An empirical validation of R-indicators

08

Fannie Cobben and Barry Schouten

Discussionpaper (08006)

Statistics Netherlands

Voorburg/Heerlen, 2008

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
2005-2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/'06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/'04–2005/'06	= crop year, financial year, etc. 2003/'04 to 2005/'06 inclusive

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

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg

second half of 2008:
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Facility Services

Cover
TelDesign, Rotterdam

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

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

Internet
http://www.cbs.nl

ISSN: 1572-0314

6008308006 X-10

AN EMPIRICAL VALIDATION OF R-INDICATORS

Summary: Response rates are often used as an indication of the quality of the survey response. They are, however, only one side to the coin; the other side is the contrast between respondents and nonrespondents, i.e. to what extent do the two groups give different answers to the survey items. Smaller differences between respondents and nonrespondents do not necessarily go together with higher response rates. Literature gives various empirical examples that show the contrary.

Complementary to the response rate there is a need for indicators that give measures of the contrast between respondents and nonrespondents. These indicators may serve as tools to assess and compare the quality of the response to different surveys. Potentially such indicators may also be used as optimisation criteria in survey designs that allow for differentiation of fieldwork protocols; so-called adaptive or responsive designs.

In an earlier paper we proposed three representativity indicators or R-indicators. In this paper we apply one of these R-indicators to a wide range of studies involving different refusal conversion and contact strategies as well as different data collection modes. We give approximations to the corresponding confidence intervals and compare the values to more detailed and elaborated analyses of the studies performed by other authors.

Keywords: Representativity; Missing data; Nonresponse; Nonresponse bias; Responsive designs.

1. Introduction

In an earlier discussion paper (Schouten and Cobben 2007) we investigated indicators that measure the dissimilarity between survey response and survey sample with respect to auxiliary variables that are available from other sources than the survey itself. We call such indicators representativity indicators or simply R-indicators. In the paper we proposed three R-indicators and identified a number of areas for future research

- Search for other promising R-indicators;
- Empirical validation of the proposed R-indicators;
- Estimators for the standard errors and confidence intervals of R-indicators;
- Interpretation and normalization of R-indicators relative to sample size.

Here, we focus on the second and third area; empirical validation, standard errors and confidence intervals.

We apply one of the proposed R-indicators to a series of studies that were conducted at Statistics Netherlands over the last three years. The objectives of each of those studies were the comparison of different data collection strategies. The studies involved different data collection modes, different amounts of prepaid incentives, different refusal conversion strategies and different contact strategies. For each of the studies a detailed analysis was done and documented. These studies are, therefore, suited for an empirical validation of the R-indicator. We compare the values of the R-indicator to the conclusions in the analyses.

R-indicators are computed based on a vector of available auxiliary variables. These are variables that are external to the survey, i.e. they can be linked to the survey sample from registers or administrative data. In this paper we consider only the case where auxiliary variable can be linked directly to the sample. R-indicators can also be computed in case only population totals are known. This will be the topic of a future paper. R-indicators are functions of the set of available auxiliary variables; they should, therefore, always be interpreted in conjunction with this set. In other words R-indicators measure the extent to which response is representative for the auxiliary variables at hand.

R-indicators are based on a realisation of a survey, i.e. on one sample and on one response from each sample unit. Consequently, R-indicators are random variables, or better are estimators of some population representativity parameter. The population parameter is the expected representativity of the response given the sampling design, given the missing-data-mechanism due to the nonresponse, and given the set of selected auxiliary variables. This means that the values of R-indicators are subject to sampling variation. As a consequence, an R-indicator has a standard error. R-indicators should, therefore, always be given together with approximate confidence intervals that reflect this uncertainty. As usual, for large survey samples confidence intervals for R-indicators will be smaller than for small survey samples.

An R-indicator may also be subject to bias depending on the strategy with which the available auxiliary variables are incorporated in the R-indicator. If an auxiliary vector is imposed beforehand, i.e. if we fix the auxiliary vector in the estimation of the R-indicators, then they are approximately unbiased. In that case the R-indicator will have no systematic deviation from the population representativity parameter it seeks to estimate. However, if the computation of R-indicators is based on model selection using some prescribed significance level, i.e. the auxiliary variables that are used in the estimation are a subset of the auxiliary vector, then the R-indicators may also be biased. Large survey samples allow for larger models than small survey samples, and as a consequence lead to lower values of the R-indicators. This can be made more intuitive by considering a very small random sample of say five persons. In such a small sample no interaction between response behaviour and household characteristics will be significant.

It is important in the following to constantly keep in mind that R-indicators themselves are estimators of some unknown population representativity parameter.

If we find a change in the values of an R-indicator, then we need to test whether the increase or decrease is significant at some level.

In this paper we compute the R-indicators for all studies using a fixed auxiliary vector composed of age, type of household, ethnic background and degree of urbanization. In some of the original studies the set of auxiliary variables used by the survey researchers was larger or different. In those studies we will also compute the R-indicators for the auxiliary vector used by the authors.

Summarizing, we answer three research questions in this paper:

1. Do R-indicators confirm findings in detailed analyses of empirical studies with different data collection strategies?
2. How is the relation between response rate and R-indicator, nonresponse bias and Mean Square Error (MSE)?
3. How is the relation between sample size and standard error for the studies that are investigated?

In section 2, we review the background to the R-indicator that we apply in this paper. In section 3, we describe the different studies and the application of the R-indicator. In section 4 we derive general conclusions about the values and standard errors of the R-indicator. Finally, in section 5 we discuss the findings.

2. R-indicators

Schouten and Cobben (2007) propose three indicators to measure the representativity of the response to a survey. The first two R-indicators are based on, respectively, the sample standard deviation and the sample variance of the estimated response probabilities. The third R-indicator employs a proportion of fit measure, e.g. Nagelkerke's pseudo R^2. For background and details we refer to their paper. In this paper we apply the first R-indicator. In this section we will show that the R-indicator can be linked directly to the nonresponse bias and the mean square error of the response mean of survey items.

In section 2.1 we give a brief review of the indicator. Next, in section 2.2 we relate the indicator to the maximal absolute bias and maximal root mean square error, which we will use as a derived quality measure. In section 2.3 we define what we call response-representativity functions. Finally, we estimate standard errors and confidence intervals in section 2.4.

2.1 Notation and R-indicator

Let $i = 1,2,3,\ldots,N$ be the labels of the units in the population. By s_i we denote the 0-1-sample indicator, i.e. in case unit i is sampled it takes the value 1 and 0 otherwise. By r_i we denote the 0-1-response indicator for unit i. If unit i is sampled and did respond then $r_i = 1$. It is 0 otherwise. The sample size is n. Next,

π_i denotes the first-order inclusion probability of unit i, and ρ_i is the probability that unit i responds in case it is sampled, i.e. $\rho_i = P[r_i = 1 | s_i = 1]$. Let $\tilde{\rho} = (\rho_1, \rho_2, \ldots, \rho_N)'$ be the vector of response probabilities.

As we do not observe ρ_i, we have to estimate its value. We do so by using a vector of auxiliary information x_i that is available for all units i in the sample. We let $\hat{\rho}_i$ denote an estimator for ρ_i that uses all or a subset of the available auxiliary variables contained in x_i. By $\hat{\bar{\rho}}$ we denote the weighted sample average of the estimated response probabilities, i.e.

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

where we use the inclusion weights π_i. (1) is an unbiased estimator of $\bar{\rho}$, the population average of the response probabilities.

