Journal of the Royal Statistical Society

SERIES A Statistics in Society

J. R. Statist. Soc. A (2013)
176, *Part* 1, *pp.* 169–189

Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction

Annemieke Luiten and Barry Schouten

Statistics Netherlands, Heerlen, The Netherlands

[Received March 2011. Final revision September 2012]

Summary. We used a tailored survey design to obtain a more representative response. Paradata from previous consumer sentiments surveys and register information were used to stratify the sample into groups that differed in contact and co-operation propensity. We approached an experimental sample of 3000 households with a Web–mail–computer-assisted telephone interviewing sequential mixed mode strategy. The choice of initial mode and the subsequent computer-assisted telephone interviewing approach were tailored to the expected contact and co-operation propensities of the sample units. In the computer-assisted telephone interviewing follow-up of non-respondents, co-operation was manipulated by assigning specific interviewers to specific sample units. Contact was manipulated by timing, spacing and prioritizing calls. The tailored fieldwork strategy was successful in significantly increasing representativeness, while maintaining the level of response and costs. Representativeness was determined by *R*-indicators.

Keywords: Adaptive design; Non-response bias; Paradata; Representative response; *R*-indicators; Tailored design

1. Introduction

For many years survey practitioners have struggled to attain high response rates as a safeguard against biased survey results. Various circumstances force us to rethink this strategy. Response rates in household surveys are becoming lower (de Leeuw and de Heer, 2002). More effort is required, so the costs of obtaining acceptable response rates rises (Starick and Steel, 2012). Also, response rates are not necessarily good indicators of non-response bias (Curtin *et al.*, 2000; Keeter *et al.*, 2000; Groves and Peytcheva, 2008; Heerwegh *et al.*, 2007).

Non-response may have different implications for different variables within one survey. The mechanisms causing non-response may be different for different groups. This implies that survey designs need to minimize potential bias across various domains of the key survey variables. Recent research addresses these issues. Groves (2006) advised to replace the blind pursuit of high response rates by informed pursuit, guided by knowledge of the relationship between response-stimulating measures, the groups that are sensitive to them and their influence on survey estimates.

Groves and Heeringa (2006) used the term ‘responsive design’ for survey designs where the status and the treatment of sample units are made dependent on an estimate of their contribution to the final survey result, relative to the costs of obtaining that result. Characteristic for the

Address for correspondence: Annemieke Luiten, Statistics Netherlands, PB 4481, NL-6401 CZ Heerlen, The Netherlands.
E-mail: a.luiten@cbs.nl

approach is that analyses of costs and errors calculated during fieldwork may lead to decisions and design alterations in mid-course (Groves and Heeringa, 2006; Mohl and Laflamme, 2007; Gambino *et al.*, 2010). Responsive survey designs are especially useful in settings where little is known about the sample beforehand or little information about the effectiveness of treatments is available from historic data. Sometimes, however, information is available on sample units from registers or prior panel rounds. Also, on-going surveys may yield information about the response propensities of groups of sample units. Such information can be used to design a tailored or differential approach before the survey starts.

Several researchers have studied the use of prior knowledge in designing differential designs. Wagner (2008) introduced the terms adaptive and dynamic design to describe differential survey designs tailored to the characteristics of sample units. Previous experience with similar sample units in similar surveys provides insight in how to treat each sample unit. Adaptive design allows treatment to vary with time, using rules specified before data collection. Peytchev *et al.* (2010) similarly described how experience in previous panels is used to prioritize sample units with a low predicted response propensity to diminish non-response bias.

Like responsive designs, adaptive and dynamic designs may also base their design on paradata such as interviewer observations on housing units or neighbourhood characteristics, and on process and administrative data produced as auxiliaries to survey data collection. Examples are the timing and outcome of call attempts, the nature of the interaction with household members, how long the interviews took, the reluctance of the interviewee and the mode of communication (Couper, 1998; Couper and Lyberg, 2005; Lepkowski *et al.*, 2010; Durrant *et al.*, 2011).

Whether designs are altered during fieldwork, or whether they are tailored to specific subgroups before fieldwork begins, what these approaches have in common is a differential fieldwork strategy, aimed at minimizing non-response bias and survey costs, while trying to maintain survey response at a level that is necessary for precise survey estimates (Schouten, 2010). Although calculating the response rate is a relatively straightforward task (e.g. American Association for Public Opinion Research (2008)), calculating costs and bias is far less so (Groves, 2004). Bethlehem (2002) defined non-response bias as the ratio of the covariance between the survey variable and the response propensity to the mean propensity. If there is no correlation between a target variable and response behaviour, the estimator is approximately unbiased. However, the stronger the relationship between a target variable and response behaviour, the larger the bias is. The size of the bias also depends on the amount of non-response.

However, one may encounter difficulties when using the formula that was proposed by Bethlehem (2002) to determine bias in a survey. The difference between respondents and non-respondents may be unknown for any number of target variables. Also, different sample estimates within the same survey can be subject to different non-response biases, making it difficult, if not impossible, to design a fieldwork strategy for minimizing overall bias.

These considerations led Schouten *et al.* (2009) to propose an alternative quality measure, the 'representativity' indicator or R-indicator, that measures the similarity between the response and the sample of a survey. The response that is obtained in a survey is defined to be representative if the individual response propensities are equal for all units in the population. Let ρ_X denote the response propensity function for variable X, say age, i.e. $\rho_X(x)$ is the probability of response of a population unit with value $X=x$. X in general is a vector of relevant auxiliary variables, e.g. age and sex. The distance between two vectors of response propensities ρ_1 and ρ_2 is expressed by using a function d:

$$d(\rho_1,\rho_2)=\sqrt{\left\{\frac{1}{N}\sum_U(\rho_{1,i}-\rho_{2,i})^2\right\}}, \tag{1.1}$$

where N is the population size, U the population and i the unit.

Given equation (1.1), the indicator of representativeness, or R-indicator, is defined as the distance between ρ_X and the survey response rate ρ:

$$R(X) = 1 - 2\,d(\rho_X, \rho_0) = 1 - 2\,S(\rho_X). \qquad (1.2)$$

d is the standard deviation S of the response propensities for different values of X. A transformation in equation (1.2) is made so that $R \in [0, 1]$. A value of 1 represents a perfect representative response, whereas a value of 0 indicates the largest possible deviation from a representative response.

The R-indicator describes the representativeness of the response given the whole of a vector of auxiliary variables. However, it is crucial to know which (vector of) X and which category within X are responsible for the deviation of representativeness when developing an adaptive or tailored fieldwork design, or when monitoring fieldwork in view of a responsive design (see for example Schouten *et al.* (2010)). For that, so-called partial R-indicators can be employed. Unconditional and conditional partial R-indicators are distinguished. Unconditional partial R-indicators describe the effect of each variable separately, whereas conditional partial R-indicators adjust the effect of one variable for the effect of other variables.

Schouten, Cobben and Bethlehem (2009) described how the R-indicator can be used as a tool for comparing different surveys, surveys over time or different data collection strategies and modes. The present study was set up as part of a large international research programme, the 'RISQ' project (`http://www.risq-project.eu`), aimed at developing R-indicators and studying their use in monitoring and controlling fieldwork. See Schouten and Shlomo (2010), Schouten, Cobben and Bethlehem (2009), Schouten, Morren, Bethlehem, Shlomo and Skinner (2009) and Shlomo *et al.* (2009a, b) for details.

