ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2016)
178, *Part* 4, *pp.* 945–961

Selection error in single- and mixed mode surveys of the Dutch general population

Thomas Klausch,

Utrecht University and Statistics Netherlands, The Hague, The Netherlands

Joop Hox

Utrecht University, The Netherlands

and Barry Schouten

Utrecht University and Statistics Netherlands, The Hague, The Netherlands

[Received November 2012. Final revision November 2014]

Summary. This study compares the extent of selection error (non-response and coverage error) evoked by the four major contemporary modes of data collection (face to face, telephone, mail and Web) and three sequential mixed mode designs (telephone, mail and Web with face-to-face follow-up) for the case of the Dutch Crime Victimization Survey. Sociodemographic characteristics and target variables from the survey serve as benchmark variables. A special two-wave experimental design allows studying design differences in selection error on Crime Victimization Survey variables independently from differences in measurement error. Despite large differences in response rates, only small or no differences in selection error between the four single-mode designs are found on both types of variable. We observe cases when the error is enlarged or mitigated in the mixed mode designs despite the fact that the designs yielded large response increases. Our results question the use of response rates to motivate the choice of mode and use of mixed mode surveys.

Keywords: Coverage error; Data quality; Mixed mode surveys; Non-response error; Representativeness; Survey modes

1. Introduction

The extent of expected coverage and non-response error is an important quality criterion of a survey design (Biemer and Lyberg (2003), page 39, and Groves (1989), pages 15–18). The 'survey mode' might play a crucial role in determining the size of these errors. Clearly, modes determine applicable sampling frames (e.g. to telephone households) and thus affect population coverage of a survey. Furthermore, modes are related to response rates (RRs), which might be indicative of the extent of non-response error evoked by a survey design. Meta-analyses have reported that face-to-face (FTF) modes yield higher RRs than telephone or mail modes (Hox and De Leeuw, 1994) and that Web modes evoke lower RRs than the other major modes (Manfreda *et al.*, 2008; Shih and Fan, 2008). Therefore, FTF surveys might result in smaller non-response error than telephone and mail, and non-response error in Web surveys might even be stronger.

Address for correspondence: Thomas Klausch, Department of Methodology and Statistics, Utrecht University, Postbus 80140, Utrecht 3508 TC, The Netherlands.
E-mail: l.t.klausch@uu.nl

 0964–1998/15/178945

Unfortunately, FTF surveys commonly incur the greatest costs during data collection. So-called 'sequential mixed mode' (MM) surveys can cope with this problem by offering inexpensive modes first (e.g. Web or mail) and later follow up only the non-respondents by a more expensive mode, such as FTF (Biemer and Lyberg (2003), page 106, De Leeuw *et al.* (2008), De Leeuw and Hox (2011), De Leeuw (2005), Dillman and Christian (2005) and Dillman *et al.* (2009b), pages 300–310). In doing so, high RRs (e.g. at the level of FTF) can typically be preserved. A prominent example is the American Community Survey, in which mail non-respondents are followed up by telephone and FTF interviews. This strategy, though participation is mandatory, strongly increases RRs by the mode switches to over 90% (US Census Bureau, 2010). Several further studies also report increases in response, when mail or Web modes are combined with telephone in sequential designs (Dillman *et al.* (2009a), Eva *et al.* (2010), Fowler *et al.* (2002), Greene *et al.* (2008) and Link and Mokdad (2006); see Millar and Dillman (2011) for a Web–mail comparison). For example, Link and Mokdad (2006) reported up to 39 percentage points additional response in telephone follow-ups of Web non-respondents. Similar results (35 percentage points) were reported by Dillman *et al.* (2009a); see De Leeuw (2005) and Lynn (2013) for further examples. These effects might imply that MM surveys can also mitigate coverage and non-response error *vis-à-vis* single-mode (SM) designs with lower RRs. Keeping these errors low is referred to as a key advantage of MM surveys in many of these studies.

However, using response and coverage rates as indicators for the risk of non-response and coverage error is based on the assumption that the quantities are inversely related. This conjecture appears contestable on theoretical and empirical grounds (Wagner, 2012). Statistically, RRs are not directly related to the size of non-response bias (see, for example, Bethlehem (1988)). Empirically, it has been shown by extensive meta-analysis that non-response bias and RRs are only weakly associated (Groves and Peytcheva, 2008; Groves, 2006). Basing design decisions on response and coverage rates, therefore, poses a potential fallacy. However, only few studies have assessed both errors empirically for multiple SM and MM surveys.

The present study addresses this issue by a large-scale ($n = 8800$) mode experiment within the Dutch Crime Victimization Survey (CVS). We consider non-response and coverage errors as compound, because practitioners generally are concerned most about the 'net effect' of both sources of error when taking design decisions. This compound is referred to as the 'selection error' of a design (and is also known as non-observation error; Groves (1989), pages 15–18). In doing so, we address two research questions (RQs):

(a) do mode-specific differences in RRs reflect differences in selection error (RQ 1);
(b) are RR increases of sequential MM designs indicative of a reduction in selection error (RQ 2)?

In particular, we first study selection error of the four major contemporary modes (FTF, telephone, mail and Web) by using a split-ballot experimental design and, subsequently, illustrate the effect of three sequential MM designs on selection error of the SM designs (telephone–FTF, mail–FTF and Web–FTF). Mode experiments of this type normally confound mode differences in selection error with mode differences in measurement error (Jäckle *et al.*, 2010; Vannieuwenhuyze and Loosveldt, 2013). For this reason, survey variables can be used only for studying mode differences in selection error (RQ 1), if it can be assumed that measurement error is absent or equal across modes. However, this condition is likely not to hold for many attitudinal or factual questions. Prior studies therefore mainly assessed selection error on sociodemographic variables assuming that sociodemographic questions are answered identically in all modes (e.g. Miller *et al.* (2002), Bälter *et al.* (2005), Link and Mokdad (2005, 2006) and Dillman *et al.* (2009a)). Using survey variables for the assessment of selection error additionally enables only

relative comparisons between modes, whereas the absolute size of selection error against the sampling frame is unknown. This objective requires sampling frame information, which is often not available to researchers (Groves, 2006).