We apply R-indicator $\hat{R}_1$ that is proposed by Schouten and Cobben (2007). For convenience we omit the index. The R-indicator is defined as

$$\hat{R}(\hat{\rho}) = 1 - 2\sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \frac{s_i}{\pi_i} (\hat{\rho}_i - \hat{\bar{\rho}})^2} = 1 - 2\hat{S}(\hat{\rho}). \tag{2}$$

In (2) the variation in the estimated response probabilities is weighted using the inclusion probabilities and is computed with respect to the average weighted response probability given by (1). By weighting estimated response probabilities we adjust for the possibly unequal inclusion probabilities. $\hat{R}$ always attains values in the interval [0,1] and lower values of $\hat{R}$ correspond to a less representative response.

The R-indicator (2) is an estimator of the population R-indicator that is based on the population standard deviation $S(\rho)$ of the 'true' response probabilities

$$R(\rho) = 1 - 2\sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (\rho_i - \bar{\rho})^2} = 1 - 2S(\rho). \tag{3}$$

$\hat{R}$ is based on a sample and it employs estimated response probabilities. As a consequence, it is a random variable with a certain precision and bias. If we were to know the unit response probabilities, i.e. if we did not have to estimate them, then $\hat{R}$ would be an unbiased estimator of (3). However, due to the estimation of the response probabilities, $\hat{R}$ may be a biased estimator of (3). This bias arises from the inevitable restriction to the set of auxiliary variables x_i in the estimation of the ρ_i's. The R-indicator does not capture differences in response probabilities within subgroups of the population other than the subgroups defined by the classes of x_i. In other words if the missing-data-mechanism is Not-Missing-at-Random, then $\hat{R}$ may be biased. The bias is unknown and cannot be estimated.

If we let $h = 1,2,\ldots,H$ denote strata defined by x_i, N_h be the size of stratum h, and $\bar{\rho}_h$ be the population average of the response probabilities in stratum h, then $\hat{R}$ is an approximately unbiased estimator of

$$R_x(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{\rho}_h - \bar{\rho})^2}\ , \tag{4}$$

in case standard models like logistic regression or linear regression are used to estimate the response probabilities. Clearly, (3) and (4) may attain different values. (3) and (4) will be approximately the same if the missing-data-mechanism is Missing-at-Random conditional on x_i.

In section 2.4 we consider standard errors and confidence intervals for (2).

2.2 Maximal absolute nonresponse bias and maximal root mean square error

The R-indicator $\hat{R}$ measures the extent to which the response composition deviates from the population composition with respect to a set of auxiliary variables. The important question arises what the value of the R-indicator implies for the bias and mean square error of survey items. Schouten and Cobben (2007) show that $\hat{R}$ induces an upper bound to the maximal absolute nonresponse bias of response means of 0-1 dummy variables. We will show that for any survey item y the R-indicator can be used to set upper bounds to the nonresponse bias and to the root mean square error (RMSE) of adjusted response means. We will use these bounds next to the R-indicator to show the impact under worst-case scenarios.

Let $\bar{Y}$ be the population mean of survey item y and $S(y)$ be the population standard deviation. A naive estimator for $\bar{Y}$ is the response-based Horvitz-Thompson estimator $\hat{\bar{y}}_{HT}$

$$\hat{\bar{y}}_{HT} = \frac{1}{N}\frac{n}{r}\sum_{i=1}^{N}\frac{r_i}{\pi_i}y_i\ , \tag{5}$$

where $r = \sum_{i=1}^{N} r_i$ is the size of the response. $\hat{\bar{y}}_{HT}$ is the response mean of the survey item y adjusted for unequal inclusion probabilities. Bethlehem (1988) refers to this estimator as the modified Horvitz-Thompson estimator. It can be shown (e.g. Bethlehem 1988, Särndal and Lundström 2005) that its bias $B(\hat{\bar{y}}_{HT})$ is approximately equal to

$$B(\hat{\bar{y}}_{HT}) = \frac{C(y,\rho)}{\bar{\rho}}\ , \tag{6}$$

with $C(y,\rho) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(\rho_i - \bar{\rho})$ the population covariance between the survey items and response probabilities. For a close approximation of the variance $\text{var}(\hat{\bar{y}}_{HT})$ of $\hat{\bar{y}}_{HT}$ we refer to Bethlehem (1988). Here, we make the simplifying

assumption that $\text{var}(\hat{\bar{y}}_{HT})$ is approximately equal to the variance of the response mean in a simple random sample without replacement, i.e.

$$\text{var}(\hat{\bar{y}}_{HT}) \cong (1-\frac{n\bar{\rho}}{N})\frac{S^2(y)}{n\bar{\rho}}. \tag{7}$$

Following the reasoning in Schouten and Cobben (2007) it can be shown that

$$|B(\hat{\bar{y}}_{HT})| \leq \frac{S(\rho)S(y)}{\bar{\rho}} = \frac{(1-R(\rho))S(y)}{2\bar{\rho}} = B_m(\rho, y). \tag{8}$$

Clearly, we do not know the upper bound $B_m(\rho, y)$ in (8) but we can estimate it using the sample and the estimated response probabilities

$$\hat{B}_m(\hat{\rho}, y) = \frac{(1-\hat{R}(\hat{\rho}))\hat{S}(y)}{2\hat{\bar{\rho}}}, \tag{9}$$

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

In a similar way we can set a bound to the RMSE of $\hat{\bar{y}}_{HT}$. It holds approximately that

$$\begin{aligned} RMSE(\hat{\bar{y}}_{HT}) &= \sqrt{B^2(\hat{\bar{y}}_{HT}) + Var(\hat{\bar{y}}_{HT})} \\ &\leq \sqrt{B_m^2(\rho, y) + (1-\frac{n\bar{\rho}}{N})\frac{S^2(y)}{n\bar{\rho}}} = E_m(\rho, y). \end{aligned} \tag{10}$$

Again, we do not know $E_m(\rho, y)$. Instead we use a sample-based estimator that employs the estimated response probabilities

$$\hat{E}_m(\hat{\rho}, y) = \sqrt{\hat{B}_m^2(\hat{\rho}, y) + (1-\frac{n\hat{\bar{\rho}}}{N})\frac{\hat{S}^2(y)}{n\hat{\bar{\rho}}}}. \tag{11}$$

The bounds $\hat{B}_m(\hat{\rho}, y)$ and $\hat{E}_m(\hat{\rho}, y)$ in (9) and (11) are different for each survey item y. For comparison purposes it is, therefore, convenient to define a hypothetical survey item. We suppose that $\hat{S}(y) = 0{,}5$, which is the maximal standard deviation of a 0-1 survey item. The corresponding bounds we denote by $\hat{B}_m(\hat{\rho})$ and $\hat{E}_m(\hat{\rho})$. They are equal to

$$\hat{B}_m(\hat{\rho}) = \frac{(1-\hat{R}(\hat{\rho}))}{4\hat{\bar{\rho}}} \tag{12}$$

$$\hat{E}_m(\hat{\rho}) = \sqrt{\hat{B}_m^2(\hat{\rho}) + (1-\frac{n\hat{\bar{\rho}}}{N})\frac{1}{4n\hat{\bar{\rho}}}}. \tag{13}$$

We will compute (12) and (13) in all studies described in section 3. We have to note that (9), (10), (12) and (13) are again random variables that have a certain precision and that are potentially biased.

2.3 Response-representativity functions

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

Clearly, lower bounds to the R-indicator depend on the survey item. Therefore, again we restrict ourselves to a hypothetical survey item for which $\hat{S}(y) = 0{,}5$.

It is not difficult to show from (12) that if we demand that

$$\hat{B}_m(\hat{\rho}) \leq \gamma, \tag{14}$$

than it must hold that

$$\hat{R} \geq 1 - 4\hat{\bar{\rho}}\gamma = r_1(\gamma, \hat{\rho}). \tag{15}$$

Analogously, using (13) and demanding that

$$\hat{E}_m(\hat{\rho}) \leq \gamma, \tag{16}$$

we arrive at

$$\hat{R} \geq 1 - 4\hat{\bar{\rho}}\sqrt{\gamma^2 - (1 - \frac{n\hat{\bar{\rho}}}{N})\frac{1}{4n\hat{\bar{\rho}}}} = r_2(\gamma, \hat{\rho}). \tag{17}$$

In (15) and (17) we let $r_1(\gamma, \hat{\rho})$ and $r_2(\gamma, \hat{\rho})$ denote lower limits to the R-indicator. In the following, we refer to $r_1(\gamma, \hat{\rho})$ and $r_2(\gamma, \hat{\rho})$ as response-representativity functions. We will compute them for the various studies in section 3.