In this paper, we describe how to obtain a more representative sample by using a tailored survey design. In an experimental setting, a standard uniform survey design was compared with a tailored adaptive design. Paradata from previous consumer sentiment surveys and information on sample units that is available in registers were used to predict the contact and co-operation propensities and at-home patterns of sample units in a new wave of the Survey of Consumer Sentiment (SCS). The tailored design sought to reduce the variability in response propensities of sociodemographic and socio-economic groups. It did so by stimulating response of sample units with low response propensity, while curbing those with high response propensity. This was done by assigning sample units to different modes (Web and mail) in an initial approach, and by differentiating the timing and number of computer-assisted telephone interviewing (CATI) contact attempts, and the interviewers assigned to specific sample units in the follow-up approach. Two constraints were important in developing the design: fieldwork should cost no more than for the standard SCS, and the response rate had to be maintained.

A major consideration in the design was to obtain a representative response in each step of the fieldwork: the first (Web–mail) wave, the CATI contact phase and the CATI co-operation phase. That meant that we would sometimes curb the chance of contact of sample units with a high contact propensity, while stimulating the same group with a low co-operation propensity.

In Section 2 we outline the design of the experiment. In Section 3 we describe the results of the experiment in terms of the response rates that were attained, the representativeness of the response in control and experimental groups, and the costs. Section 4 discusses the findings.

2. Method

We used the SCS as a vehicle for the experiment. The SCS is an on-going cross-sectional CATI survey, conducted among 1500 households of whom a listed telephone number can be found.

Questions may be asked of any person in the household core (the head of household or partner). The questionnaire takes about 8 min to complete. Questions are related to sentiments about the household's economic situation and expenditure. Fieldwork is conducted in the first 10 work days of each month. This experiment was conducted in October and November 2009, alongside the regular SCS, during the same 10-day fieldwork periods, with a similar sampling method and sample size, and the same interviewers. The regular SCS served as the control group for the experimental manipulations.

To achieve better representativeness at the same costs, we chose a mixed mode design for the experiment, in which a mail and/or Web round was followed by a CATI follow-up of non-respondents. Mail and Web questionnaires cost less to administer than CATI questionnaires and can reach respondents who are otherwise difficult to contact and/or convince to co-operate. As it is not feasible to conduct a mixed mode design as well as the CATI follow-up within 10 days, we did the Web–mail part of the survey a fortnight before the first 10 days of the month that is traditionally reserved for the SCS. Sample units received an advance letter with a Web link and/or a mail questionnaire. 1 week later, we sent a reminder. Another week later, we started the CATI follow-up.

The fieldwork strategy of the experiment was based on the response propensities of sample units in two data sets. First, historic SCS data were used to identify groups with low, medium and high contact and co-operation propensities in this telephone survey. The data set contained paradata about the response behaviour of about 18000 sample units. We determined for all sample units whether they were contacted and co-operated, how many attempts were needed and at what time these attempts were made. The propensity to respond in either a Web or mail mode was gauged from the paradata of another survey: the safety monitor in 2008 (Kraan *et al.*, 2009). In 2008 the previously single-mode annual computer-assisted personal interviewing survey was redesigned to a mixed mode Web, mail, CATI and computer-assisted personal interviewing design with a net sample size of 62803 respondents.

First the sample units were invited to complete a Web questionnaire. A mail questionnaire was available on request. Groups with a high propensity to co-operate in the SCS turned out also to have a high propensity to co-operate in Web surveys, whereas the opposite was true for groups with a low co-operation propensity in the SCS. The Web response in the safety monitor of the group with a high propensity to co-operate in the SCS was 31.3% , whereas the Web response in the group with a low propensity to co-operate in the SCS was 4.8%. In contrast, the mail response was relatively high in the group with low CATI co-operation propensity (13.5%), against 6.4% in the group with the high CATI co-operation propensity. Our conclusion was that we needed a Web questionnaire to cut the costs of the tailored design, but also a mail questionnaire to obtain co-operation from households with the lowest co-operation propensities.

2.1. Web–mail wave

With the aim of representativeness in the first wave in mind, sample units with a low co-operation propensity received a mail questionnaire, sample units with a high co-operation propensity received an invitation to the Web survey, and the middle groups were given a choice. The historic safety monitor data showed that we should not expect a substantial Web response from the group with the lowest co-operation propensity but could expect a relatively high mail response. This group mainly consists of elderly people, often without access to the Web (Statistics Netherlands, 2011), and (first-generation) ethnic minorities. We expected that the shorter, simpler advance letter of the mail-only condition and the short, simple paper questionnaire could persuade this difficult group to participate. The group with the highest co-operation propensity hardly used the mail option in the safety monitor. Because we could expect a relatively high response in the Web mode from this group, and because of cost considerations, we decided not to send this group a paper questionnaire. The groups between these two extremes were given the choice and received an invitation to the Web survey as well as a questionnaire on paper.

2.2. Telephone wave

Non-response from the first wave was followed up by CATI. We attempted to stimulate co-operation and contact for groups with low co-operation and contact propensities, and to curb those for groups with high co-operation and contact propensities.

To influence the chance of making contact, we defined different call schedules for the different contact propensity groups. Groups with a high contact propensity were primarily called during the day and were started later in the fieldwork period. Apart from freeing valuable capacity for evening calls, this was also cost effective, as daytime shifts are paid 20% less than evening shifts. Households with the lowest contact propensity, however, were to be called in every shift (morning, afternoon and evening), every day of the fieldwork period. The group with the low–middle propensity was called in the evening for the first two contact attempts. Subsequent attempts were made alternating between day and evening. The group with the high–middle contact propensity received the same default treatment as the control group: the regular SCS.

The rationale for these call schedules was based on analyses of paradata of historic SCS data and computer-assisted personal interviewing surveys (van Veen, 2004; Luiten *et al.*, 2007). The group with a high contact propensity consisted largely of elderly people, who can be reached during the day and usually need only one or two contact attempts: hence the decision to call during the day and to start fieldwork later. The strategy for the group with the lowest contact propensity, to call every day in every shift, obviously optimizes the chance of contact. The third group consisted largely of working households with younger children. There was a greater chance of contacting them in the evening but, if the first two attempts failed, we would spread the calls.

We manipulated the assignment of sample units to specific interviewers to influence the probability of co-operation. On the basis of SCS paradata, interviewers were classified according to their response rates achieved in 2008 and 2009 (82%, 76%, 72% and 66%). The best interviewers called the households with the lowest co-operation propensity. The interviewers with the lowest response rates called those with the highest propensity. And the group in between called on the middle group. The hypothesis was that low co-operation propensity would be stimulated, and high propensity curbed. If their workload permitted, interviewers could always call ‘easier’ addresses, but never ‘more difficult’ ones. The assignment of groups of addresses to groups of interviewers was handled by the CATI management system. See the Blaise CATI guide (Westat, 2004) for details on creating differential call schedules and allocating specific interviewers to specific addresses.