For similar reasons, the potential of sequential MM designs for reducing selection error (RQ 2) could only be assessed in rare circumstances (De Leeuw, 2005). The scarce empirical evidence suggests that MM designs might increase selection error on some variables, while having no effect or decreasing it on others (Voogt and Saris, 2005; Dillman *et al.*, 2009a; Link and Mokdad, 2006). However, only a very limited number of benchmark variables could be studied so far in this line of research (mainly sociodemographics; Voogt and Saris (2005) considered voting turnout).

The present study entails three key improvements over earlier empirical approaches. Firstly, we improve on data quality by using a large ($n = 8800$) random-probability sample from the Dutch general population and by using sociodemographic variables from the national register as benchmarks. Measurement error is known to be small in the Dutch register and can be ignored in the present study for this reason (Bakker, 2012). In addition, these data allow an assessment of the size of selection error against the sampling frame and hence absolute conclusions about the size of accumulated selection error of the SM and MM survey designs under study.

Secondly, we extend analyses from sociodemographics to survey target variables, while addressing the confounding of measurement and selection error by using a within-subject experimental design (Klausch *et al.*, 2014). In this design, we remeasured key target variables in a single mode (FTF) after the split-ballot mode experiment that was described above. In doing so, we avoided the confounding problem, because the remeasured target variables could only show measurement error of FTF surveys regardless of the mode that was assigned earlier. We applied a non-response correction procedure to extrapolate conclusions from the reinterview to the population by using the register as a source of auxiliary information.

Thirdly, we improve on the statistical methodology. Prior research has considered selection error in isolated univariate analyses of different benchmark variables. Drawing conclusions about the effect of mode is difficult, however, if findings differ by variables. To answer the RQs it is then more important, whether SM and MM designs still cause 'systematic differences' in selection error, even though there might be variable-specific variations. Conclusions, such as 'FTF surveying evokes on average less selection error than other designs', are only defendable if information about selection error can be combined across variables. Approaches to combine evidence for selection error across sets of variables are reviewed and applied in the present study.

2. Research design

We conducted a mode experiment with two waves within the national CVS that was carried out by Statistics Netherlands in 2011. In the first wave, a probability person sample of 8800 individuals was drawn from the national register, which is a list of all individuals living in the Netherlands that can be considered free of coverage error. The sample was randomly split between four modes: FTF, telephone, mail and Web (2200 individuals each). This design allowed estimating and comparing selection error for a set of benchmark variables across modes (RQ 1). The benchmark measures that were applied in the present study were partly available from the register and partly collected during the second wave, as described in detail next. Afterwards, we provide details on the MM data used for RQ 2 and provide details on fieldwork procedures.

2.1. Measures for studying selection error

Selection error was studied on two types of variable: sociodemographics and target variables from the CVS. A set of six sociodemographic variables was available from the national register: sex, age, income, ethnicity, marital status and size of household. We additionally included two geographical indicators (degree of urbanization and living in one of the three large cities of the Netherlands). These variables represent a high quality (i.e. small measurement error; see Bakker (2012)), exogenous benchmark that is available for all units regardless of being non-respondents or respondents in any mode.

To cope with the difficulty of confounded selection and measurement error on target variables that were observed during the first wave, we administered another survey to the same sample about 4–6 weeks later. This 'second wave' approached every individual again using only the FTF mode, regardless of the mode that was assigned at wave 1 and being respondent or non-respondent. A core set of questions from the CVS was repeated in this survey. These second-wave FTF measurements could now be regarded as a benchmark exhibiting equivalent measurement error regardless of the mode that was assigned at wave 1 (i.e. measurement error of the FTF mode). The wave 2 data could be compared across wave 1 respondents and non-respondents to assess selection error. These evaluations were not confounded with measurement error anymore, since all variables were measured by FTF surveying in wave 2.

The set of repeated CVS variables was rich, including 22 attitudinal and factual questions. Attitudinal questions were asked about the social quality, security and problems of the neighbourhood. Factual questions concerned the time of past police contact and past victimization to different forms of crimes.

However, the second wave, like any survey, suffered from unit non-response (48.6%). Since we needed the full second-wave sample as a benchmark for valid inference about the population, the missing data problem had to be adequately addressed. We used multiple imputations of unit non-response for this, as explained in detail in Section 3.4 (Rubin (1987), pages 15–17, Schafer and Graham (2002) and van Buuren (2012), pages 25–49). The sociodemographics from the register could be used to build the imputation models, because they were available for all units in the sampling frame. A *caveat* of using reinterview data as the benchmark is that respondents at wave 1 are less likely to participate repeatedly and thus FTF response at wave 2 is lower than at wave 1. We give detailed attention to this potential problem in Section 2.3.

2.2. Mixed mode data collection

After having assessed the difference in selection error between modes at the first wave, we considered the change that was brought to the selection error by approaching wave 1 non-respondents (including non-covered units) by FTF surveying that was used during the second wave (RQ 2). The two-wave experimental design that was used to collect FTF measurements of target variables thus additionally made available data that strongly resembled three sequential MM surveys: in particular, telephone–FTF, mail–FTF and Web–FTF.

2.3. Fieldwork

Fieldwork took place in the period from April to June 2011. First, we posted personalized invitation letters to all mail addresses. These letters differed only by information on how to participate in the survey: the mail condition contained a paper questionnaire, letters in the Web condition mentioned a hyperlink to the on-line survey and letters in telephone and FTF interviewing informed individuals about forthcoming interviewer contact. Incentives were not offered in any of the conditions. Individuals were not informed about wave 2 at this

point to avoid biasing response samples at wave 1 for respondents who preferred an FTF mode.