2.4 Standard error and confidence interval

We are interested in the precision of the R-indicator. If we want to compare its values for different surveys or data collection strategies, we need to estimate the standard error of the R-indicator.

The R-indicator $\hat{R}$ involves the sample standard deviation of the estimated response probabilities. This means that there are two random processes involved. The first process is the sampling of the population. The second process is the response mechanism of the sampled units. If the true response probabilities were known, then

drawing a sample would still introduce uncertainty about the population R-indicator and, hence, lead to a certain loss of precision. However, since we do not know the true response probabilities, these probabilities are estimated using the response. This gives an additional loss of precision.

An analytical derivation of the standard error of $\hat{R}$ is not straightforward due to the estimation of the response probabilities. In this paper we, therefore, resign to numerical approximations of the standard error.

As we do not want to formulate a model for the response probabilities, we estimate the standard error of the R-indicator by non-parametric bootstrapping (Efron and Tibshirani 1993). The non-parametric bootstrap estimates the standard error of the R-indicator by drawing a number $b = 1,2,\ldots,B$ of so-called bootstrap samples. These are samples drawn independently and with replacement from the original dataset, of the same size *n* as the original dataset. On every bootstrap sample *b* the R-indicator is calculated. We thus obtain *B* replications of the R-indicator; $\hat{R}_b^{BT}$, $b = 1,2,\ldots,B$. The standard error for the empirical distribution of these *B* replications is an estimate for the standard error of the R-indicator, that is

$$s_R^{BT} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{R}_b^{BT} - \hat{\bar{R}}^{BT}\right)^2} \qquad (18)$$

where $\hat{\bar{R}}^{BT} = \frac{1}{B}\sum_{b=1}^{B}\hat{R}_b^{BT}$ is the average estimated R-indicator.

As a rule of thumb, the number of bootstrap replications that is needed for a good estimate of the standard error, i.e. little bias and small standard deviation, very seldom exceeds $B = 200$ (Efron and Tibshirani 1993, p. 52). In the approximations in section 3 we took $B = 200$ for all studies. We experimented with the number of bootstrap samples B and found that in all cases the estimate of the standard error had converged at much smaller values than $B = 200$.

We determine $100(1-\alpha)\%$ confidence intervals by assuming a normal approximation of the distribution of $\hat{R}$ employing the estimated standard errors using (18)

$$CI_\alpha^{BT} = (\hat{R} \pm \xi_{1-\alpha} \times s_R^{BT}) \qquad (19)$$

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

3. Application of the R-indicator

In this section we apply the R-indicator to five studies that investigate different refusal conversion techniques, combinations of data collection modes and contact strategies. The first two studies both involve the Dutch Labour Force Survey (LFS). The first of the two LFS studies is an investigation of both the call-back approach

(Hansen and Hurwitz 1946) and the basic-question approach (Kersten and Bethlehem 1984). The second of the LFS studies employs pre-paid incentives in different amounts. The third and fourth study both deal with mixed-mode data collection designs applied to the Dutch Safety Monitor survey and the Dutch Informal Economy survey.

The first four studies were not explicitly directed at improving the representativity of response. The final and fifth study is different from the others as this study was explicitly based on R-indicators. In the Dutch Consumer Confidence survey different contact strategies were attempted that aimed at optimisation of the representativity of the response. In these studies we explicitly tried to enhance response rather than reduce nonresponse.

In sections 3.2 to 3.5 we pay a closer look at each of the studies in connection to the representativity of their different fieldwork strategies. First, in section 3.1 we give a brief description of the available auxiliary variables that were linked from the sampling frame and from administrative data.

3.1 Auxiliary variables

R-indicators employ information that is auxiliary to the survey and that is available for all sampled units. The motivation for this is that nonresponse can be viewed as a second phase in the survey, where the sampling design constitutes the first phase. In studying and measuring the impact of nonresponse on the representativity we are interested in the second phase. However, R-indicators can be extended to the situation where only population totals are given instead of sample totals. In these situations the R-indicators need to be adjusted for the sampling design so that we isolate the impact of the nonresponse.

In this paper, we restrict ourselves to auxiliary information at the sample level. In all studies we were able to link the sample directly to various external registers and administrative data.

We again stress that R-indicators cannot be viewed separately from the auxiliary information that is employed. In fact, the value of the R-indicator should always be given together with the set of auxiliary variables.

Table 3.1.1: The auxiliary variables used for the application of the R-indicator in all surveys except for the LFS and Consumer Confidence survey

Variable	*Categories*
Age in 6 classes	≤ 25, 26 – 35, …, 56 – 65, 65 +
Ethnic group in 3 classes	Native; Western non-native; non-Western non-native
Degree of urbanization in 5 classes	Very high, High, Average, Low, Very low
Household type in 5 classes	Single, Couple, Couple with children, Single parent, Other

Table 3.1.2: The auxiliary variables used for the application of the R-indicator in the LFS and Consumer Confidence survey

Variable	*Categories*
Average age in 6 classes	≤ 25, 26 – 35, …, 56 – 65, 65 +
Ethnic group in 4 classes	Native; Western non-native; non-Western non-native, Mixture
Degree of urbanization in 5 classes	Very high, High, Average, Low, Very low
Household type in 6 classes	Single, Couple, Couple with children, Single parent, Other, Mixture

To each of the samples we linked the same set of auxiliary variables: age, ethnic group, degree of urbanization and household type. However, we have to make a distinction between household and individual surveys with respect to the categories of these variables. In Tables 3.1.1 and 3.1.2 we show the categories for, respectively, individual and household surveys.

Two of the surveys that were involved in the studies are the Labour Force survey (LFS) and the Consumer Confidence survey (CCS). Both are household surveys and the samples consist of addresses whereas for all the other surveys the samples consist of persons. In the LFS and CCS all households living at the sampled addresses are asked to complete the questionnaire. We consider the response to these surveys, therefore, as a feature of the households living at the addresses. Consequently, the selected auxiliary variables are aggregated to the household level. Instead of individual characteristics, we take the characteristics of the cores of the households, i.e. the heads of the households and the partners if present. The auxiliary variable age becomes the average age of the household cores at the address. The auxiliary variable ethnic group gets an additional category in case the household cores are a mixture of ethnicities. The auxiliary variable degree of urbanization is not affected by the aggregation to the household level. The auxiliary variable household type is not affected unless there is more than one household at the address. We added an additional category for addresses with more than one household.

In the original LFS studies of sections 3.2 and 3.3 the set of auxiliary variables was different from the set of four variables in Table 3.1.2. We also compute the R-indicators for the sets of auxiliary variables that were used in those studies. In sections 3.2 and 3.3 these auxiliary variables are described in detail.

3.2 Labour Force Survey; follow-up study 2005

From July to December 2005 Statistics Netherlands conducted a large scale follow-up of 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) or a basic-question approach (Kersten and Bethlehem 1984). The samples consisted of LFS households that

refused, were not processed or were not contacted in the LFS of the months July – October. In the design of the follow-up study we used the recommendations in the studies by Stoop (2005) and Voogt (2004).