2.3. Selection of auxiliary variables for the tailored design

The objective in this experiment is to improve the representativeness of the survey response. But for which variables do we want the response to be representative? Auxiliary variables may relate to response behaviour, to one or more of the key survey variables or to the main publication domains. By the last we mean subpopulations that appear as marginals in publication tables and other publication statistics. When the assessment of response representativeness is used to compare multiple surveys, then it is necessary to select variables that relate to response behaviour, and are generally available in many surveys (Schouten, Cobben and

Table 1. Linked data to the SCS

Variable	*Categories*
Household level	
Ethnic group	Native Dutch, Morrocan, Turkish, Suriname–Netherlands Antilles, other non-western, other western, mixed and unknown: for the present analyses aggregated to native, ethnic minority, mixed and unknown
Sex	All male, all female, mixed, unknown
Average age of household core	15–30, 31–44 and 45–65 years; over 65 years, unknown
Type of household	Single, partners without children, partners with children, single parents, unknown
Postal code area level	
Degree of urbanization	Very strong, strong, moderate, low, not urban, unknown
Percentage non-western non-natives	Very high, high, average, low, very low, unknown
Average monthly income	Quartiles

Bethlehem, 2009). For use in tailored survey designs, it is important that variables relate either to the key survey variables or to the main publication domains (Bethlehem and Schouten, 2009).

Both SCS and experimental samples were linked to the social statistical database of Statistics Netherlands. This database is an integrated register based on registrations of all kinds of subjects. It contains administrative information on individuals, households, jobs, benefits, pensions and income. The sample addresses were matched on the basis of a precise combination of address, house number and date of contact. The variables that were used for this experiment are related to the key variables of the SCS. Table 1 summarizes them.

The registers contain information on individuals. As the SCS is a household survey, the individual level information was aggregated to household core level (head of household and partner). So, the variables ethnic group and sex have a category to indicate a mixture of the categories on the personal level (e.g. mixed native–ethnic minority). Some information is available only at the postal code level, which is quite a narrow geographical area around the sample unit's house.

Each variable has a category 'information not available'. This is concerned with linking sample units to registers. Registers are never entirely up to date: people move, dwellings are built or demolished, and unregistered people may lead to unavailable information at the individual, household or postal code level. Rather than treating these absent data as missing values, they are incorporated as meaningful values. The amount of absent data for each category is about 5%, with the exception of ethnic group, where it amounts to almost 11%.

2.4. Defining groups with differential contact and co-operation propensities

We determined which groups are over- or under-represented in the historical SCS data by calculating partial R-indicators. Contact and co-operation propensities were calculated separately because measures to stimulate contact may be different from measures to stimulate co-operation. See Table 2 for these partial R-indicators. These propensities were then projected on the sample units for the experiment. Co-operation was defined according to American Association for Public Opinion Research definition COOP2 (American Association for Public Opinion Research, 2008) as the number of complete and partial interviews divided by the number of interviews (complete plus partial) plus the number of non-interviews that involve the identi-

Table 2. Unconditional and conditional *R*-indicators for historic SCS data

	Unconditional R-indicators (× 1000)			*Conditional R-indicators (× 1000)*		
	Contact	*Co-operation*	*Response*	*Contact*	*Co-operation*	*Response*
Age (years)	*31*	*62*	*47*	*18*	*35*	*17*
<30	−23	3	−8	*5*	12	0
30–44	−13	29	16	13	27	2
45–64	8	18	22	5	29	12
⩾ 65	14	−51	−38	8	58	14
Sex	*33*	*53*	*66*	*7*	*1*	*7*
Male(s)	−28	−16	−35	3	0	3
Mixed	17	30	38	0	0	0
Female(s)	−8	−40	−41	2	0	2
Household composition	*35*	*61*	*75*	*10*	*13*	*19*
Single	−17	−40	−52	1	2	5
Partners, with children	9	12	15	4	2	5
Partners, no children	14	32	40	3	1	5
Single parent	−7	−1	−2	1	6	12
Ethnic group	*27*	*24*	31	*8*	9	*14*
Native Dutch	25	17	19	2	1	4
Foreign	−11	−14	−23	5	7	16
Mixed	1	10	8	1	0	1
Income in quartiles	*11*	*71*	*64*	*6*	*30*	*25*
< 1600	*5*	−55	−47	1	51	33
1600–1900	4	9	3	0	17	9
1900–2300	9	39	38	0	19	18
>2300	2	22	20	2	2	1
Urban density	*37*	*28*	*31*	*9*	*9*	*10*
Very strongly urban	−14	−8	−18	4	0	4
Strongly urban	3	3	4	1	3	4
Medium urban density	3	5	6	0	0	0
Low urban density	3	6	6	0	0	0
No urban density	7	−5	2	1	4	1
No information available	−33	−25	−23	0	0	0
% non-western foreigners in area	*12*	*16*	*18*	*2*	*10*	*8*
Less than 5%	7	−1	4	0	1	1
5–10%	−4	4	4	0	1	1
10–20%	−5	7	2	0	3	2
20% and more	−5	−2	−6	0	0	0
No information available	−6	−14	−16	0	6	3

fication of and contact with an eligible respondent (refusal and break-off plus other). Contact was defined according to American Association for Public Opinion Research definition CON1, which assumes that all cases of indeterminate eligibility are eligible.

We used a sum score to determine whether the expected contact and co-operation propensity was low, medium or high. For example, the partial *R*-indicators showed that elderly households, low income households, households of non-Dutch origin, households in a neighbourhood with a high percentage of people of non-Dutch origin and singles were less likely to participate than

Table 3. Co-operation and contact rates in groups with low, medium and high co-operation and contact propensity in historic SCS data

Propensity	*Co-operation (%)*	*Propensity*	*Contact (%)*
Low co-operation propensity	56.5	Low contact propensity	88.5
Low–medium co-operation propensity	67.5	Low–medium contact propensity	92.5
High–medium co-operation propensity	72.5	High–medium contact propensity	94.2
High co-operation propensity	78.5	High contact propensity	95.7

the other households. A household would receive a 'risk point' for each of these socio-demographic groups that it belonged to. The more risk points, the lower the co-operation propensity is, i.e. a low income elderly household had a lower co-operation propensity than a high income elderly household.

We did a similar exercise with chance of contact. Young households, singles or partners without children, households in highly urban areas, ethnic minority households and households living in neighbourhoods with a high percentage of ethnic minorities appeared to have a low contact propensity. Again, the more 'risk points' the lower the contact propensity is. On the basis of these analyses, each sample unit was classified as having a high, medium–high, medium–low or low contact propensity and having a high, medium–high, medium–low or low co-operation propensity. Table 3 shows non-contact and co-operation rates for these groups in historic SCS data.

2.5. *Fieldwork in the control group*

The regular SCS is a single-mode—telephone-only—survey. No information on the characteristics of the households is available beforehand. All households have an equal probability of being selected in the day batch, although households with whom appointments are made are prioritized. About 80% of the fieldwork is performed during the evening shifts. During daytime shifts, an interviewer is present to call appointments made for daytime. He or she uses spare time to phone other numbers. Interviewers are assigned to the SCS on the basis of availability, not ability or experience.

Supervisors determine daily whether the work is progressing well and whether it makes sense to call an address additional times. The decision is based on the overall response rate. An advance letter is sent some days before starting fieldwork, which is the same as in the experimental group. No incentives are given or promised, and no attempt is made to convert refusals in the regular survey or the experiment.

