Component of Statistics Canada Catalogue no. 11-522-X
Statistics Canada's International Symposium Series: Proceedings

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

Indicators for the Representativeness of Survey Response

by Jelke Bethlehem, Fannie Cobben, and Barry Schouten

2009

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2008
Data Collection: Challenges, Achievements and New Directions

Indicators for the Representativeness of Survey Response

Jelke Bethlehem, Fannie Cobben, and Barry Schouten[1]

Abstract

Many survey organizations use the response rate as an indicator for the quality of survey data. As a consequence, a variety of measures are implemented to reduce non-response or to maintain response at an acceptable level. However, the response rate is not necessarily a good indicator of non-response bias. A higher response rate does not imply smaller non-response bias. What matters is how the composition of the response differs from the composition of the sample as a whole. This paper describes the concept of R-indicators to assess potential differences between the sample and the response. Such indicators may facilitate analysis of survey response over time, between various fieldwork strategies or data collection modes. Some practical examples are given.

Key Words: Non-response, Missing data, Representativity, Survey quality, Indicators.

1. Introduction

1.1 The need for survey quality indicators

Most surveys suffer from non-response. This is the phenomenon that sample elements do not provide the required information. Non-response may seriously affect the quality of the outcomes of a survey. Estimates of population characteristics will be biased if, due to non-response, some groups in the population are over- or underrepresented, and these groups behave differently with respect to the survey variables.

Survey agencies often use the survey response rate as an indicator of survey quality. However, a low response rate does not necessarily imply that the accuracy of survey estimates is poor. If non-response is ignorable, i.e. there is no correlation between response behaviour and the survey variables, estimates will still be unbiased. Indeed, the literature on survey methodology contains ample examples showing that response rates by themselves are poor indicators of non-response bias, see e.g. Curtin, Presser and Singer (2000), Groves, Presser and Dipko (2004), Groves (2006), Groves and Peytcheva (2006), Keeter et al. (2000), Merkle and Edelman (2002), Heerwegh et al. (2007) and Schouten (2004).

Focus on just the response rate as an indicator of survey quality can be misleading. This is illustrated by an example from the 1998 Dutch POLS survey (short for Permanent Onderzoek Leefsituatie or Integrated Survey on Household Living Conditions in English). Table 1.1-1 contains estimates of two population quantities: the percentage of people receiving some form of social allowance and the percentage of people having at least one parent that was born outside the Netherlands. Both variables are taken from a register and are artificially treated as survey questions. Therefore sample percentages are also available. These sample percentages are given in Table 1.1-1. After one month of fieldwork the response rate was 47.2%, while after the full two month period the rate had increased to 59.7%. The mode of data collection in the first month was CAPI (Computer Assisted Personal Interviewing). Non-respondents were approached in the second month with CATI (Computer Assisted Telephone Interviewing) if they had a listed, land-line phone. Otherwise, CAPI was used again. The second month of fieldwork increased the

[1]Jelke Bethlehem, Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands (jbtm@cbs.nl); Fannie Cobben, Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands (fcbn@cbs.nl); Barry Schouten, Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands (bstn@cbs.nl)

response by 12.5%. However, this did not result in better estimates. The bias of the estimators increased after the second month.

Table 1.1-1:
Response means in POLS after the first and second month of data collection.

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

There is a need for additional survey quality indicators that provide more insight in the possible risk of biased estimators. This paper describes such indicators. They are called *R-indicators*. The R stands for 'representativity'. R-indicators measure how representative the survey response is, or to say it differently, how the composition of the response differs from that of the sample.

R-indicators can be used in many different ways. One way is to inspect the survey data after completion of the fieldwork. But they can also play an important role during data collection. By monitoring the fieldwork, data collection efforts can be targeted at obtaining a response the composition of which does not deviate too much from that of the complete sample (or the population). R-indicators can be useful both for social survey and economic surveys. R-indicators can not only be applied in survey data collection, but also to establish the quality of register data.