The LFS is a monthly household survey. In 2005 the sample size was approximately 6500 addresses per month. The target population consists of all inhabitants of the Netherlands of 15 years and older, except for people living in institutions. The main objective of the LFS is a set of statistics about the employment status of persons and households. Most statistics concern the population of 15 – 64 years. The households are interviewed face-to-face in CAPI. Proxy interviewing is allowed under certain circumstances. The LFS is a rotating panel. Each household is asked whether it is willing to participate in four CATI interviews with time lags of three months. Hence, the last interview is 12 months after the CAPI interview. The study involved nonresponse to the CAPI interview.

The main characteristics of the call-back and basic-question approaches applied to the LFS are given in Table 3.2.1. For details we refer to Schouten (2007) and Cobben and Schouten (2007). The call-back approach employed the original household questionnaire in CAPI, while the basic-question approach used short questionnaires in a mixed-mode setting. The mixed-mode design involved web, paper and CATI. CATI was used for all households with a listed phone number. Households without a listed phone number received an advance letter, a paper questionnaire and a login to a secured website containing the web questionnaire. Respondents were left the choice to fill in either the paper or web questionnaire.

Table 3.2.1: Characteristics of the two approaches in the follow-up study.

Call-back approach	*Basic-question approach*
• LFS questionnaire to be answered by all members of the household in CAPI • 28 interviewers geographically selected from historically best performing interviewers • Interviewer was different from interviewer that received non-response • Interviewers received additional training in doorstep interaction • Extended fieldwork period of two months • Interviewer could offer incentives • Interviewers could receive a bonus • A paper summary of the characteristics of the non-responding household was sent to the interviewer • Allocation of address one week after non-response	• A strongly condensed questionnaire with key questions of the LFS which takes between 1 and 3 minutes to answer or fill in • Mixed-mode data collection design using web, paper and CATI • The questionnaire was to be answered by one person per household following the next birthday method • The timing is one week after the household is processed as a nonresponse

In the LFS study a number of municipalities was omitted from the samples as these requested large travelling times due to the small number of participating

interviewers. The remaining municipalities contained both rural areas and strongly urbanized areas. Consequently, the sample size of the LFS pilot was somewhat smaller and equalled $n = 18074$, of which 11275 households responded. The nonresponding households were stratified according to the cause of nonresponse. Households that were not processed or contacted, and households that refused were eligible for a follow-up. Households that did not respond due to other causes like illness were considered ineligible. In total 6171 households were eligible. From these households two simple random samples were drawn of size 775. In the analyses the non-sampled eligible households were left out. The sampled eligible households received a weight accordingly. The 11275 LFS respondents and the 628 ineligible households all received a weight one. This implies that the inclusion probabilities in (1) and (2) are unequal for this example.

Schouten (2007) compared the LFS respondents to the converted and persistent nonrespondents in the call-back approach using a large set of demographic and socio-economic characteristics. The auxiliary variables in section 3.1 were a subset of this set. He used logistic regression models to predict the type of response. He concluded that the converted nonrespondents in the call-back approach are different from the LFS respondents with respect to the selected auxiliary variables. Furthermore, he found no evidence that the converted nonrespondents were different from persistent nonrespondents with respect to the same characteristics. These findings lead to the conclusion that the combined response of the LFS and call-back approach is more representative with respect to the selected auxiliary variables.

The additional response in the basic question approach was analyzed by Cobben and Schouten (2007) using the same set of auxiliary variables and employing the same logistic regression models. For this follow-up the findings were different for households with and without a listed phone number. When restricted to listed households, they found the same results as for the call-back approach; the response becomes more representative after the addition of the listed basic-question respondents. However, for the overall population, i.e. including the unlisted households, the converse was found. The basic-question approach gives 'more of the same' and, hence, sharpens the contrast between respondents and nonrespondents. Combining LFS response to basic-question response leads to a less representative composition. In the logistic regression models by Cobben and Schouten (2007) the 0-1 indicators for having a listed phone number and having a paid job gave a significant contribution.

Table 3.2.2 shows the weighted sample size, response rate, $\hat{R}$, $CI_{0.05}^{BT}$, $\hat{B}_m$ and $\hat{E}_m$ for the response to the LFS, the response of the LFS combined with the call-back response and the response of the LFS combined with the basic-question response. The standard errors are relatively large with respect to the studies in subsequent sections due to the weighting.

There is an increase in $\hat{R}$ when the call-back respondents are added to the LFS respondents. As both the response rate and the R-indicator increase, the maximal

absolute bias $\hat{B}_m$ decreases. The confidence intervals $CI_{0.05}^{BT}$ for the LFS response and the combined LFS and call-back response overlap. However, the R-indicators $\hat{R}$ for the two groups are dependent as they concern the same sample. The bootstrapped 95%-confidence interval for the difference between the R-indicator for the combined response and the LFS response equals (0,9%, 6,6%). As this interval does not cover 0%, the composition of the combined response has improved significantly at the 5% level by adding the call-back response.

Table 3.2.2: Weighted sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, LFS plus call-back, and LFS plus basic-question.

Response	*n*	*Rate*	$\hat{R}$	$CI_{0.05}^{BT}$	$\hat{B}_m$	$\hat{E}_m$
LFS	18074	62,2%	82,4%	(79,7 - 84,8)	7,1%	7,1%
LFS + call-back	18074	76,9%	86,1%	(83,3 -88,9)	4,5%	4,5%
LFS + basic-question	18074	75,6%	82,8%	(80,2 - 85,0)	5,7%	5,7%

There is a small increase in $\hat{R}$ when we compare the LFS response to the combined response with the basic-question approach. This increase is not significant. $\hat{B}_m$ slightly decreases. In Table 3.2.3 this comparison is restricted to households with a listed phone number. The R-indicator in general is much higher than for all the households. Because the sample size is now smaller, the estimated standard errors are larger as is reflected in the width of the confidence interval. $\hat{B}_m$ is decreased. For the combined response in the LFS and the basic-question approach, we see a slight increase of $\hat{R}$ but again this increase is not significant. $\hat{B}_m$ decreases.

Table 3.2.3: Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, and LFS plus basic-question restricted to households with listed phone numbers.

Response	*n*	*Rate*	$\hat{R}$	$CI_{0.05}^{BT}$	$\hat{B}_m$	$\hat{E}_m$
LFS	10135	68,5%	87,5%	(84,3 – 90,8)	4,6%	4,6%
LFS + basic-question	10135	83,0%	87,8%	(84,8 – 90,8)	3,7%	3,7%

Cobben and Schouten (2007) and Schouten (2007) used much larger sets of auxiliary variables than the set of four variables used in Tables 3.2.2 and 3.2.3. The additional variables are given in Table 3.2.4. They selected variables in logistic regression models for response probabilities in case these variables gave a significant contribution at the 5% level.

Table 3.2.4: The additional auxiliary variables in the studies by Schouten (2007) and Cobben and Schouten (2007).

Variable	*Categories*
Household has a listed phone number	Yes, No
Region of the country	North, East, West, South
Province and 4 largest cities	12 provinces, Amsterdam, Rotterdam, The Hague, Utrecht
Gender	Male, Female, Mix
Average house value at zip code level	11 classes
At least one member of household core is self-employed	Yes, No
At least one member of household core has a subscription to the CWI	Yes, No
At least one member of household core receives social allowance	Yes, No
At least one member of household core has a paid job	Yes, No
At least one member of household core receives disability allowance	Yes, No

Table 3.2.5: Weighted sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, LFS plus call-back, and LFS plus basic-question for the extended set of auxiliary variables.

Response	*n*	*Rate*	$\hat{R}$	$CI_{0.05}^{BT}$	$\hat{B}_m$	$\hat{E}_m$
LFS	18074	62,2%	80,1%	(77,5 – 82,7)	8,0%	8,0%
LFS + call-back	18074	76,9%	85,1%	(82,4 – 87,8)	4,8%	4,9%
LFS + basic-question	18074	75,6%	78,0%	(75,6 – 80,4)	7,3%	7,3%

Table 3.2.6: Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, and LFS plus basic-question restricted to households with listed phone numbers and for the extended set of auxiliary variables.