During the subsequent fieldwork of 4 weeks that was used for wave 1, individuals in the two self-administered modes received up to two mailed reminders before being classified as wave 1 non-respondents. In the telephone condition, multiple call attempts were made to individuals with a known landline telephone number. 28.5% of all people who were allocated to the telephone condition were not covered by lists that are available to Statistics Netherlands, however (or numbers turned out incorrect during fieldwork). Non-covered people were classified as telephone non-respondents directly. In the FTF condition, finally, the interviewers tried to establish contact on location for a maximum six times.

Table 1 provides information on the response turnout of the two-wave design. The second column shows RRs to the first wave, the third column RRs to the second wave and the fourth column the MM RRs. Here the term RR refers to the proportion of respondents out of all eligible units covered by the person sampling frame including units that were not covered by telephone (also discussed as the 'realization rate' in Skalland (2011); this definition deviates from the American Association for Public Opinion Research 'RR1' standard, which would exclude non-telephone households in the telephone sample before computing telephone RRs). Originally, 2200 individuals were assigned to each condition, but frame errors (e.g. due to relocation) reduced these numbers slightly (eligible units indicated in parentheses next to the RRs). In the FTF mode, 64.3% of eligible people responded, 48.2% in the telephone, 49.8% in the mail and only 28.7% in the Web mode. This order of RRs matches experience of earlier studies, stressing the point that the highest response can be expected when offering FTF as a response mode and the lowest when offering Web (see Section 1). Section 4 assesses whether these differences in RRs also imply mode differences in selection error addressing RQ 1 (the American Association for Public Opinion Research RR1 standard for the telephone sample was 67.7%, which excludes non-telephone households before calculating RR1; although this RR was substantially higher, our RQs consider the combined effect of non-response and non-coverage and so do the RRs that are reported in Table 1).

The second wave followed 4–6 weeks after the first. Because of cost, a smaller sample of wave 1 units was randomly selected (80%; $n = 6803$). Telephone households, furthermore, were oversampled for reasons that were unrelated to the present study. All subsequent analyses involving wave 2 used design weights to adjust for this overrepresentation. The weighted RRs

Table 1. RRs of wave 1 and 2 by mode of administration, and the change induced by an FTF follow-up to non-respondents at wave 1 (MM response)†

Mode assigned at wave 1	*RR, wave 1 (%)*	*RR, wave 2 (%)*	*RR, MM (%)*	*ΔRR*
FTF	64.3 (2081)	53.2 (1639)	(71.1)‡ (1639)	(6.7)‡
Telephone	48.2 (2062)	50.1 (1658)	65.2 (1658)	16.6
Mail	49.8 (2182)	51.9 (1760)	67.3 (1760)	18.0
Web	28.7 (2199)	50.5 (1746)	59.6 (1746)	30.7
Total	47.5 (8524)	51.4 (6803)	65.7 (6803)	18.1

†Sample sizes *n* are given in parentheses.
‡For completeness the increase in response rates for the wave 1 FTF– wave 2 FTF combination is reported. In this case the wave 2 FTF survey represents a simple non-response follow-up survey for first-wave FTF non-respondents and not an MM design.

at wave 1 in this subsample closely followed the RR that is indicated in the second column of Table 1 for the full sample.

Contrary to wave 1, no separate invitation letters were sent at the outset of the second wave. Instead, FTF interviewers were instructed to explain to individuals the need for (repeated) participation in the CVS. This procedure was chosen to keep the perceived survey burden for respondents at the first wave low, which could have been increased significantly by communicating the second wave as a separate survey.

The third column of Table 1 shows RRs at the second wave. 3498 individuals responded, which represent 51.4% of all eligible people ($n = 6803$). Although the response at wave 2 was overall lower than for FTF surveying at wave 1 (64.3%), the variation of wave 2 RR across modes assigned at wave 1 was not significant ($\chi^2 = 4.07$; degrees of freedom df $= 3$, not significant). Hence we may assume that the unwillingness for repeated participation at wave 2 was distributed evenly across wave 1 modes. There was also no systematic difference in sociodemographic indicators between the FTF response sample at wave 1 and wave 2 (Klausch *et al.*, 2013; Schouten *et al.*, 2013). Both conjectures support the claim that the full second wave was very similar to a standard FTF survey and there was no dependence of response on modes that were assigned earlier.

Finally, the fourth and the fifth columns of Table 1 illustrate the change in RRs that was caused by adding wave 2 FTF respondents who were non-respondents or not covered at wave 1 to the mode-specific response samples (i.e. MM response; RQ 2). All MM designs yielded a strong response increase (ΔRR). This change was particularly pronounced for the Web mode (30.7). About half of the MM Web sample was represented by second-wave FTF respondents. In the telephone and mail samples this effect was still strong (16.6 and 18.0 respectively). The power of the mode switch is stressed by comparing these figures with the FTF condition, where the response was only marginally increased (6.7%). These RR shifts reflect experiences that have been reported in the literature (see Section 1) and can be considered typical for sequential MM surveys of this type. In the analyses we assess the implications for selection error (RQ 2). However, before we present these results, the next section describes our statistical approach.

3. Statistical methodology

In this section, we describe how selection error was studied statistically. Answering the RQs required two steps. First, we measured selection error in each mode and compared it across the SM designs. Second, we assessed the effect of the MM extension on the selection error of the SM design (i.e. we considered whether it was decreased, increased or stayed equal). In doing so, it was relevant to specify what we mean by selection error. We see three relevant ways of specifying selection error:

(a) absolute selection error per benchmark variable,
(b) relative selection error per benchmark variable and
(c) absolute selection error over sets of benchmark variables.