It is the objective of the RISQ project to develop and to test R-indicators. RISQ stands for Representativity Indicators for Survey Quality. Five partners participate in this project: Statistics Netherlands, Statistics Norway, The Statistical office of Slovenia, the University of Southampton (UK) and the University of Leuven (Belgium). The RISQ project is financed by the 7th Framework Programme of the European Union. More information can be found on www.r-indicator.eu.

1.2 The concept of representativity

The concept of representativity is often used in survey research, but usually it is not clear what it means. Kruskal and Mosteller (1979a, 1979b and 1979c) present an extensive overview of what representative is supposed to mean in non-scientific literature, scientific literature excluding statistics and in the statistical literature. They found the following meanings for 'representative sampling': (1) general acclaim for data, (2) absence of selective forces, (3) miniature of the population, (4) typical or ideal case(s), (5) coverage of the population, (6) a vague term, to be made precise, (7) representative sampling as a specific sampling method, (8) as permitting good estimation, or (9) good enough for a particular purpose. They recommended not using the word *representative*, but instead to specify what one means.

To be able to define an indicator for representativity, the concept of representativity is defined here as the absence of selective forces. Suppose a probability sample s of size n is selected without replacement from a finite population U of size N. The sample can be seen as a vector of N indicators $s = (s_1, s_2, \ldots, s_N)$, where the indicator $s_k = 1$ if element k is selected in the sample, and where $s_k = 0$ otherwise (for k = 1, 2, …, N).

The phenomenon of non-response is modelled by introducing response probabilities. Every element k in the population is assumed to have a certain, unknown, probability ρ_k of responding when selected in the sample. The response to the survey can be represented by the vector of indicators $r = (r_1, r_2, \ldots, r_N)$, where $r_k = 1$ if element k was selected in the sample ($s_k = 1$) and did respond. Otherwise, $r_k = 0$. It follows that $\rho_k = P(r_k = 1 \mid s_k = 1)$.

It is clear that there are no selective forces if all response probabilities are equal. This observation forms the basis of the first definition of representativity.

Definition 1.2.1
The response to a survey is called *strongly representative* with respect to the sample if the response probabilities of all elements in the population are equal and if the response of an element is independent of the response of all other elements. In other words:

$$\rho_k = P(r_k = 1 \mid s_k = 1) = \rho, \text{ for } k = 1, 2, \ldots, N. \tag{1.2.1}$$

Note that strong representativity corresponds to a missing data mechanism that is called *Missing Completely at Random* (MCAR) for every target variable *Y*. It means that non-response does not cause estimators to be biased. This is an appealing definition, but it is not very useful since in practice it is not possible to compare individual response probabilities. Therefore, it cannot be established whether the survey response is strongly representative. To solve this problem a weaker definition of representativity is introduced.

Suppose there is a categorical auxiliary variable X having L categories. It divides the population into L strata (sub-populations). The number of elements in stratum h is denoted by N_h, for $h = 1, 2, \ldots, L$. It is assumed that this variable has been measured in the survey and that its value is available for each respondent and non-respondent. The response probability of element k in stratum h is defined by ρ_{hk}.

Definition 1.2.2
The response to a survey is called *weakly representative* with respect to the sample for auxiliary variable X if the average response probability is the same in each stratum, i.e.

$$\overline{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho, \tag{1.2.2}$$

for $h = 1, 2, \ldots, L$.

Weak representativity means that it is not possible to distinguish respondents from non-respondents just using information with respect to X. If the response is weakly representative with respect to many auxiliary variables X, and there exist strong relationships between the target variables and these auxiliary variables, biases of estimates will be small. It is possible to estimate the means of the response probabilities in the strata, and therefore the assumption of weak representativity can be checked in practice.

2. R-indicators

2.1 The case of known response probabilities

R-indicators measure how far the composition of the response to a survey deviates from the original sample. If all response probabilities are equal, the response is strongly representative, and there will be no systematic differences between the composition of the response and the sample. If the response probabilities are not equal, it is important to establish to what extent the composition of the response is affected. This is accomplished by defining a distance function that measures how far the individual response probabilities differ from the mean response probability.