Response	*n*	*Rate*	$\hat{R}$	$CI_{0.05}^{BT}$	$\hat{B}_m$	$\hat{E}_m$
LFS	10135	68,5%	86,3%	(83,1 – 89,5)	5,0%	5,1%
LFS + basic-question	10135	83,0%	87,5%	(84,3 – 90,7)	3,8%	3,8%

Tables 3.2.5 and 3.2.6 are similar to Tables 3.2.2 and 3.2.3. However, now the extended set of auxiliary variables is used in the computation of the R-indicators. In the estimation of the response probabilities only those variables are included that give a significant contribution at the 5% level. In all cases the R-indicators are smaller. This is expected as a larger set of auxiliary variables is used and more

differences in response probabilities can be captured. The differences in the R-indicator between the various response groups are sharpened in Table 3.2.4 and 3.2.5 and are bigger than in Tables 3.2.2 and 3.2.3. Again the difference between the LFS response and the LFS plus call-back response of 5,0% is significant at the 5% level. However, the differences in the R-indicator for the LFS plus basic question response in Tables 3.2.5 and 3.2.6 are not significant at the 5% level.

We find in the example of the LFS follow-up that the R-indicators confirm the conclusions for the call-back approach and the basic question approach. Furthermore, the increase in the R-indicator that follows by adding the call-back response is significant at the 5% level.

3.3 Labour Force Survey; pilot incentives 2005

In November and December 2005, Statistics Netherlands conducted a pilot to evaluate the effect of pre-paid incentives in the LFS. There were four separate samples in the pilot that received different amounts of incentives: no incentive, an incentive of 5 post stamps, 10 post stamps or 20 post stamps. The objective of the pilot was to evaluate the effect of incentives on the response percentage and the composition of the response. Wetzels and Schmeets (2006a and b) give a detailed description of the pilot.

Since the sample of addresses that received a pre-paid incentive of 20 stamps is small, only 250 addresses, we leave this group out of the analysis. Table 3.3.1 shows the sample sizes (at individual level) and the response rates of the other three groups. As expected the response rates go up with the number of stamps.

Wetzels and Schmeets (2006 b) compared the incentive groups one-by-one to the group without incentives. To this end they added each of the corresponding samples separately to the sample that received no incentives. Next, they constructed logistic regression models for the 0-1 response indicator and used the 0-1 indicator for incentives as one of the explanatory variables. The other explanatory variables were a mix of demographic and socio-economic variables originating from administrative data. As the response rate increases when an incentive is given, the 0-1 indicator for incentive enters their models in all cases. Furthermore, Wetzels and Schmeets found that offering incentives affects also the composition of the response. The 0-1 indicator for incentive interacts with the auxiliary variables ethnic group and degree of urbanization. Both 5 and 10 stamps reduce the selectivity with respect to degree of urbanization but increase the selectivity with respect to ethnic group. This effect is strongest for an incentive of 5 stamps.

We computed R-indicators for each of the groups. However, the set of auxiliary variables used by Wetzels and Schmeets (2006 b) was slightly different from the set of variables that we selected in Table 3.1.1. Table 3.3.1 contains their variables. Tables 3.3.2 and 3.3.3 contain the R-indicators, respectively, from Table 3.1.1 and from Table 3.3.1.

In Table 3.3.2 we see that the incentives decreased the R-indicator. However, the decrease is not significant at the 5% level for both groups. $\hat{B}_m$ increases slightly for the group with 5 stamps and is approximately the same for 0 and 10 stamps. Table 3.3.3 gives a different picture. Again the R-indicator for the incentive of 5 stamps is smallest, but the difference with zero incentive group is not signifcant anymore. The R-indicator for the incentive of 10 stamps is now bigger than that of the zero incentive group.

Table 3.3.1: The auxiliary variables used by Wetzels and Schmeets (2006b) in their analysis of the LFS incentives study.

Variable	*Categories*
Age in 5 classes	<34, 35-44, 45-54, 55-64, >65
Ethnic group in 3 classes	Native; Western non-native; non-Western non-native
Living in Amsterdam, Rotterdam or The Hague	Yes, No
Income in 3 classes	Not available, Below average, Above average
Household size in 5 classes	1, 2, 3, 4, >4

Table 3.3.2: Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for response following from no stamps, 5 stamps and 10 stamps.

Response	*n*	*Rate*	$\hat{R}$	$CI^{BT}_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
No stamps	11774	66,6%	85,5%	(83,8 - 87,2)	5,4%	5,5%
5 stamps	5906	72,2%	82,1%	(79,9 - 84,3)	6,2%	6,3%
10 stamps	5982	73,8%	84,2%	(81,9 - 86,5)	5,4%	5,4%

Table 3.3.3: Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for response following from no stamps, 5 stamps and 10 stamps for the alternative set of auxiliary variables.

Response	*n*	*Rate*	$\hat{R}$	$CI^{BT}_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
No stamps	11774	66,6%	84,1%	(82,5 – 85,8)	6,0%	6,0%
5 stamps	5906	72,2%	81,1%	(78,9 – 83,3)	6,6%	6,6%
10 stamps	5982	73,8%	84,6%	(82,3 – 86,9)	5,2%	5,3%

The R-indicators indicate that the composition of the response to a pre-paid incentive of 5 stamps is less balanced than those of 0 and 10 stamps. These findings confirm the analyses by Wetzels and Schmeets. The difference between the R-

indicators of the groups that received no stamps and 10 stamps is smaller and its sign depends on the variables used in the estimation of the response probabilities.

3.4 Safety Monitor and Informal Economy; pilots mixed-mode 2006

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

The regular Safety Monitor surveys individuals of 15 years and older in the Netherlands about issues that relate to safety and police performance. The majority of the sample units are approached in the first quarter. In the remaining quarters small samples are allocated to fieldwork. In the second to fourth quarters of each year statistics are only disseminated at the national level. In the first quarter statistics are detailed to a low regional level. The Safety Monitor is a mixed-mode survey. Persons with a listed phone number are approached by CATI. Persons that cannot be reached by telephone are approached by CAPI. In the 2006 pilot, the possibility to use the internet as one of the modes in a mixed-mode strategy was evaluated. Persons in the pilot were first approached with a web survey. Nonrespondents to the web survey were re-approached by CATI in case they had a listed phone number and by CAPI otherwise. In Table 3.4.1 we give the response rates for the normal survey, the pilot response to the web only, and the response to the pilot as a whole.

The response to the web survey alone is low. Only 30% of the persons filled in the web questionnaire. This implied that close to 70% of the sampled units was re-allocated to either CAPI or CATI. This resulted in an additional response of approximately 35%. The overall response rate is slightly lower than that of the normal survey.

Fouwels et al. (2006) performed a univariate analyses of response compositions. They argue that the response rate is lower for the pilot but that this decrease is quite stable over various demographic subgroups. They observe a univariately declining response rate for categories of the auxiliary variables in Table 3.1.1. However, they do find indications that the response becomes less representative in case the comparison is restricted to the web respondents only. This holds, as expected, especially for the age of the sampled persons; elderly people more often do not have access to the web and are less acquainted with using computers.

Table 3.4.1 shows that the R-indicator for the web response is lower than that of the regular survey. The corresponding p-value is close to 5%. As a consequence of both a low response rate and a low R-indicator, the maximal absolute bias $\hat{B}_m$ is more than twice as high as for the regular survey. However, for the pilot as a whole both the R-indicator and $\hat{B}_m$ are similar to the regular survey. The only difference is the width of the confidence interval. Due to the smaller sample size of the pilot the estimated standard errors are larger than in the regular survey.

Table 3.4.1: Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for response for the regular Safety Monitor, the pilot with web only and the pilot with web and CAPI/CATI follow-up.