Absolute and relative selection errors are specific to a given variable. Analyses based on these statistics thus may lead to variable-dependent conclusions about the RQs. The third approach therefore tries to generalize conclusions on absolute selection error across sets of variables to answer the RQs. After introducing these perspectives, we provide details on how a multiple-imputation technique was used to allow inference to the population for target variables measured in wave 2.

3.1. Absolute selection error

Suppose that a survey design using mode M evokes selection error on a given variable Y. Then the distribution of Y differs between response and non-response groups identified by response mechanism S. We can write the conditional distribution of Y as $P(Y|S, M)$. Selection error is absent in a specific mode $M = i$, when Y is conditionally independent of S:

$$P(Y|S, M = i) = P(Y|M = i). \tag{1}$$

Our benchmark variables Y (e.g. sociodemographics) were available on discrete measurement levels with differing numbers of categories. In principle, definition (1) could be applied to each category-specific estimate of selection error by assessing its size and comparing it across designs. However, this procedure would lead to a large number of estimates increasing with the number of categories, variables and design comparisons.

We therefore used a summary statistic for the strength of absolute deviations from independence (1) in mode $M = i$ across all categories of Y on the variable level, which is called the absolute selection error. Our analyses are based on Pearson's χ^2-statistic which sums the squared category-specific deviations from independence (i.e. absence of selection error) across cells of the contingency table of Y and S. However, χ^2 must be scaled to be comparable across tables. A useful measure for this purpose is Cramér's V, scaling χ^2 to the interval of 0 and 1. V has an effect size interpretation (e.g. 0.10–0.30 small selection error, 0.30–0.50 medium and 0.50–1.00 large). V, however, is by far not the only measure of strength of association that could be applied. Alternatives are represented by, for example, the dissimilarity index (Agresti (2002), pages 329–330) or Cohen's w (Cohen (1977), pages 222–224). All analyses were also executed for the alternative measures with equivalent results. We present results on V, because of the clear effect size interpretation of the index.

Furthermore, testing independence on the basis of Pearson's χ^2 or likelihood ratio statistics indicates the statistical significance of associations. However, in particular in large samples even small deviations from independence can lead to significant tests. In these cases, effect size measures such as V are more informative about the strength of deviation from independence (i.e. absolute selection error).

To answer RQ 1 for a survey variable Y, we estimate V for each mode and compare it across modes. To answer RQ 2, the change in V by the FTF follow-up is assessed.

3.2. Relative selection error

It is a disadvantage of measures of absolute selection error that the direction (sign) of error is ignored across categories. Equivalent Cramér's V-indices, therefore, suggest neither that selection error is present on the same categories nor that error has the same sign. Instead V merely indicates that across-categories absolute deviations from expected frequencies are equal. For this reason it is additionally informative to consider relative selection error, which is based on the independence relationship

$$P(Y|S = 1, M) = P(Y|S = 1). \tag{2}$$

The left-hand side of the equation considers the response distribution of Y conditionally on mode. If the response distribution depends on mode M, there may be both positive and negative deviations from the unconditional response distribution $P(Y|S = 1)$. The conditional response distribution can be thought of as a contingency table of M and Y with frequency counts from the response sample only. For example, consider a table of categorized income bounds against the four modes that were assessed in this study. Each cell contains the number of respondents

in a particular mode and an income group. The level of dependence of mode and income consequently indicates the strength of relative differences in selection error, which again can be measured by Cramér's V for the contingency table.

However, absolute differences against the population cannot be seen from relative selection error (i.e. definition (1) is needed). For example, even if equation (2) holds and there are no relative differences in selection error, there can still be selection error against the population if $P(Y|S=1) \neq P(Y)$. Since this information is required to answer the RQs (for example, for RQ 2 we need to know whether selection error was increased or decreased by the MM follow-up, which cannot be seen from relative selection error), we focused on design differences in absolute selection error in all analyses and considered relative selection error as secondary information.

3.3. Absolute selection error for sets of benchmark variables

With this level of analysis, we want to extract the systematic effect of mode on multiple variables. The benchmark variables may show a diffuse picture in their absolute selection error over designs due to factors that are related to variable content. In this case, however, it becomes difficult to attribute the differences in selection error to the mode. Assessing systematic differences in absolute selection error aims to isolate the joint variance in absolute selection error across variables, which is due to the design and not variable content. In the literature, for example, there were inconsistent findings across variables when assessing the effect of MM on selection error (see Section 1). In such a situation it is helpful to extend variable level analyses to a summarizing view on selection error across variables. For this, we chose two different strategies.

First, we compared the distribution of all variable-specific estimates of absolute selection error across the SM designs addressing RQ 1 (i.e. all Cramér's V-estimates from definition (1)). The variation of V-statistics across variables and their central tendency allows conclusions about which design 'on average' causes less absolute selection error and how much variance can be attributed to variable content. To assess RQ 2, the effect of the MM follow-up on the SM distribution of V was considered.

Second, we applied an index which summarizes deviations from independence in definition (1) across multiple variables simultaneously: the 'representativeness (R-) indicator' (Schouten *et al.*, 2009, 2011; Shlomo *et al.*, 2012). R-indicators are a new class of quality indicators defined on the variance of response probabilities represented by $P(S=1|Y, M=i)$. If selection error is absent (definition (1)), it holds that

$$P(S=1|Y, M=i) = P(S=1|M=i). \tag{3}$$

In other words, response probabilities then are constant across Y. Therefore, deviations from constant response probabilities, measured by their variance, are informative about the extent of selection error on Y. Schouten *et al.* (2009) have shown that, when estimating $P(S=1|Y, M=i)$ for a set of multiple benchmark variables Y by means of a logit model, the variance of response probabilities indicates the extent of selection error on all underlying variables. The R-indicator is then defined as

$$R_{M=i} = 1 - 2\,\mathrm{sd}\{P(S=1|Y, M=i)\} \tag{4}$$

where 'sd' denotes standard deviation. R varies between 0 and 1, where values close to 1 indicate absence of selection error ('representative response'). To assess RQ 1, R can then be compared across the SM designs to assess systematic differences in selection error (i.e. for all benchmark variables Y simultaneously). To assess RQ 2, the change in R by the MM follow-up is considered.