3. Results

3.1. *Response*

Table 4 shows response results for the regular SCS and the experiment.

In both months, the number of response cases was slightly higher in the experimental group, but the difference was not significant. Truly ineligible cases (0.8% in the control group and 0.5% in the experimental group) were collapsed with cases of unknown eligibility.

The ineligible cases turned out not to be households. The cases of unknown eligibility have disconnected telephone numbers, or numbers that do not belong to the sample address. As these cases are not followed up, they are called ineligible but are counted as a non-response.

Table 4. Response results in the SCS and experimental group

Category	*SCS*		*Experiment*	
	N	*%*	*N*	*%*
Ineligible	225	7.5	144	4.8†
Non-contact	196	6.5	183	6.1
Not present during fieldwork period	73	2.4	62	2.1
Not able (ill, dementia)	115	3.8	122	4.1
Language problems	40	1.3	26	0.9
Refusal	467	15.6	548	18.3‡
Response	1884	62.8	1915	63.8
Response Web–mail			1081	36.0
Response CATI			834	27.8

†$p < 0.001$.
‡$p < 0.05$.

The collapsed amount of ineligibility was significantly less in the experiment than in the control group ($\chi^2_{(1)} = 19.37$; $p < 0.000$). As the percentage of ineligibility is typically stable within the SCS across months (the mean percentage of ineligible sample units in 2009 is 8.4%, standard deviation 1.3), the lower percentage of ineligible addresses in the experimental group can only be attributed to mail or Web participation of households that we would otherwise not have been able to reach, because their number had been disconnected. Analysis of the paradata of the historic SCS shows that disconnected numbers are found mostly in households with a high non-contact and non-co-operation propensity. Sending a mail questionnaire to these addresses contributed substantially to a better representative response. Surprisingly, the number of refusals was higher in the experiment than in the control group ($\chi^2_{(1)} = 4.23$; $p < 0.05$). In Section 3.5 we shall elaborate on this result. No differences were found in the other non-response categories.

3.2. Predicted contact and co-operation

Before interpreting the results of the experimental manipulation, we evaluated whether the estimated co-operation and contact propensities proved to be predictive of the actual outcomes. Table 5 shows co-operation and contact rates of propensity groups in the regular SCS. The prediction proved to be quite accurate: so, the higher the predicted contact propensity, the

Table 5. Co-operation and contact rates in groups with low, low–medium, high–medium and high contact and co-operation propensities in the SCS control group

Propensity	*Co-operation (%)*	*N*	*Propensity*	*Contact (%)*	*N*
Low co-operation propensity	62.7	630	Low contact propensity	84.2	814
Low–medium co-operation propensity	68.4	493	Low–medium contact propensity	94.5	455
High–medium co-operation propensity	75.3	674	High–medium contact propensity	95.7	896
High co-operation propensity	79.2	1100	High contact propensity	96.9	732

higher the actual contact rates and, the higher the predicted co-operation propensity, the higher the actual co-operation rate.

3.3. Representativeness

In this section we examine the effect of the adaptive design on measures of representativeness, the R-indicator and partial R-indicators. Table 6 shows the value of the R-indicator for the response compared with the sample, as well as the R-indicator for each step in the fieldwork process: the representativeness of the eligible part of the sample, of those contacted from those eligible, of those able to co-operate from those contacted (sample units able to co-operate speak the language sufficiently well and are not too ill), and of those actually co-operating from those able to co-operate. As Table 6 shows, the R-indicator of each subsequent step is higher in the experiment than in the control group, with the exception of 'being able to co-operate'. Only for the R-indicator of response do the confidence intervals not overlap, however, and the one-sided null hypothesis $H_0 : R_{\text{control}} - R_{\text{pilot}} \geqslant 0$ is rejected at the 5% level.

Analysis of the partial R-indicators shows how the experimental manipulations affected sample composition (Table 7). Partial indicators ideally have values equal to 0. Large unconditional partial indicators show that a variable has a strong effect on representativeness. Negative values indicate under-representation; positive values over-representation. Large conditional partial indicators correspond to a large effect even after conditioning on the other auxiliary variables. Contrary to the situation for the unconditional partial indicators, a positive or negative sign cannot be assigned to conditional partial indicators. This is because the sign may be different for each subclass of X. In some subclasses a certain age of the head of the household may have a positive effect on the response whereas in others it may have a negative effect.

The following section illustrates the use of partial R-indicators in evaluating the effects of the experimental manipulation. We shall not go into detail for all variables and all columns but illustrate the interpretation with the variables sex and age. No estimator for the variance of a partial R-indicator has yet been developed (the methodology to calculate confidence intervals became available in the second half of 2012), so we cannot draw strong conclusions about the extent of the deviation from the representative response. The analysis illustrates, however, that all auxiliary variables in the experiment deviate less from representativeness than in the control group in the unconditional analysis, and also mostly so in the conditional analysis.

Table 6. R-indicators and 95% confidence interval CI for eligible, contacted, able and co-operating cases and overall response in the SCS and experiment

Case	*SCS*			*Experiment*		
	N	*R*	*CI*	*N*	*R*	*CI*
Sample	3000			3000		
Eligible	2774	0.84	(0.813–0.865)	2856	0.85	(0.856–0.905)
Contacted	2578	0.83	(0.801–0.856)	2673	0.89	(0.842–0.895)†
Able	2350	0.86	(0.832–0.881)	2463	0.85	(0.831–0.877)
Co-operating	1884	0.87	(0.842–0.896)	1915	0.89	(0.862–0.911)
Response	1884	0.77	(0.743–0.799)	1915	0.85	(0.821–0.872)‡

†$p < 0.10$.
‡$p < 0.05$.

Table 7 shows the results of the analysis of unconditional partial R-indicators for each step in the fieldwork process, and the conditional partial R-indicators for response only. For each auxiliary variable, the italic value is the composite contribution of the variable to representativeness; the other values describe the positive or negative contribution of the categories of the variable. For example, in the 'response' columns, the unconditional partial R-indicators show that the variable with the largest deviation from representativeness is sex, both in the control group and, to a lesser extent, in the experiment.

The category level information shows that households that are all male and all female are under-represented, whereas mixed gender households are over-represented. As single-sex households are mostly single households, and often either young or old, conditioning on the other auxiliary variables removes part of the influence of the variable, although sex is still the variable with the largest deviation in the control group.

Inspection of the category level variables shows that this is caused by the under-representation of single males. In the experiment, deviation from representativeness of sex has become smaller, and especially the males are far better represented. The unconditional R-indicators show that men in the experiment co-operated relatively well in the Web–mail first round and were also over-represented in overall co-operation. They were still under-represented in contact, however. This results in a nearly perfect representation of men in the response of the experiment.

Another illustration concerns age, which was a category that was explicitly targeted in the design. Young households were stimulated in the contact phase, whereas elderly households were stimulated in the co-operation phase. The partial R-indicators show that the adaptive design was successful in augmenting representativeness of this variable. The unconditional R-indicators at the (italic) variable level for contact, co-operation and finally response are lower for the experiment than for the control group. Unconditionally, the under-representation of the young households is lower in the experiment than in the control group (−12 *versus* −41). The conditional R-indicators show that the young households hardly differ from perfect representation in the final response. Co-operation of the elderly households was −26 in the control group but more representative (−7) in the experiment for the unconditional R-indicators.