Suppose, for the time being, that the individual response probabilities $\rho_1, \rho_2, \ldots, \rho_N$ of all elements in the population are known. Then the standard deviation

$$S(\rho) = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (\rho_k - \overline{\rho})^2}, \tag{2.1.1}$$

of the response probabilities can be computed. This is a distance function. $S(\rho) = 0$ if all response probabilities are equal, and the value of $S(\rho)$ will be larger as there is more variation in the values of the response probabilities. It is not difficult to prove that the maximum value of $S(\rho)$ is equal to 0.5. An R-indicator can now be defined as

$$R(\rho) = 1 - 2S(\rho) \tag{2.1.2}$$

This R-indicator assumes values in the interval [0, 1]. A value of 1 implies strong representativity. The smaller its value is, the more the response composition deviates from that the sample composition.

2.2 The case of unknown response probabilities

The values of the individual response probabilities are unknown in practice. This problem is solved by estimating the response probabilities. This can be accomplished if proper auxiliary information is available, i.e. variables that have been measured for both respondents and non-respondents. Several techniques can be used, for example logistic or probit models (see Agresti, 2002) or CHAID classification trees (see, Kass, 1980). Suppose $\hat{\rho}_1, \hat{\rho}_2, ..., \hat{\rho}_n$ are the estimated response probabilities of the sample elements. Then the mean response probability can be estimated by

$$\hat{\bar{\rho}} = \frac{1}{N} \sum_{i=1}^{n} \frac{\hat{\rho}_i}{\pi_i}, \tag{2.2.1}$$

where π_i is the first order inclusion probability of sample element *i*. The R-indicator (2.1.2) can now be estimated by

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1} \sum_{i=1}^{n} \frac{(\hat{\rho}_k - \hat{\bar{\rho}})^2}{\pi_i}}\ . \tag{2.2.2}$$

Note that expression (2.2.2) involves two estimation steps. The first one is estimation of the response probabilities and the second one is estimation of the standard deviation. The consequences of this are discussed in the final section.

2.3 An example

From July to December 2005 Statistics Netherlands conducted a large scale follow-up among the non-respondents in the Dutch Labour Force Survey (LFS). In the study two samples of non-respondents in the LFS were approached once more using either a call-back approach (Hansen and Hurwitz, 1946) with the full LFS questionnaire or a basic-question approach (Kersten and Bethlehem, 1984) with a very short questionnaire containing only a few basic questions. Some results are summarized in Table 2.3-1. For more details see Schouten (2007) and Cobben and Schouten (2007). CAPI was used in the call-back approach, and the basic-question approach used a mixed-mode design which involved web, paper and CATI. The R-indicators were estimated using logistic regression models that included a large number of explanatory variables that measured demographic, geographic and socio-economic characteristics of the households. The final column of Table 2.3-1 contains 95% confidence intervals for the R-indicator that have been computed using a bootstrap technique.

Table 2.3-1
R-indicators for the LFS, LFS including call-back, and LFS including basic-question approach.

Response	Response rate	R-indicator	Confidence interval
LFS	62.2%	0.80	(0.78 ; 0.83)
LFS + call-back	76.9%	0.85	(0.82 ; 0.88)
LFS + basic-question	75.6%	0.78	(0.76 ; 0.80)

The value of the R-indicator for the initial LFS response is equal to 0.80, which is lower than the ideal value of 1.00. So this response is not completely representative. Application of the call-back approach increases the response rate from 62.2% to 76.9%. The value of the R-indicator also increased, from 0.80 to 0.85. The confidence intervals (almost) do not overlap. This indicates that the additional response improves the composition of the data set.

Application of the basic-question approach results in a different conclusion. Although the response rate increases from 62.2% to 75.6%, the value of the R-indicator decreases from 0.80 to 0.78. The intervals for the initial LFS and the LFS including basic-question approach overlap. Apparently, the basic-question approach does not improve the composition of the data set. This approach gives 'more of the same' and, hence, sharpens the contrast between respondents and non-respondents.

3. Properties of R-indicators

3.1 Dependence on auxiliary variables and sample size

Estimated response probabilities are used to compute the R-indicator. Estimation of these probabilities is based on a logistic regression model using a set of auxiliary variables as explanatory variables. This implies that the R-indicator measures the deviation from weak representativity and not from strong representativity. Indeed, this approach is not capable to detect and quantify differences between individual response probabilities within the classes obtained by crossing the auxiliary variables.