As the response probabilities are estimated from a logit model, R-indicators additionally control for the multicollinearity in the benchmark variables, isolating the covariance with S that is uniquely caused by an underlying selection mechanism. Therefore, analyses of R and mean Cramér's V might differ when considering multiple Y. For single variables Y, however, Cramér's V and R are asymptotically equivalent in large samples except for a scaling constant (Schouten *et al.*, 2009). R-indicators, however, still lack a clear effect size interpretation like V but feature the advantage to summarize the variance of response probabilities over multiple variables into a single number. Variance estimation for R-indicators was discussed in Shlomo *et al.* (2012).

3.4. Inference to the population for wave 2

As pointed out in Section 2.3, the second wave suffered from unit non-response (48.6%). We imputed the missing cases by an algorithm called 'multiple imputations by chained equations' using sociodemographic frame information as auxiliary information (van Buuren and Groothuis-Oudshoorn (2011) and van Buuren (2012), pages 20–49). This method assumes that units were missing at random in wave 2 given register information (Little and Rubin (2002), pages 12–19). It is important to note that inference to the population is only valid given this assumption.

10 multiply imputed data sets were analysed separately and then pooled by taking the mean of V and R-indicators across the imputed data sets (Rubin (1987), page 90, and Schafer and Graham (2002)). P-values for tests of independence in multiply imputed contingency tables were computed following a procedure for the pooling of χ^2-distributed statistics described by Li and colleagues (Li *et al.* (1991), Schafer (1997), pages 115–116, and van Buuren (2012), page 159). The standard error of R-indicators is estimated by pooling the between- and within-imputation variances of R by using Rubin's rules (Rubin (1987), page 90).

4. Results

We present analyses of selection error on eight register variables and 22 CVS target variables. First we consider RQ 1, estimating absolute and relative selection error in each of the four SM designs per benchmark variable and across sets of variables. Subsequently, RQ 2 is assessed by illustrating the effect of the three MM designs on absolute selection error of the SM designs.

4.1. Single-mode differences in absolute selection error (research question 1)

RQ 1 asked whether the order of RRs reflected the extent of selection error of the four SM surveys: FTF, telephone, mail and Web. We found the largest response in the FTF mode, followed by mail and telephone (Section 2.3, Table 1). The response was lowest in the Web mode. Table 2 provides Cramér's V-statistics for eight register variables. It is instructive to consider first relative selection error (formula (2)). Here, V is estimated for each contingency table of a benchmark variable and mode conditional on response in wave 1 ($n = 4048$). We found significant differences in selection error across modes on all except the 'urbanization' characteristic. However, the small size of V (less than 0.10 in all cases) suggests that the category-specific differences in selection error were small across modes.

This analysis is still uninformative about the absolute size of selection error on variable level. For this purpose V is estimated, separately for each mode, from the contingency tables of the response indicator and the sociodemographics (formula (1)). We found statistically significant selection error in all modes for most variables (the absolute selection error columns). The size of effects can be classified as 'small' in all cases, however ($V < 0.30$). The strongest differences in V were found for the income distribution and the two regional indicators. On the income

Table 2. Absolute and relative selection error on eight sociogeographical indicators from the national register (Cramér's *V*)†

Variable	*Absolute selection error V*				*Relative selection error V, all modes*
	FTF	*Telephone*	*Mail*	*Web*	
Gender	0.046‡	0.036	0.050‡	0.050‡	0.045‡
Age	0.092§	0.125§§	0.184§§	0.128§§	0.051§
Income	0.039	0.087§	0.141§§	0.174§§	0.050‡
Ethnicity	0.167§§	0.228§§	0.167§§	0.094§§	0.045§
Marital status	0.090§§	0.149§§	0.131§§	0.111§§	0.049§
Household size	0.078§	0.167§§	0.165§§	0.139§§	0.040‡
Urbanization	0.171§§	0.145§§	0.057	0.037	0.036
Urban cities	0.147§§	0.113§§	0.038‡	0.028	0.043‡
Sample size	2081	2062	2182	2199	4048

†Categories of discrete variables: gender, two categories (male and female); age, six categories (15–24, 25–34, 35–44, 45–54, 55–64 and 65 years or higher); income, seven categories (no income, up to €30000, €30000–45000, €45000–60000, €60000–100000, €100000 and above, and missing); ethnicity, four categories (Dutch, Western, foreigner and non-Western foreigner); marital status, four categories (married or partnership, single, divorced or widowed and missing); household size, three categories (one person, two and three or more people); urbanization, five categories (very strong, strong, moderate, little and none); urban cities, three categories (living in Amsterdam, Rotterdam or Utrecht and living elsewhere).
‡$p < 0.05$.
§$p < 0.01$.
§§$p < 0.001$.

variable, the Web mode showed the strongest selection error followed by mail, whereas the telephone and FTF modes showed weaker or no selection error. In contrast, the FTF and telephone modes showed stronger error on the regional indicators. Compared across variables, the risk for selection error that is implied by the RRs was not reflected by the order of *V*. 'Income' was the only variable that showed the expected order (i.e. FTF showed least and Web surveying most selection error). Furthermore, the size of the *V*-statistics indicated small effect sizes throughout, suggesting that no mode stood out strongly in terms of accumulated selection error. Put differently, each mode seemed to have weaknesses and strengths in terms of sociodemographical reach, but differences were generally small.