Response	*n*	*Rate*	$\hat{R}$	$CI_{0.05}^{BT}$	$\hat{B}_m$	$\hat{E}_m$
Regular	30139	68,9%	81,4%	(80,3 - 82,4)	6,8%	6,8%
Pilot – web	3615	30,2%	77,8%	(75,1 - 80,5)	18,3%	18,4%
Pilot – web plus	3615	64,7%	81,2%	(78,3 - 84,0)	7,3%	7,4%

The findings in Table 3.4.1 do not contradict those of Fouwels et al. (2006). We also find that the web response in the pilot has a less balanced composition whereas the composition of the full pilot response is not markedly worse than that of the Safety Monitor itself.

The second mixed-mode pilot in 2006 concerned the Informal Economy. This is a new survey that is set up at Statistics Netherlands. The 2006 survey served as a pilot for subsequent years. The target population consists of individuals of 16 years and older. Questions about unpaid labour, or moonlighting, are the main interest of this survey. In the pilot, two samples are selected. One sample is approached by CAPI, the other by a combination of a web- and a paper questionnaire. Nonrespondents to the web/paper survey that have a listed phone number are re-approached by CATI. Nonrespondents without a listed number are not re-approached. We consider three groups: the CAPI respondents, the web/paper respondents, and the web/paper respondents supplemented by the CATI response in the re-approach. See Table 3.4.2 for the results of this pilot.

Gouweleeuw and Eding (2006) compared the three groups univariately with respect to age, gender, level of education and ethnic group. They found small differences with respect to age and gender. However, with respect to ethnic group they concluded that the composition of the CAPI response is more comparable to the population than those of the other two groups. With respect to level of education they found that all groups deviate from the population but in different ways. More high educations are found in the web/paper and web/paper/CATI group while low educations are overrepresented in the CAPI group. From the univariate comparisons it is, therefore, not immediately clear to which degree the overall compositions of the response are different.

Table 3.4.2 gives the R-indicators for the three groups with their corresponding confidence intervals and maximal absolute bias. The response rate in the paper/web group is considerably lower than in the other samples. However, the R-indicator of the web/paper group is significantly higher than in the other groups at the 5% level. This does not contradict the findings by Gouweleeuw and Eding (2006), but it is somewhat surprising as the response rate is much lower for this group.

Furthermore, despite of the lower response rate, the maximal absolute bias $\hat{B}_m$ does not differ considerably from that of the other groups. Hence, from these results we may conclude that with respect to the four selected auxiliary variables, there is no reason to favour the other groups to the web/paper alternative.

Table 3.4.2: Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for response for the Informal Economy survey by CAPI, web and paper only and web and paper with CATI follow-up.

Response	*n*	*Rate*	$\hat{R}$	$CI_{0.05}^{BT}$	$\hat{B}_m$	$\hat{E}_m$
CAPI	2000	56,7%	77,2%	(73,0 - 81,4)	10,1%	10,2%
Web/paper	2001	33,8%	85,1%	(81,5 - 88,7)	11,0%	11,2%
Web/paper + CATI	2001	49,0%	78,0%	(74,4 - 81,6)	11,2%	11,3%

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

3.5 Consumer Confidence Survey; pilot contact strategies 2006

The Consumer Confidence Survey (CCS) is a CATI household survey. First, each month a sample of addresses is drawn. Next, listed telephone numbers are linked to the addresses. In case no listed number is available for the address, the address is not forwarded to CATI and is filtered out. The original monthly sample size is chosen in such a way that approximately 1500 households remain that have a listed phone number. The CCS questionnaire deals with respondent opinions about the past, current and future state of the economy and is important input to the Consumer Confidence Index that is published each month. The target population of the survey consists of members of household cores, i.e. in case a household is called then the interviewer asks the head of the household or his/her partner to answer the questions. It is important to remark that due to the screening the actual target population consists of members of household cores that can be reached by a listed phone number. In the following the R-indicators thus correspond to a different population than the other studies. From research we know that having a listed phone number is a distinctive characteristic of a household.

In an earlier study, see Van der Grijn et al. (2006), the CCS samples of 2003 and 2004 were used to estimate and predict response probabilities and to construct contact strategies. For this purpose CATI fieldwork process data was linked to the samples. These data contained the interviewer, the number of calls and the outcome

of each call. Censored geometric regression models were fitted to the 2003 and 2004 data. The models were extended to account for unreachable households, i.e. households that have a zero contact probability during the fieldwork period. For reachable households, i.e. households with a positive contact probability, it was assumed that making contact in a certain contact attempt is independent from the outcome of another contact attempt. The final models formed the basis for the estimation of individual contact and cooperation probabilities. Van der Grijn et al. (2006) employed the estimated probabilities to construct contact strategies that aimed at a minimum number of call attempts given different constraints on the values of the R-indicator and response rate. In order to reach this goal the contact strategies were differentiated with respect to known auxiliary variables. Households with a low estimated contact probability received more calls than households with a high contact probability. The contact probabilities were estimated for four day parts: 10 - 5 pm, 5 - 7 pm, 7 - 8 pm and 8 - 10 pm. The auxiliary variables that were used in the estimation of response probabilities were household type, region of the Netherlands, survey month, age crossed with marital status, gender, number of jobs in the household core, average house value of zip code and ethnic background.

In November and December 2006, parallel to the CCS, a pilot was conducted using two independent samples. In both months the pilot sample sizes were 1500 like the normal CCS. In November the contact strategy was constructed according to the strategy proposed in Van der Grijn et al. (2006). In December the contact strategy was based on educated guesses about the most effective timing of calls given the auxiliary information about the household composition and the employment status. Both pilot contact strategies allocated only one or two call attempts in about 95% of the sampled cases. As a result the number of contact attempts was reduced by one third in November and by one fifth in December relative to the CCS.

Table 3.5.1: Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for response for the CCS pilot and the regular CCS in November and December

Response	*n*	*Rate*	$\hat{R}$	$CI_{0.05}^{BT}$	$\hat{B}_m$	$\hat{E}_m$
Pilot November	1493	53,7%	81,4%	(76,5 - 86,4)	8,7%	8,8%
Pilot December	1500	53,7%	79,6%	(74,8 - 84,4)	9,5%	9,7%
CCS November	1500	67,3%	84,1%	(79,7 - 88,6)	5,9%	6,1%
CCS December	1500	66,9%	81,0%	(76,1 - 85,8)	7,1%	7,3%

Table 3.5.1 shows the response rates, R-indicators, confidence intervals and maximal absolute bias for the months November and December of the CCS and pilot.

Detailed results of the 2006 pilots are given in Luiten et al. (2007). They observed that the response rates in the pilots were considerably lower than those in the CCS. Lower response rates for the pilots were expected as it was anticipated that the independence assumption for reachable households is too optimistic. Differences of this size were not expected, however. Luiten et al. conclude that the independence assumption of the success of subsequent calls given auxiliary variables does not hold; even if a class of unreachable households is included. A failure to contact a household on a certain day part, is informative of the success rate in a next call on the same day part on a different day. The pilot contact strategies were especially directed at improving the composition of the response. As the strategies resulted in lower contact rates than anticipated, they may also be less successful in improving the response compositions. Luiten et al. find evidence that with respect to the auxiliary variables household type, average age and region the pilot compositions of the response are less balanced than those in the CCS. However, the compositions improve for the number of jobs in the household and the ethnic background. For the variables average house value and gender the results are different for November and December. In November the representation of average house value is somewhat better, while in December the gender composition is closer to the sample. Hence, there is no clear overall improvement with respect to the selected auxiliary variables.