3.4. Maximum bias

Schouten, Cobben and Bethlehem (2009) showed that, for any survey item y, the R-indicator can be used to approximate an upper bound to the non-response bias, in the case that y covaries maximally with the available (vector of) X. They used this upper bound to evaluate the effect under worst-case scenarios and to derive acceptable values for the R-indicator. The bias of y-variables that are not fully explained by X may be smaller or larger. The maximum bias provides intuition about how R-indicators relate to bias and is most useful in surveys with many y-variables. The upper bound of the bias is approximated by

$$\frac{|B(\hat{\bar{y}})|}{S(y)} \leqslant \frac{S(\hat{\rho})}{\hat{\bar{\rho}}} = \frac{1-R(\hat{\rho})}{2\hat{\bar{\rho}}} = B_m(\hat{\rho}, y). \tag{3.1}$$

The maximum bias in the control group is 0.18; in the experimental group it is 0.12. This means that the non-response bias is expected to be at most 18% of the standard deviation of any item in the control group and 12% in the experiment.

The lower maximum bias in the experiment indicates that the more representative response influenced the estimates. Whether the experimental estimates were less biased than those of the control group cannot be ascertained with certainty, however.

Table 7. Unconditional and conditional partial indicators for the SCS and experiment†

	Unconditional R-indicators, SCS (×1000)				*Unconditional R-indicators, experiment (×1000)*					*Conditional R-indicators (×1000)*	
	Eligible	*Contact*	*Co-operation*	*Response*	*Response Web–mail*	*Eligible*	*Contact*	*Co-operation*	*Response*	*Response, SCS*	*Response, experiment*
Age (years)	*59*	*52*	*35*	*58*	*62*	*43*	*33*	*21*	*36*	*24*	*13*
< 30	−26	−41	17	−25	−10	−18	−12	14	−10	25	3
30–44	−17	−13	14	1	−34	−11	−23	12	0	5	8
45–64	13	8	7	29	3	9	15	−8	22	21	6
⩾65	26	22	−26	−10	43	20	10	−7	−10	6	1
No information available	−42	−17	−2	−43	−28	−30	−8	3	−25	2	1
Sex	*142*	*108*	*41*	*209*	*118*	*105*	*71*	*58*	*134*	*31*	*12*
Male(s)	−18	−37	−1	−43	13	−5	−30	25	−3	40	6
Mixed	21	27	7	54	2	15	20	−9	30	27	1
Female(s)	−2	−8	−13	−38	0	−1	−5	−4	−28	26	8
No information available	−51	−27	5	−53	−28	−37	−9	0	−34	0	0
Household composition	*51*	*49*	*18*	*88*	*53*	*40*	*38*	*29*	*52*	*22*	*18*
Single	1	−28	0	−37	4	0	−27	10	−33	6	14
Partners, with children	15	20	2	32	37	18	14	−16	17	4	5
Partners, no children	11	20	7	37	−32	3	18	−3	25	4	4
Single parent	−16	−5	−16	−28	3	−13	−7	22	4	4	11
No information available	−45	−29	−5	−57	−20	−33	−11	4	−26	31	1
Ethnic group	*57*	*33*	*13*	*71*	*35*	*42*	*15*	*17*	*43*	*17*	*15*
Native Dutch	14	10	−6	15	1	12	5	−1	13	8	7
Mixed	8	3	8	23	21	7	3	−8	9	3	2
Foreign	−19	−16	7	−38	−2	−16	−11	14	−22	18	14
No information available	−51	−27	5	−53	−28	−37	−9	0	−34	0	0

(*continued*)

Table 7 *(continued)*

	Unconditional R-indicators, SCS (×1000)				*Unconditional R-indicators, experiment (×1000)*					*Conditional R-indicators (×1000)*	
	Eligible	*Contact*	*Co-operation*	*Response*	*Response Web–mail*	*Eligible*	*Contact*	*Co-operation*	*Response*	*Response, SCS*	*Response, experiment*
Income in quartiles (€)	*51*	*29*	*14*	*67*	*49*	*37*	*14*	*14*	*54*	*21*	*28*
< 1600	−1	−4	−5	−21	−10	3	2	−5	−23	13	24
1600–1900	5	4	−9	−7	−13	−1	−7	−7	−11	5	15
1900–2300	4	3	1	12	−4	5	5	0	14	2	8
>2300	10	7	9	31	36	9	4	11	30	22	28
No information available	−49	−28	4	−53	−28	−35	−10	−1	−34	0	2
Urban density	*18*	*30*	*16*	*32*	*28*	*31*	*15*	*26*	*24*	*14*	*14*
Very strongly urban	−6	−21	9	−19	5	−6	−13	18	0	0	8
Strongly urban	−11	−6	−1	−16	14	4	2	0	0	10	1
Medium urban density	6	0	7	13	−9	6	4	6	10	5	3
Low urban density	9	11	−9	12	5	5	4	−16	−2	2	6
No urban density	2	16	−6	10	−4	7	4	−5	5	1	1
No information available	4	5	0	5	−20	−28	−4	1	−21	1	2
% non-western foreigners in area	*47*	*38*	*35*	*60*	*33*	*34*	*18*	*18*	*25*	*15*	*5*
Less than 5%	14	17	−4	22	14	12	9	−7	8	2	1
5–10%	9	3	−2	8	6	8	−1	1	6	1	0
10–20%	−3	−4	−16	−15	−9	−1	−15	7	0	7	0
20% and more	−12	−23	30	−13	−12	−11	−3	14	−6	3	1
No information available	−41	−26	−1	−51	−24	−29	−3	−4	−22	10	1

†Values in italics are the composite contributions of the variable to representativeness.

3.5. Experimental manipulations

The experiment consisted of three manipulations: adding a mode, manipulation of the chance of contact in the CATI part and manipulation of the chance of co-operation, again in CATI. This section describes the effect of these measures on the subsequent distribution of responses. Response, co-operation and contact rates are used to illustrate the effect of the manipulations on representativeness.

3.5.1. *Adding a mode*

The number of ineligible cases was significantly lower in the experimental condition due to the Web–mail first round of data collection, which contributed significantly to the better representativeness of the experimental response. As Table 7 shows, males and single parents were better represented as a result of the added mode. Adding mail as a mode resulted in very high co-operation of elderly people in the first round. This did not lead to over-representation, however, owing to the curbing measures that were taken in the subsequent CATI round.

Because of the low predicted Web participation in the groups with low co-operation propensity, a mail questionnaire was added to the design. Table 8 shows that this measure succeeded in securing a fairly balanced first-round response. The response of the high co-operation propensity group that was given the Web option only lagged behind the high–medium propensity group. The latter had been given a choice of mode (odds ratio 1.635; standard error 0.101; $p < 0.001$). It even lagged marginally significantly behind the groups with low–medium (odds ratio 1.215; standard error 0.108; $p < 0.10$) and low co-operation propensity (odds ratio 1.215; standard error 0.109; $p < 0.10$). Compared with the Web–mail first round of the safety monitor however, where the response of the high propensity group was 38%, whereas the response of the low propensity group was 18%, the variability in response across groups is substantially less. When given the choice, 81% of households chose the mail questionnaire. The higher preference for the mail option is found repeatedly in research (e.g. Shih and Fan (2007) and Millar and Dillman (2011)).