Subsequently, we extended this analysis to selection error on target variables from the CVS surveyed during the second wave in FTF mode. Table 3 shows the 22 variables sorted in groups of questions on the 'social quality of the neighbourhood', 'problems in the neighbourhood', 'summary ratings about the neighbourhood' and 'factual questions'. The factual questions about past victimization are aggregated from multiple questions about victimization to diverse forms of crimes. As explained in Section 3.4, these statistics are based on pooled multiply imputed data sets. Inference to the population is only valid given that units were missing at random on sociodemographics. In addition, we assessed selection error against wave 2 FTF respondents only, i.e. deleting wave 2 unit non-respondents listwise. The findings of this analysis were equivalent to those shown here.

The relative selection error was small and insignificant for nearly all variables except for 'safety feeling', suggesting that any differences in absolute selection error perhaps were likewise small (Table 3). Considering absolute selection error, we found that this was indeed true for many variables. However, in some cases *V* in FTF and telephone mode appeared somewhat higher

Table 3. Absolute and relative selection error (Cramér's *V*) on 22 CVS target variables measured in wave 2 (estimates based on 10 multiply imputed data sets)

	Absolute selection error V†				*Relative selection error V*†, *all modes*
	FTF	*Telephone*	*Mail*	*Web*	
Questions 'social quality neighbourhood'‡					
State of roads, walkways and squares	0.052	0.035	0.043	0.058	0.022
Good playgrounds for children	0.039	0.060	0.030	0.041	0.031
Good provisions for younger people	0.029	0.024	0.033	0.048	0.039
People know each other well	0.089§	0.089§§	0.054	0.051	0.025
People treat each other well	0.110§	0.148	0.060	0.090§§	0.050
Nice neighbourhood with solidarity	0.096§§	0.055	0.060	0.051	0.032
Feel at home with people	0.118§	0.090§§	0.084§§	0.064	0.031
Have a lot of contact with people	0.072	0.055	0.049	0.050	0.023
Satisfied with population composition	0.113§	0.106§	0.089§	0.045	0.033
*Questions 'neighbourhood problems'***					
Plastering on walls and/or buildings	0.096§§	0.057	0.089§§	0.050	0.032
Harassment by groups of young people	0.129	0.102§	0.063	0.062	0.033
Drunken people on the streets	0.088§	0.083§§	0.097§§	0.060	0.045
Unpleasant people on the streets	0.125§	0.099§	0.078	0.068	0.020
Junk on the streets	0.094§§	0.064	0.053	0.052	0.035
Dog excrement on the streets	0.031	0.042	0.036	0.071	0.035
Destruction of telephone boxes, etc.	0.050	0.029	0.028	0.023	0.022
Drug problem	0.121*	0.074	0.090	0.046	0.035
Summary ratings about neighbourhood					
Safety feeling: insecure††	0.130*	0.059§§	0.011	0.028	0.061§§
Quality-of-life rating‡‡	0.115§	0.121	0.104§	0.088	0.027
Factual questions (yes or no)+					
Police contact (past 12 months)	0.061	0.053	0.056	0.033	0.029
Victim of crime (past 12 months)	0.075§§	0.102§	0.073	0.059	0.034
Victim of violent crime (past 12 months)	0.079	0.105§	0.064§§	0.050	0.034
Sample size	2081	2062	2182	2199	4048

†Mean weighted estimates of Cramér's V across 10 multiple imputations, significance levels based on an adjusted χ^2-test of independence according to Li *et al.* (1991). All questions were surveyed with separate 'don't know' answer categories. On most variables these categories were very sparse (less than 1% of cases) and were imputed.
‡Question with five rating scale answer categories from 'completely disagree' to 'completely agree' and 'don't know'. To avoid cell sparseness, the categories 'completely agree' and 'agree' as well as 'completely disagree' and 'disagree' had to be merged in the analyses.
§$p < 0.01$.
§§$p < 0.05$.
*$p < 0.001$.
**Questions with three rating scale answer categories: 'happens frequently', 'happens sometimes' and 'happens rarely or never'.
††Question 'Do you sometimes feel insecure?' with two answer categories: yes or no.
‡‡Question on a summary rating about the neighbourhood on a 10-point scale with a 'don't know' category. To avoid cell sparseness, recoded to four categories (1–6, 7, 8 and 9–10).
+Dichotomous measures about police contact within past 12 months and two summary indices based on multiple questions about victimization.

