistics Netherlands.

Cochran, W. (1968) '*The effectiveness of adjustment by subclassification in removing bias in observational studies*' Biometrics, Vol. 24, p. 205 – 213.

Dehjia, R. and Wahba, S. (1999) *'Causal effects in non-experimental studies: re-evluating the evaluation of training programs'* JASA, p. 1053 – 1062.

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.

Rosenbaum, P.R. and Rubin, D.B. (1983) 'The central role of the propensity score in observational studies for causal effects' Biometrika, 70, p. 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 (387), p. 516-524.

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