Suppose the classes are defined by one auxiliary variable X having L categories. Let N_h be the size of class h, and let $\overline{\rho}_h$ be the population mean of the response probabilities in stratum h. If a standard model like logistic regression is used, the estimated R-indicator is a consistent estimator of

$$R_X(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{h=1}^{H} N_h(\overline{\rho}_h - \overline{\rho})^2}\ , \tag{3.1.1}$$

$R_X(\rho)$ measures the variation of the response probabilities between classes X. If the within class variation is assumed to be zero in all classes, $R_X(\rho)$ is equal to $R(\rho)$.

Dependence of the R-indicator on the set of auxiliary variables used has implications for comparing different data sets (e.g. over time or over domains). One approach could be to fix the set of auxiliary variables beforehand and keep them the same for all data sets. The maximum possible set of variables should be chosen for this. Due to over-fitting the (estimated) standard error may be affected, but estimated response probabilities will be unbiased. Another approach could be to attempt to find the best model for each data set using model selection techniques. This makes the models dependent on the sample size: the larger the sample the more variables in the model will have a significant contribution. Small samples simply not allow for proper estimation of response probabilities. Small samples will lead to a more optimistic view on representativity.

3.2 Relationship with maximal possible bias

The definition of the R-indicator does not involve any information about the target variables of the survey. Nevertheless, there is a relationship between the value of the R-indicator and the largest possible bias of estimates for target variables. Suppose the objective of the survey is to estimate the population mean Y of some target variable $\overline{Y}$. For sake of simplicity, simple random sampling without replacement is assumed. In the case of full response, the sample mean $\overline{y}$ is an unbiased estimator of the population. It can be shown (see e.g. Bethlehem, 1988 or Särndal & Lundström, 2005) that this estimator is biased in case of non-response and that the bias is approximately equal to

$$B(\overline{y}) = \frac{Cov(Y,\rho)}{\overline{\rho}} = \frac{Cor(Y,\rho)S(Y)S(\rho)}{\overline{\rho}}\ , \tag{3.2.1}$$

where $Cov(Y,\rho)$ is the population covariance between the values of the target variable and the response probabilities, $Cor(Y,\rho)$ the population correlation coefficient and $S(Y)$ is the population variance of the target variable. Since the value of the correlation coefficient is restricted to the interval [-1, 1], the maximum value of the absolute bias is equal to

$$| B(\bar{y}) | \leq \frac{S(\rho)S(y)}{\bar{\rho}} = \frac{(1 - R(\rho))S(y)}{2\bar{\rho}} = B_{max}(Y, \rho) \,. \qquad (3.2.2)$$

This upper bound cannot be computed in practical situations, but it can be estimated using the sample data and the estimated response probabilities. Expression (3.2.2) computes the maximum possible bias given the value of the R-indicator. Conversely, it is possible to define beforehand a maximum allowed absolute bias and compute a minimum value of the desired R-indicator.

4 The R-indicator in practice

4.1 Use of the R-indicator

The R-indicator proposed in this paper is promising because it can be estimated using sample data and it allows for easy interpretation. Computation of its value is reasonably straightforward with standard software like SPSS, SAS or STATA. Of course, the value of this indicator is only meaningful if accurate estimates of response probabilities can be obtained. The R-indicator as defined in (2.2.2) is just one example of such an indicator. Other R-indicators can be constructed by choosing different functions measuring the variation in response probabilities. For example Särndal & Lundström (2008) attempt to determine the best auxiliary variables for non-response reduction using a criterion based on a linear model for the response weights (i.e. the inverse response probabilities). They propose a quantity q^2 that ranks different sets of auxiliary variables with respect to their ability to reduce bias. For a fixed set of X's, this quantity can be used as an R-indicator.