3.5.2. *Manipulating chance of contact*

The higher R-indicator for the contact phase shows that the manipulations of contactability were successful in attaining a more representative contacted sample. Table 9 illustrates these findings with the contact rates for the SCS, compared with total contact rates for the experiment and the contact rates for the CATI part of the experiment separately. Table 9 shows that contact rates were somewhat higher in the experiment than in the SCS for the lower contact propensity groups, and somewhat lower for the high propensity group. A logistic regression analysis on

Table 8. Response on either Web or mail questionnaire by co-operation propensity

	Web or mail response (%)	*N*
Low co-operation propensity	35.1	304
Low–medium co-operation propensity	35.1	326
High–medium co-operation propensity	42.1	224
High co-operation propensity	30.8	227

Table 9. Contact rate per propensity category for the SCS, the experiment and the CATI part of the experiment

Contact propensity	*SCS (%)*	*N*	*Experiment*		*Experiment*	
			Total (%)	*N*	*CATI (%)*	*N*
Low contact propensity	84.2	640	87.1	657	79.4	413
Low–medium contact propensity	94.5	858	96.6	951	94.8	610
High–medium contact propensity	95.7	415	93.7	443	91.2	317
High contact propensity	96.9	794	95.3	804	91.7	459

contact rate with propensity group as factor showed a significant interaction between propensity group and experimental condition ($\text{Wald}_{(3)} = 10.39$; $p < 0.05$). Although the variability in contact rate in the experiment was reduced, compared with the SCS control group, we failed to obtain representative contact in the CATI part of the experiment. After the first-wave Web–mail response, the remaining group of non-respondents in the group with the lowest contact probability lagged behind considerably in contact rate, even with one call in every shift, every day.

3.5.3. Manipulating chance of co-operation

The chance of co-operation was manipulated by having the best interviewers call addresses with the highest chance of refusal, whereas the addresses with the highest chance of co-operation were called by the least successful interviewers. Analysis of the fieldwork verified that the fieldwork strategy was applied as planned and that the mean level of interviewer capacity was comparable in the experiment and the SCS.

Although the number of co-operating sample units was slightly higher in the experiment, the co-operation rate was somewhat lower ($\chi^2_{(1)} = 4.23$; $p < 0.05$).

The R-indicator for co-operation (Table 6) showed that there was hardly any difference in the distribution of participation for the experiment and control group. Table 10 illustrates this finding with the co-operation rates per propensity group for the experiment, its CATI part and the SCS. Like the findings concerning contact, co-operation in the experiment is higher for the two groups with the lower co-operation propensity and lower for the two groups with the higher

Table 10. Co-operation rate by co-operation propensity for the SCS, the CATI part of the experiment and the experiment total

Co-operation propensity	*SCS (%)*	*N*	*Experiment*		*Experiment*	
			Total (%)	*N*	*CATI (%)*	*N*
Low co-operation propensity	62.7	630	65.1	619	43.8	392
Low–medium co-operation propensity	68.4	493	71.4	639	52.8	415
High–medium co-operation propensity	75.3	674	72.8	744	50.3	418
High co-operation propensity	79.2	1100	74.7	995	62.8	691

Table 11. Response by co-operation propensity for the experiment and the SCS

	Response for the experiment				*Response for the SCS*			
	Low (%)	*Low–medium (%)*	*High–medium (%)*	*High (%)*	*Low (%)*	*Low–medium (%)*	*High–medium (%)*	*High (%)*
Not able (ill, not present)	12.3	6.9	2.5	2.7	12.1	7.3	4.5	3.5
Language problems	2.4	1.1	0.5	0.0	4.3	1.0	0.6	0.2
Refusal	15.7	13.3	20.8	21.2	14.4	16.8	16.9	15.5
N	619	639	744	995	630	493	674	1100
Co-operation rate COOP2	65.1	71.4	72.8	74.7	62.7	68.4	75.3	79.2
Co-operation rate COOP3	78.4	81.1	76.2	76.9	78.2	76.4	79.8	82.5

co-operation propensity. A logistic regression on co-operation rate with propensity group as factor showed a significant interaction between propensity group and experimental condition ($\text{Wald}_{(3)} = 10.21$; $p < 0.05$). The interaction signified that having the best interviewers call the hardest cases did not bring about the expected rise in co-operation. But having the lesser interviewers call the easy cases brought about a significant decline in co-operation in this group. The difference was not enough, however, to bring about the desired change in co-operation representativeness.

Some light is shed on the issue of why the best interviewers could not secure a higher co-operation rate by studying Table 11, which shows response results for the experiment and SCS by co-operation propensity.

The first co-operation rate in Table 11, which is also shown in Table 10, shows again that, as predicted, the co-operation rate is higher when the estimated participation propensity is higher, in both the experiment and the control group. Prediction of participation propensity was based on the calculation of co-operation according to COOP2 (American Association for Public Opinion Research, 2008), as co-operation of contacted eligible sample units. However, as Table 11 shows, prediction of co-operation appears to be strongly correlated with the ability to participate, and with the existence of language problems. In the experiment, the percentage of sample units who could not participate ranges from 2.7% in the group with high co-operation propensity to 12.3% in the group with low co-operation propensity, and language problems range from absent to 2.4%. The control group shows similar patterns. The second co-operation rate in Table 10 shows co-operation of eligible, contacted and able sample units (COOP3). With this calculation, the propensity differences all but disappear. If the only difference between groups in the level of co-operation is related to the ability to co-operate, a different intervention is needed, e.g. using translated questionnaires and bilingual interviewers.

3.6. Costs

One of the aims of this experiment was to raise the quality of data while maintaining or ideally lowering costs. We took two potentially cost saving measures: the Web round and more daytime interviewing, the latter because interviewers at Statistics Netherlands receive 20% more pay for working in the evenings. Mail questionnaires are about 50% cheaper than CATI interviewing but, in a mixed mode experiment, a part of the sample will be addressed in both modes, thereby adding to the costs, unless the mail response is substantial.

Table 12. Itemized total costs for the SCS and experiment

	Results for the SCS			*Results for the experiment*		
	N	*Rate*	€	*N*	*Rate*	€
Postage advance letters	3000	0.36	1071	1032	0.36	368
Postage advance letters + mail questionnaire				1968	0.69	1358
Reminders				2318	0.36	828
Printing costs for mail questionnaire				1968	2.16	4244
Interviewer hours at evening rate	334	36	12024	159	36	5724
Interviewer hours at daytime rate	105	30	3150	86	30	2573
Data entry hours				48	33	1571
Total			16245			16665

To compare the costs of the experiment with those of the control group, we considered the costs of observation and data processing, notably postage and printing costs for the advance letters, reminders and paper questionnaires (including labour and machine depreciation), data entry for the paper questionnaires and the interviewers' time, differentiated by shift. Table 12 shows the total costs and the items contributing to total costs.

As Table 12 shows, the costs of the experiment are marginally higher (2.6%) than those of the SCS in those 2 months.