Table 3.5.1 shows that the values of the R-indicator of the pilot groups are lower than those of the CCS. However, the differences are not significant at the 5% level. We must note that only three of the auxiliary variables that were used in Luiten et al. (2007) are used in this paper. If we assemble all auxiliary variables that were used in the construction of the pilot contact strategies then differences in the values of the R-indicator are very small.

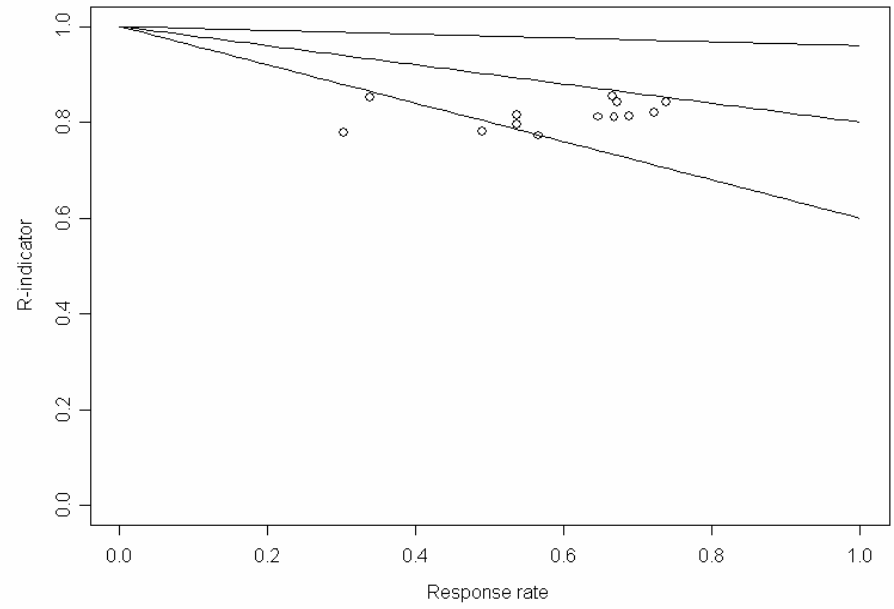
4. Maximal absolute bias, response-representativity functions and standard errors

Now that we have discussed the separate studies we can combine the results to see whether there are relations between sample size, response rate, maximal absolute bias and the R-indicator. We omit the LFS follow-up study as this study involved a design with unequal inclusion probabilities. In all studies the sample size leads to maximal root mean square errors that are only slightly larger than the maximal absolute bias. For this reason we restrict attention to response-representativity function $r_1(\gamma, \rho)$.

Figure 4.1 is a plot of the values of the R-indicator for the various studies versus the corresponding response rates. There is no clear relation between the R-indicator and the response rate for the present studies. A higher response rate does not necessarily imply a higher R-indicator. In Figure 4.1 we also depicted the response-representativity function $r_1(\gamma, \rho)$ for $\gamma = 1\%$, 5% and 10%. The response-

representativity function sets a lower bound to the R-indicator by requiring that the maximal absolute bias is smaller than γ .

Figure 4.1: The R-indicator versus the response rate for all studies except the LFS follow-up study. The lines correspond to the response-representativity function $r_1(\gamma, \rho)$ *for* $\gamma = 1\%$ *(top), 5% (middle) and 10% (bottom).*



All studies have R-indicators that lie below the 1% and 5% functions, but most R-indicators lie above or around the 10% line. This means that a maximal absolute bias of 5% is not guaranteed by any of the studies, but a maximal absolute bias of 10% is attained by many studies. Not surprisingly, Figure 4.1 also shows that a higher response rate implies that the risk of nonresponse bias is reduced.

We can look more closely at the maximal absolute bias by plotting $\hat{B}_m$ against the response rate., see Figure 4.2. This picture again confirms that a higher response rate corresponded to a lower risk of nonresponse bias in the present studies.

Finally, we move to the standard errors. In this paper we used a non-parametric bootstrap to approximate those errors. In Figure 4.3 the approximated errors are plotted against the sample size. The standard error is proportional to one over the square root of the sample size, as expected.

Figure 4.3 gives some guidance as to how large the sample size should be chosen in order to find significantly different R-indicators or maximal absolute biases. This only holds for different surveys with independent samples, since we can view the R-indicators as two independent random variables. In case different data collection strategies are applied to all units in one sample, then one needs to account for the dependence between the R-indicators. Clearly, more research is necessary to derive approximate closed forms for the standard error.

Figure 4.2: The maximal absolute bias $\hat{B}_m$ versus the response rate for the various studies except the LFS follow-up study.

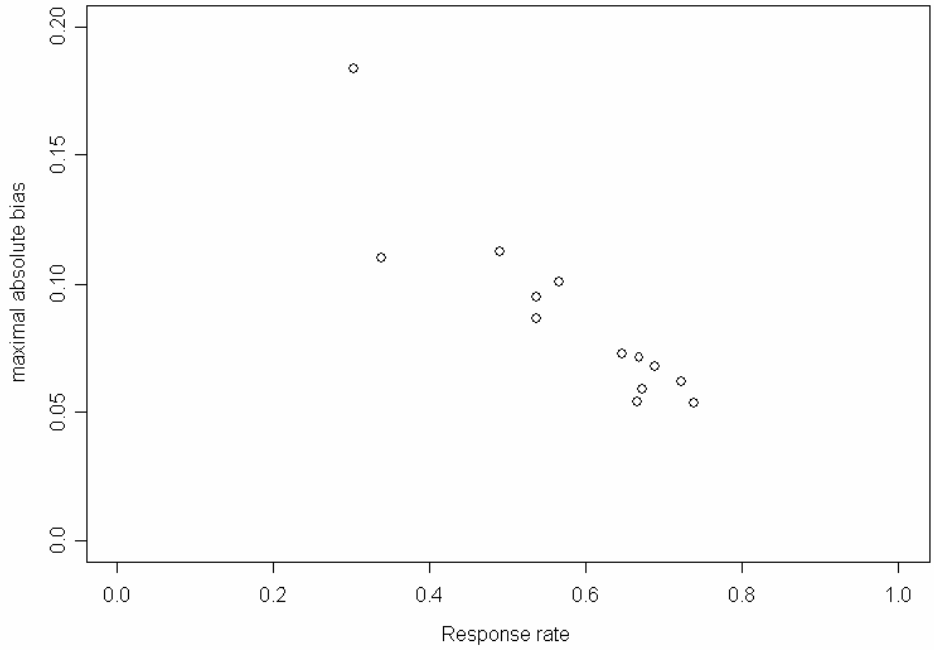
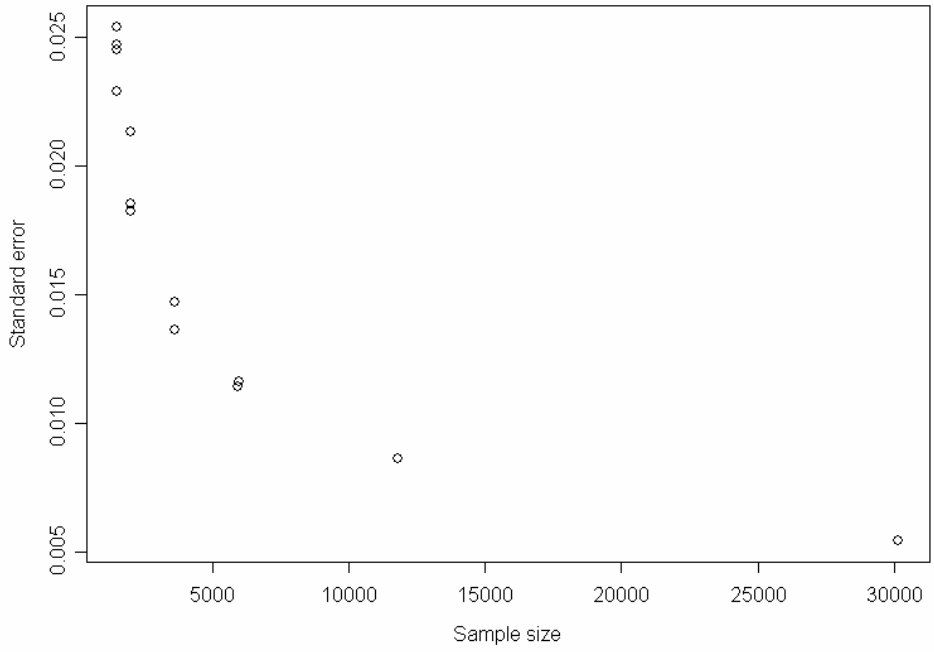


Figure 4.3: The standard error versus the sample size for the various studies except the LFS follow-up study.