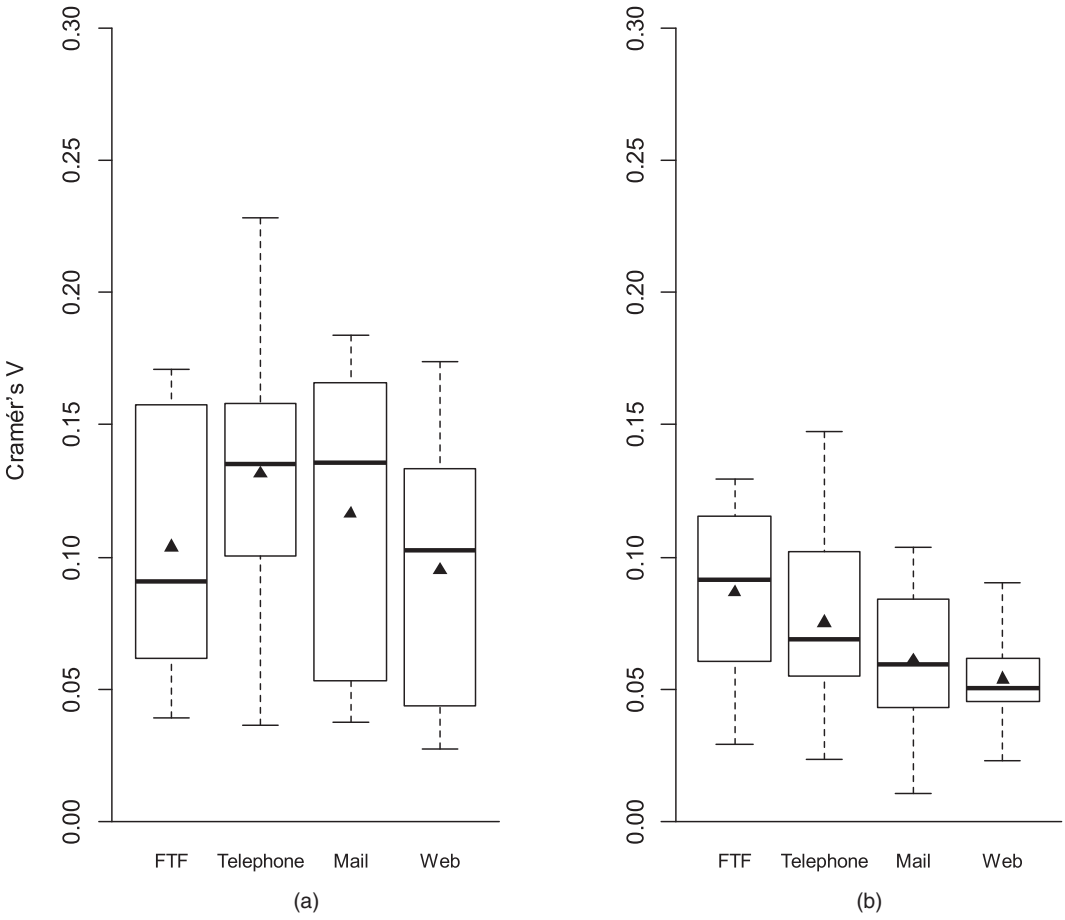

Fig. 1. Empirical distribution of Cramér's *V*-statistic from Tables 2 and 3 (measuring absolute selection error) for the four SM designs (▲, means; —, medians): (a) sociodemographics; (b) CVS variables (mutiple imputation)

than in the mail and Web modes, suggesting slightly stronger selection error on these variables (e.g. 'safety feeling' and 'unpleasant people on the streets').

The large number of V-estimates for sociodemographics and CVS variables and somewhat differing findings across variables still aggravated generalizing conclusions about the effect of mode on selection error. Therefore, it was instructive additionally to assess absolute selection error across these sets of variables (see Section 3.3). We first visualized the empirical distribution of variable-specific absolute selection error by using boxplots of the V-estimates from Tables 2 and 3 in Fig. 1. Each boxplot contains the same estimates as listed in Tables 2 and 3, where triangles denote means and bars medians. This illustration allowed three additional insights: first, the error was somewhat larger for the sociodemographic indicators than for the CVS variables. Second, on sociodemographics, mail and telephone surveying evoked somewhat higher selection error, on average, than the FTF and Web modes. However, the large ranges of all boxplots indicated strong variance across variables, as noted above. Third, the selection error of CVS variables was slightly stronger in the FTF and telephone than in the mail and Web modes. Again there was considerable variance across CVS variables.

Table 4. Representativeness (*R*-) indicators for sociodemographics and target variables by SM and MM designs†

	Results for SM designs				*Result for MM designs*		
	FTF	*Telephone*	*Mail*	*Web*	*Telephone + FTF*	*Mail + FTF*	*Web + FTF*
Sociodemographics	0.768 [0.728, 0.809]	0.698 [0.660, 0.736]	0.719 [0.681, 0.757]	0.798 [0.759, 0.836]	0.751 [0.706, 0.796]	0.811 [0.766, 0.855]	0.803 [0.757, 0.850]
CVS variables	0.831 [0.765, 0.897]	0.823 [0.761, 0.885]	0.857 [0.789, 0.926]	0.882 [0.828, 0.937]	0.789 [0.731, 0.846]	0.842 [0.777, 0.907]	0.825 [0.753, 0.898]
Response rate‡	64.3	48.2	49.8	28.7	65.2	67.3	59.6

†For CVS variables, the mean *R* across 10 multiply imputed data sets is reported. Variances of *R* were estimated by using Rubin's rule for multiply imputed data sets (Rubin (1987), page 90).
‡Based on Table 1.

In considering these results it is important to note again the magnitude of all Cramér's V-estimates, indicating small effect sizes (selection error) in all cases. Thus, no mode stood out greatly in terms of absolute error. Remarkably, however, the Web mode evoked, on average, less error than the telephone and mail modes despite its lower RR (28.7%).

Finally, we abstracted from the level of variables to a summary score for absolute selection error by using the *R*-indicator (formula (4)). The left-hand part of Table 4 shows *R* for the SM designs. All estimates were somewhat larger for the CVS variables than for sociodemographics, suggesting that the selection error was smaller, overall, on target variables, which reflected findings from Fig. 1. Also the order of *R*-estimates relates to the findings from Fig. 1. The Web (0.798) and FTF (0.768) modes achieved slightly more representative responses than telephone (0.698) and mail (0.719) on sociodemographics. On CVS variables the differences were less pronounced, where the Web (0.882) mode achieved a slightly more representative response than the telephone (0.831) and FTF (0.823) modes. Clearly, the order of magnitude of these estimates did not reflect the order of RRs that are shown again in the last row of Table 4.