R-indicators like the one proposed in (2.2.2) can be used in several different ways in the survey process. A number of possibilities are described here:

- Monitoring the survey process. Already during data collection it can become clear whether or not the composition of the collected data differs from that of the initial sample. The outcomes of this monitoring process may help to underpin a decision to initiate additional efforts to obtain data for specific groups in the target population; Such an approach will also be useful in evaluating additional data collection efforts such as re-approaching a sample of non-respondents (Hansen & Hurvitz, 1946) or a basic question approach (Kersten & Bethlehem, 1984);

- Controlling the survey process. Use of an R-indicator already during the data collection phase may reveal that the composition of the collected data may deviate more and more from representativity. This could lead to a decision to focus the remainder of the data collection process on groups that are under-represented. Groves & Heeringa (2006) call these mid-survey decisions to change the design "responsive survey design". Another way to use the R-indicator to control the survey process is to analyze representativity for a previous version of the survey. Results of such an analysis may provide input for implementing an improved data collection strategy for a new survey.

- Selection of auxiliary variables for non-response correction. Estimation of response probabilities is based on models involving auxiliary variables. Variables that significantly contribute to predicting response probabilities are also important in non-response correction techniques like adjustment weighting. Indeed, Särndal & Lundström (2008) use their q^2 indicator as variable selection tool.

- Analysis of surveys. The R-indicator can be used as a simple analysis tool providing insight in possible problems due to non-response. Like the response rate, it is a quality indicator. The R-indicator can also be very useful for comparing surveys over times or comparing survey data for different domains or regions. Particularly for a multi-national survey like the European Social Survey (ESS, see www.europeansocialsurvey.org) can help to show that country differences may have to be attributed to differences in fieldwork procedures and fieldwork results. Of course, this is only possible if response probabilities can be properly estimated in all participating countries. This may prove not to be very simple as it requires relevant auxiliary information to be available in

all these countries. It may help to define a minimum shared set of relevant auxiliary variables that should be measured in each survey.

Use of R-indicators is not restricted to social surveys. They can also be very useful for business surveys. An example is given in section 4.2. Another possibility is to apply R-indicators in the collection and processing of register data. Countries like the Scandinavian countries and The Netherlands rely more and more on register data and other administrative data for their official statistics. R-indicators can play a role here as well as a quality control tool, particularly with respect to registers that are filled over a period of time, for which auxiliary information is available (e.g. data from a previous year), and for which the data are already used for compiling statistics before it is complete.

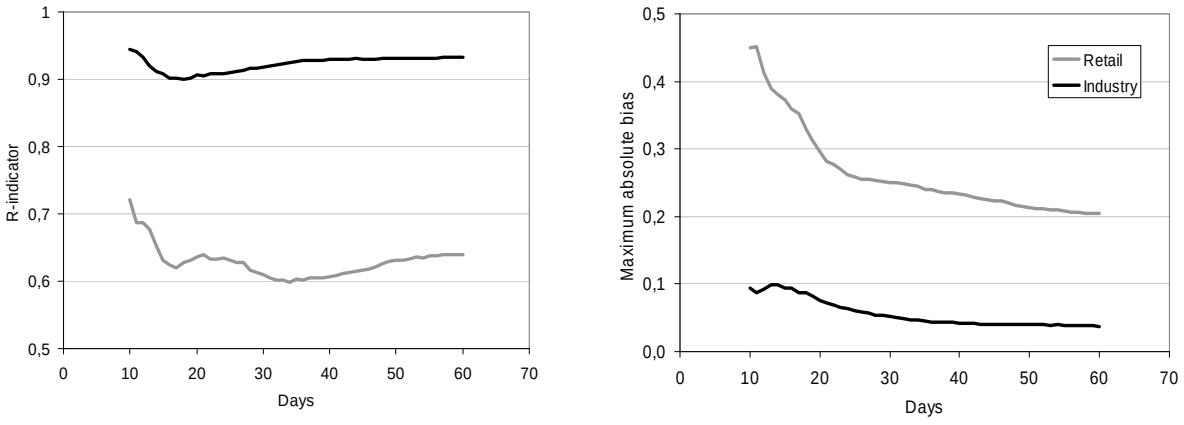
4.2 Another example

Like other national statistical institutes, Statistics Netherlands publishes time series of short term economic statistics. Data are collected on a monthly basis. The demand for timely statistics puts high pressure at data collection. To stop data collection early will general mean that less data will have been collected, and that quality of statistics is at stake.