5. Discussion

The main objective of this paper is to find empirical support that R-indicators are valuable tools in the comparison of different surveys and data collection strategies. We consider R-indicators to be useful if they confirm findings in elaborate analyses of studies that involve multiple surveys or strategies. We investigated an R-indicator proposed by Schouten and Cobben (2007). This R-indicator uses the sample variation in estimated response probabilities to measure the deviation from an optimal composition of response. We applied this R-indicator to a series of studies that were conducted at Statistics Netherlands in recent years, and that were thoroughly investigated by other authors. These are a follow-up study of the LFS, a study using incentives in the LFS, two pilots using mixed-mode data collection in the Safety Monitor survey and the Informal Economy survey, and a study employing different contact strategies in the Consumer Sentiments survey.

An important feature of R-indicators is their dependence on the set of available auxiliary variables. An R-indicator should always be interpreted in conjunction with this set. For this reason, we selected four auxiliary variables that were available for all studies: average age, degree of urbanization, household composition and ethnic background. By selecting the same set of auxiliary variables, we can directly compare the various studies.

The values of the R-indicator do not contradict the findings in the studies. For some studies, like the two LFS studies, the values clearly conform to the conclusions made in those studies. In some studies, results are more mixed, but this is mostly due to the restricted set of auxiliary variables that we used. We, therefore, conclude that R-indicators can be valuable tools but that it is important to constantly keep in mind that R-indicators must be viewed relative to the set of auxiliary variables.

The application of the R-indicator also showed that there is no clear relation between response rate and representativity of response. Larger response rates do not necessarily lead to a more balanced response. Not surprisingly, we do find that higher response rates reduce the risk of nonresponse bias. The higher the response rate, the smaller the maximal absolute bias of survey items.

An R-indicator is an estimator of a population measure of representativity and is, therefore, a random variable. One part of the randomness is due to the use of samples instead of the population as a whole, and another part of the randomness is due to the estimation of response probabilities. As a consequence, the R-indicators have a certain accuracy; the larger the sample the more precise the R-indicator. We, thus, need approximations of the standard errors of R-indicators in order to be able to compare different surveys or strategies. It was another objective of this paper to get some feeling about the standard errors of R-indicators.

Application to the selected studies learned that standard errors do decrease with increasing sample size as expected but they are still relatively large for modest sample sizes. For example for a sample size of 2000 we found a standard error of approximately 2%. Hence, if we assume a normal distribution, then the 95%

confidence interval has an approximate width of 8%. The study with the largest sample size of about 30000 units led to a standard error of approximately 0.5% and a corresponding 95% confidence interval that is approximately 2% wide. The standard errors are larger than we expected and also larger than we hoped for. To some extent this false expectation is due to the fact that the available auxiliary variables only explain a relatively small proportion of the variance in response behaviour. As a consequence shifts in R-indicators are relatively small as well. We do not have a good explanation, however, at this point, why the standard errors are relatively big.

This paper is a first empirical study of R-indicators and their standard errors. Much more theoretical and empirical research is necessary to get a grip on R-indicators and their properties. First, we did not consider survey items at all. Clearly, it is imperative that we do this in the future. However, as we already argued, R-indicators are dependent on the set of auxiliary variables. It can, therefore, be conjectured that, as for nonresponse adjustment methods, the extent to which R-indicators predict nonresponse bias of survey items is completely dependent on the missing-data mechanism. In a missing-data mechanism that is strongly non-ignorable, R-indicators will not do a good job. However, without knowledge about the missing-data mechanism no other indicator will. For this reason we constructed the notion of maximal absolute bias, as this gives a limit to nonresponse bias under the worst-case-scenario. A second topic of future research is a theoretical derivation of the standard error of the R-indicator used in this paper. We believe that the non-parametric bootstrap errors are good approximations. However, if we want R-indicators to play a more active role in the comparison of different strategies, then we need (approximate) closed forms. Third, and most importantly, we will need to investigate the relation between the selection and number of auxiliary variables and the standard errors of the R-indicator. We conjecture that the response-based R-indicators converge quite rapidly to the true population R-indicators when auxiliary variables are added. However, this conjecture needs to be confirmed by empirical studies.

References

Bethlehem, J.G. (1988), Reduction of nonresponse bias through regression estimation, Journal of Official Statistics, 4 (3), 251 – 260.

Brakel, J. van den, Berkel, K. van, Janssen, B. (2007), Mixed-mode experiment bij de Veiligheidsmonitor Rijk, Technical paper DMH-2007-01-23-JBRL, CBS, Heerlen.

Cobben, F., Janssen, B., Berkel, K. van, Brakel, J. van den (2007), Statistical inference in a mixed-mode data collection setting, Paper to be presented at ISI 2007, August 23- 29, 2007, Lisbon, Portugal.

Cobben, F., Schouten, B. (2007), The basic-question-approach – an application to non-respondents in the Dutch Labour Force Survey, Discussion paper, CBS, Voorburg.

Efron, B., Tibshirani, R.J. (1993), An introduction to the bootstrap, Chapman & Hall/CRC.

Fouwels, S., Janssen, B., Wetzels, W. (2006), Experiment mixed-mode waarneming bij de VMR, Technical paper SOO-2007-H53, CBS, Heerlen.

Gouweleeuw, J., Eding, H. (2006), De Informele Economie: analyse en vergelijking van de mixed-mode en face-to-face respons, Technical paper, CBS, Voorburg.

Groves, R.M., Heeringa, S.G. (2005), Responsive design for household surveys: tools for actively controlling survey nonresponse and costs, Technical paper, University of Michigan and Joint Program in Survey Methodology, USA.

Grijn, F. van der, Schouten, B., Cobben, F. (2006), Balancing representativity, costs and response rates in a call scheduling strategy, Paper presented at 17th International Workshop on Household Survey Nonresponse, August 28-30, Omaha, NE, USA.

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

Heij, R. de (2007), Measuring the underground economy at Statistics Netherlands: a progress report, Technical paper, CBS, Voorburg.

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

Kohler, U. (2007), Surveys from inside: An assessment of unit nonresponse bias with internal criteria, Survey Research Methods, 1, no 2, 55 – 67.

Luiten, A., Schouten, B., Cobben, F., Grijn, F. van der (2007), Balancing representativeness, costs and response rates in a call scheduling strategy, Discussion paper, CBS, Voorburg.

Särndal, C., Lundström, S. (2005), Estimation in Surveys with Nonresponse, Wiley Series in Survey Methodology, John Wiley & Sons, Chichester, England.

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

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

Stoop, I. (2005), Surveying nonrespondents, Field Methods 16, 23 – 54.

Voogt, R. (2004), I am not interested: nonresponse bias, response bias and stimulus effects in election research, PhD dissertation, University of Amsterdam, Amsterdam.

Wetzels, W., Schmeets, H. (2006a), Eerste adviesnota “Experiment Postzegel Beloning”, Technical paper, CBS, Heerlen.

Wetzels, W., Schmeets, H. (2006b), Tweede adviesnota “Experiment Postzegel Beloning”, Technical paper, CBS, Heerlen.