4. Summary and discussion

We described a tailored fieldwork strategy to obtain a better representative sample at comparable response and cost levels. The results showed that the tailored fieldwork strategy was successful in maintaining the level of response, while significantly augmenting representativeness, even within the very short fieldwork period of the SCS. A longer fieldwork period would have provided more possibilities to vary the number and spacing of calls, and to make use of the paradata becoming available during fieldwork. The tailored design was slightly more expensive. We shall comment on the expense first, and then discuss the findings regarding representativeness.

The experiment was somewhat more costly than the regular SCS because of several circumstances. First, 81% of people who were given the choice between a Web and mail questionnaire chose mail. This meant that more money had to be spent on data entry than expected. Second, although the paper questionnaire, including extra postage and subsequent data entry, costs only about half of what a CATI sample unit in the control group costs, there were many sample units who did not respond in the cheaper mode and had to be called in the more expensive CATI mode. This resulted in a higher per-unit cost. The experiment was designed to incorporate a larger number of (cheaper) day calls, targeting groups with a high contact propensity. Although 11% more daytime calls were made in the experiment than in the control group (35% *versus* 24%), this difference was not enough to offset the mechanisms that were described above.

The addition of the paper questionnaire was the obvious cause of the relatively high costs. We could easily have achieved lower costs by using only a Web approach, followed by a CATI non-response follow-up. We did not do this because we sought representativeness within each step of the fieldwork. For the same reason we stimulated co-operation in some of the same groups in which we curbed contact. For example, an elderly person who refuses is a different person from an elderly person who cannot be contacted and may have a differential influence on potential bias. By adding the mail questionnaire we succeeded in obtaining a far more balanced first-wave response than if we had used a Web-only approach.

Representativeness was measured with the R-indicator, which measures the distance between the mean response level and the response of subgroups defined by the auxiliary variables in the research. Partial R-indicators were first used to examine which groups were under- and over-represented in the SCS, and later to study the effect of the experimental manipulations on representativeness within auxiliary variables. The choice of auxiliary variables is paramount in designing tailored designs. First, it is imperative that variables are known for all sample units. Second, they need to be related to key variables of the survey and main domains of interest in publications. The broader the subject of a survey, the more general the auxiliary variables need to be (Bethlehem and Schouten, 2009). The choice of auxiliary variables also influences the conclusions that can be drawn about representativeness. Representativeness is not an absolute given but depends on the auxiliary variables in the model. A survey could have a very high R-indicator and still contain biases on variables for which correlating paradata or other auxiliary variables are not available.

The auxiliary variables that were chosen in this experiment all relate to the key variables of the SCS. The finding that the response composition was more representative with regard to these variables can be generalized to variables that were not part of the design. Schouten and Cobben (2012) show that the design in this experiment was successful in reducing non-representative response on variables other than those used to differentiate subgroups, specifically, ownership of a company car, business type of the person in the household with the largest job and job number and sizes in the household. Although these variables were not used in the tailored survey design, they are associated with the selected design variables age, ethnicity, income, type of household and urbanization. If tailored or other adaptive survey designs are to be promising extensions of sampling designs, then the indicator values should also be better for variables that were not involved in the adaptation.

Statistics Netherlands is allowed by law to use registers to link to survey results. However, the number of register variables is limited. If matching with register data is impossible, or if the available variables are not related to key variables, the only option is to resort to paradata like observations of sample units and/or their environment that are expected to relate strongly to the main survey variables.

A key question is whether the higher representativeness that was found in the experimental group is due to the experimental manipulation of response propensity, the introduction of a second mode or the longer fieldwork period. The manipulations are partly confounded and the independent effects of each of the treatments cannot be disentangled completely. Undoubtedly, adding a mode helped in improving representativeness: we have shown that the number of households that could not be approached as a result of disconnected telephone numbers was significantly reduced in the experimental group, bringing in households with a low response propensity. The mode offered was differentiated according to response propensity, as a result of which the first-wave response was fairly balanced for the four propensity groups. Consequently, the sample for the CATI reapproach was also balanced. The differential response in the CATI wave for the four propensity groups is therefore also the result of the experimental CATI manipulations. Additional support for the contention that the mere introduction of a second mode does not in itself result in a better representative response is found in the analysis of several redesigns of Statistics Netherlands surveys, where (computer-assisted personal interviewing) unimode designs were replaced by mixed mode designs. In all of these redesigns, adding a mode led either to a slight reduction in representativeness or to a comparable level, but never to augmented representativeness (Banning *et al.*, 2011; Cobben, 2011). The longer fieldwork period did not lead to higher overall contact rates, compared with the regular SCS. Although in the experimental group higher contact rates were realized for the groups with the lowest contact propensity, lower rates were attained for the groups with the highest contact propensity, thereby reducing variability in contact rates across the groups. The longer fieldwork period in itself cannot account for the differentiation in these results.

The representativeness of the experimental group was augmented especially as a result of more representative eligible and contacted cases. The manipulation of co-operation had less effect. This result may have been influenced by the introduction of a Web–mail first round, filtering away the ‘easiest’ respondents, leaving the interviewers to deal with a relatively uniform difficult group of initial non-respondents. In other words, the co-operation propensity of the remaining group may have been different from the expected propensity. With increasing experience in mixed mode survey methodology it will be possible to gauge the influence of different modes on co-operation propensity of different groups. Another explanation is to be found in the definition of co-operation that we used in this study. When co-operation was defined conditionally on the ability to co-operate, co-operation propensity was not predictable with the auxiliary variables that were available in this study. As a result of this finding, Luiten and Cobben (2010) analysed a large database that consisted of numerous surveys with a variety of topics, lengths, modes and sampling types, and containing an extensive number of auxiliary variables. Again, co-operation conditional on ability to co-operate could not be predicted. That is not to say that co-operation cannot or should not be influenced, but rather that co-operation has other underlying dimensions than socio-economic or demographic correlates. Present attempts to find and incorporate paradata that relate both to response propensity and to substantive variables may fill this gap (Schouten, 2010; Kreuter *et al.*, 2010).

In this paper we set out to show that it is possible to attain a more representative sample while keeping response and cost levels the same. We found that we could. Far more research is needed, however. We need more experience with design variations to find out whether it can be done even better, and to learn what works best for which groups. The field of adaptive design is only just starting.

Groves (2006) set in motion an awareness among survey practitioners that we need to think in terms of non-response bias as much as in terms of response rates. This can only be accomplished when survey designs are aimed at reducing bias, which in turn means that sample units should not be treated in a uniform fashion. Adaptive, tailored or responsive survey designs are a means to accomplish this end. R-indicators may help in drafting these designs by selecting subsets of cases that need extra attention, in monitoring the fieldwork and in gauging the maximum bias in a given survey. Future research should focus on determining which groups are susceptible to which treatments, and how differential treatment relates to the reduction of non-response bias.

Acknowledgements

The experiment in this paper was part of the project RISQ: ‘Representativeness indicators for survey quality’, financed by the seventh framework programme of the European Union, Co-operation Programme, Socio-economic Sciences and the Humanities, Provision for Underlying Statistics. The authors thank Willem Wetzels for his help in designing the experiment and Hans Schmeets for his helpful comments on earlier drafts.