De Nooij (2008) investigated the representativity of some short term statistics as a function of the length of the fieldwork period. He focused on two branches: industry and retail trade. To estimate the response probabilities in 2007, three auxiliary variables were used: VAT in the previous year (2006), activity, and size class.

The left-hand part of Figure 4.2-1 contains a graph containing for the two branches the values of the R-indicator as a function of the number of days in the data collection period. A first conclusion is that the R-indicators for industry are substantially higher than for the retail branch. The explanation is that the industry branch is, for a large part, a homogeneous group of large companies. Therefore, it does not make a lot of difference which companies report early and which ones late. The retail branch is much more heterogeneous with a lot of small companies. The group of early reporters is apparently not so representative. The situation does not improve in the rest of the data collection. After two months the representativity still has not improved.

Figure 4.2-1
R-indicators and maximum absolute bias for retail and industry by number of fieldwork days



The graph of the R-indicator for industry seems to indicate that, from a point of view of representativity, it is not useful to continue data collection after, say, 40 days. After that day the R-indicator does not change any more. However, one should realize that the maximum value of the bias is larger for a smaller sample size.

The right-hand part of Figure 4.2-1 contains the maximum absolute bias excluding the standard deviation $S(Y)$, making it independent of the auxiliary variable used. The maximum absolute bias almost does not change any more

after 40 days. Both graphs together suggest it might not be unreasonable to stop data collection after 40 days. The situation is different for the retail trade. As the maximum absolute bias keeps decreasing, there is no reason to stop early.

References

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation, *Journal of Official Statistics*, 4, 251-260.

Cobben, F. (2007). A follow-up with basic questions of nonrespondents to the Dutch Labour Force Survey, Discussion paper 07011, Statistics Netherlands, Voorburg, The Netherlands.

Curtin, R., Presser, S. and Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment, *Public Opinion Quarterly*, 64, 413-428.

De Nooij, G. (2008). Representativity of short term statistics, Technical Report, Statistics Netherlands, Voorburg, The Netherlands.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys, *Public Opinion Quarterly*, 70, 646–675.

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

Groves, R.M. and Peytcheva, E. (2006). The impact of nonresponse rates on nonresponse bias: a meta-analysis, *17th International Workshop on Household Survey Nonresponse*, Omaha, NE, USA.

Groves, R. M., Presser, S. and Dipko, S. (2004). The role of topic interest in survey participation decisions, *Public Opinion Quarterly*, 68, 2-31.

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

Heerwegh, D., Abts, K. and Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: do our efforts pay off?, *Survey Research Methods*, 1, 3–10.

Keeter, S., Miller, C., Kohut, A., Groves, R.M. and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey, *Public Opinion Quarterly*, 64, 125-148.

Kersten, H.M.P. and Bethlehem, J.G. (1984). Exploring an reducing the nonresponse bias by asking the basic question, *Statistical Journal of the United Nations Economic Commission for Europe*, 2, 369–380.

Merkle, D.M. and Edelman, M. (2002). Nonresponse in exit polls: a comprehensive analysis, *Survey Nonresponse* (Eds. R.M. Groves et al.), New York: John Wiley & Sons, 243-258.

Särndal, C.E. and Lundström (2005). *Estimation in surveys with nonresponse*, New York: John Wiley & Sons.

Särndal, C.E. and Lundström (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator, *Journal of Official Statistics*, 34, 167-191.

Schouten, B. (2004). Adjustment for bias in the Integrated Survey on Household Living Conditions (POLS) 1998, Discussion paper 04001, Statistics Netherlands, Voorburg, The Netherlands.

Schouten, B. (2007). A follow-up of nonresponse in the Dutch Labour Force Survey, Discussion paper 07004, Statistics Netherlands, Voorburg, The Netherlands.

Schouten, B. and Cobben, F. (2007). R-indicators for the comparison of different fieldwork strategies and data collection modes, Discussion paper 07002, Statistics Netherlands, Voorburg, The Netherlands.