References

American Association for Public Opinion Research (2008) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, 5th edn. Lenexa: American Association for Public Opinion Research.

Banning, R., Cobben, F. and en Leufkens, K. (2011) Kwaliteitsmeting van EBB Kernvariabelen in Eerste Peiling 2010 (Measuring quality of substantive Labour Force Survey variables, first wave 2010). *Internal Report.* Statistics Netherlands, The Hague.

Bethlehem, J. (2002) Weighting non-response adjustment based on auxiliary information. In *Survey Non-response* (eds R. M. Groves, D. Dillman, J. Eltinge and R. Little). New York: Wiley.

Bethlehem, J. and Schouten, B. (2009) Representativeness indicators for measuring and enhancing the quality of survey response. *RISQ Deliverable 9.* (Available from `www.risq-project.eu/papers`.)

Cobben, F. (2011) Responsgedrag in de Gezondheidsenquête 2010 (Response behaviour in the Health Survey 2010). *Internal Report*. Statistics Netherlands, The Hague.

Couper, M. (1998) Measuring survey quality in a CASIC environment. *Jt Statist. Meet. American Statistical Association, Dallas*. (Available from `http://www.amstat.org/Sections/Srms/Proceedings`.)

Couper, M. and Lyberg, L. (2005) The use of paradata in survey research. In *Proc. 55th Sessn International Statistical Institute, Sydney*.

Curtin, R., Presser, S. and Singer, E. (2000) The effects of response rate changes on the index of consumer sentiment. *Publ. Opin. Q.*, **64**, 413–428.

Durrant, G. B., D'Arrigo, J. and Steele, F. (2011) Using paradata to predict best times of contact, conditioning on household and interviewer influences. *J. R. Statist. Soc.* A, **174**, 1029–1049.

Gambino, J., Laflamme, F. and Wrighte, D. (2010) Responsive design at Statistics Canada: development, implementation and early results. *21st Int. Wrkshp Household Survey Non-response*, *Nürnberg, Aug. 30th–Sept. 1st.*

Groves, R. M. (2004) *Survey Errors and Survey Costs*. Hoboken: Wiley.

Groves, R. M. (2006) Non-response rates and non-response bias in household surveys. *Publ. Opin. Q.*, **70**, 646–675.

Groves, R. M. and Heeringa, S. G. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. *J. R. Statist. Soc.* A, **169**, 439–457.

Groves, R. and Peytcheva, E. (2008) The impact of non-response rates on non-response bias: a meta analysis. *Publ. Opin. Q.*, **72**, 167–189.

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

Keeter, S., Miller, C., Kohut, A., Groves, R. M. and Presser, S. (2000) Consequences of reducing non-response in a national telephone survey. *Publ. Opin. Q.*, **64**, 125–148.

Kraan, T., van den Brakel, J., Buelens, B. and Huys, H. (2009) Social desirability bias, response order effects and selection effects in the new Dutch Safety Monitor. *Federal Committee on Statistical Methodology Res. Conf., Nov. 2nd–4th.*

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M. and Raghunathan, T. E. (2010) Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *J. R. Statist. Soc.* A, **173**, 389–407.

de Leeuw, E. and de Heer, W. (2002) Trends in household survey non-response: a longitudinal and international perspective. In *Survey Non-response* (eds R. M. Groves, D. Dillman, J. Eltinge and R. Little). New York: Wiley.

Lepkowski, J., Axinn, W., Kirgis, N., West, B., Kruger Ndiaye, S., Mosher, W. and Groves, R. (2010) Use of paradata in a responsive design framework to manage a field data collection. *Survey Methodology Working Paper 10-012*. National Survey of Family Growth.

Luiten, A. and Cobben, F. (2010) Predicting co-operation in survey research. *21st Int. Wrkshp Household Survey Non-response*, *Nürnberg*.

Luiten, A., Schouten, B. and Cobben, F. (2007) Experiments in balancing response, representativeness and costs in a CATI survey. *18th Int. Wrkshp Household Survey Non-response, Southampton*.

Millar, M. and Dillman, D. (2011) Improving response to web and mixed-mode surveys. *Publ. Opin. Q.*, **75**, 249–269.

Mohl, C. and Laflamme, F. (2007) Research and responsive design options for survey data collection at Statistics Canada. *Proc. Am. Statist. Ass.*, 2962–2968.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010) Reduction of non-response bias in surveys through case prioritization. *Surv. Res. Meth.*, **4**, 21–29.

Schouten, B. (2010) Dynamic adaptive survey designs; from paradata to data. *21st Wrkshp Household Survey Non-response Nürnberg, Aug. 30th–Sept. 1st.* (Available from `http://cbsh1sps/sites/TmoMeth/WaarPers/03%20Waarnemingsmethoden%20personen/2010/NR%20workshop%20paper%20-%20Schouten.pdf`.)

Schouten, B. and Cobben, F. (2012) Non-representative survey response: to balance, to adjust or both? *Discussion Paper*. Statistics Netherlands, The Hague.

Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the representativeness of survey response. *Surv. Methodol.*, **35**, 101–113.

Schouten, B., Luiten, A., Loosveldt, G., Beullens, K. and Kleven, Ø. (2010) Monitoring and changing data collection through R-indicators and partial R-indicators. *RISQ Deliverable*. (Available from `www.risq-project.eu/papers`.)

Schouten, B., Morren, M., Bethlehem, J., Shlomo, N. and Skinner, C. (2009) How to use R-indicators? *RISQ Deliverable*. (Available from `www.risq-project.eu/papers`.)

Schouten, B. and Shlomo, N. (2010) Indicators for representative response. *Quality in Official Statistics Conf., Helsinki, May*.

Shih, T.-H. and Fan, X. (2007) Response rates and mode preferences in web-mail mixed-mode surveys: a meta-analysis. *Int. J. Internet Sci.*, **2**, 59–82.

Shlomo, N., Skinner, C., Schouten, B., Bethlehem, J. and Zhang, L.C.(2009a) Statistical properties of representativity indicators. *RISQ Deliverable*. (Available from www.risq-project.eu/papers.)

Shlomo, N., Skinner, C., Schouten, B., Carolina, T. and Morren, M. (2009b) Partial indicators for representative response. *RISQ Deliverable*. (Available from www.risq-project.eu/papers.)

Starick, R. and Steel, J. (2012) Does increased effort lead to a less representative response?: selected case studies from the Australian Bureau of Statistics. *Eur. Conf. Quality in Official Statistics*, *Athens, May 29th–June 1st*.

Statistics Netherlands (2011) Media en ICT: Gebruik Televisie, Krant, PC en Internet (Media and ICT: use of television, paper, PC and Internet). Statistics Netherlands, The Hague. (Available from http://statline.cbs.nl/StatWeb/publication/?DM=SLNL&PA=70655ned&D1=76-77,90&D2=37-41,52&D3=a&HDR=T&STB=G1,G2&VW=T.)

van Veen, P. (2004) Contact and participation probabilities in CAPI and CATI surveys. *Masters Thesis*. University of Utrecht, Utrecht.

Wagner, J. (2008) Adaptive survey design to reduce non-response bias. *PhD Thesis*. University of Michigan, Ann Arbor.

Westat (2004) *Blaise CATI Guide*. Rockville: Westat.