4.2. Effect of the mixed mode designs on absolute selection error (research question 2)

The three sequential MM designs greatly increased RRs by the FTF follow-up survey to approximately the level of an SM FTF survey (64.3%; see Table 4). The second RQ asked whether this shift in RRs caused a mitigation of selection error. To asses this question, we first considered the change in *R*-indicators (Table 4, right-hand column). On sociodemographics, we noted an upward trend for the telephone mode (from 0.698 to 0.751) and mail (from 0.719 to 0.811), indicating a reduction in selection error, but the Web mode was unchanged (from 0.798 to 0.803). However, for the CVS variables we did not find any increased *R*-indicators. In fact, the somewhat reduced *R* for the telephone (from 0.823 to 0.789), mail (from 0.857 to 0.842) and Web (from 0.882 to 0.825) modes suggested a slight increase in selection error. It appeared that the MM design was capable of reducing error in some, but not all, cases.

Comparing the MM designs with the SM FTF survey offered an explanation for these findings. The *R* for FTF sampling was higher than for telephone and mail for sociodemographics, but it

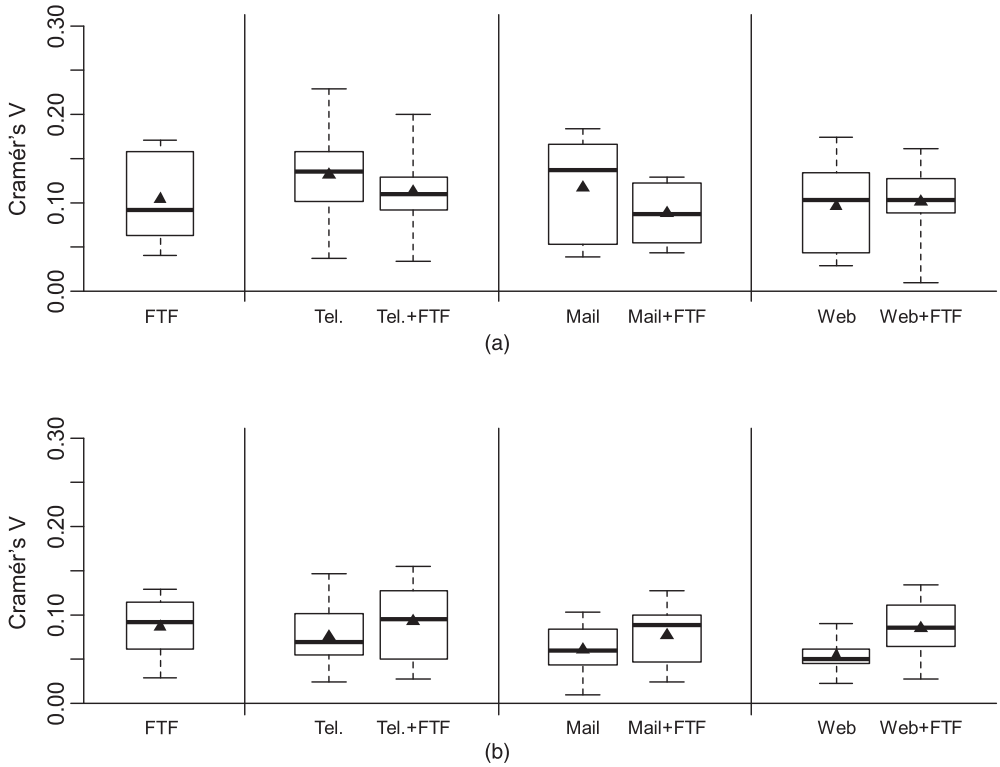

Fig. 2. Comparison of Cramér's *V*-statistics (measuring absolute selection error) between the SM and MM designs (▲, means; —, medians): (a) sociodemographics; (b) CVS variables (mutiple imputation)

was not higher than for the Web mode. For CVS variables the FTF *R* was smaller than those for the mail and Web modes. Therefore, it appeared that *R* for the MM designs trended towards that for the SM FTF survey. This seemed plausible as also the RRs of the MM designs were increased to the level of that for the FTF mode but did not clearly exceed it.

We investigated this idea in more detail by using the distribution of variable-specific *V*-indices measuring absolute selection error (equation (1)) shown in Fig. 2 contrasting the SM and MM designs. Clearly, absolute selection error on sociodemographics of MM mail and MM telephone sampling showed a downward shift towards the level of SM FTF sampling. The Web results remained unchanged at the level of FTF sampling in central tendency. However, considering CVS variables we found a strong change in *V*-statistics in the Web mode, which increased to the level of the SM FTF results. These trends were also visible for telephone and mail surveying.

These effects suggest that the absolute selection error of the three SM samples was turned more similar to the SM FTF sample by the FTF follow-up. However, the error was not mitigated on all variables (RQ 2). Rather, the FTF follow-up also increased absolute selection error on some of the variables (e.g. CVS variables for the Web mode). In the supplemental on-line material we demonstrate the details of this process and provide a summary of these analyses here. If the telephone, mail or Web modes exhibited smaller error than the FTF mode, the selection error was increased by the MM extensions. Furthermore, MM sampling was only capable of decreasing error, if the FTF mode showed smaller error than the telephone, mail or Web modes. These effects were found for both types of variable and all MM designs. Even for the Web condition, for which no change in *R* (Table 4) and central tendency of *V* (Fig. 2) was found on sociodemographics, the

error was changed towards FTF surveying. However, it was both mitigated on some variables and increased on others. Therefore, we could not observe an overall change on systematic level though it became apparent when considering all variable-specific changes.

5. Discussion

Response and coverage rates are often assumed to indicate the risk for selection error of a survey design and as such they are established as central indicators of survey quality. Strong enhancements in RRs, furthermore, are often referred to as an advantage of MM surveys, promising to keep selection error small besides reducing costs. The present study assessed the legitimacy of this common practice empirically, reporting two key findings.

First, we found that RRs did not indicate the extent of selection error of the SM designs (RQ 1). On sociodemographics, the FTF and, surprisingly, Web modes evoked least selection error (greater representativeness as measured by the *R*-indicator), whereas the error on the telephone and mail modes was slightly stronger. On CVS variables, the size of selection error and mode differences was virtually negligible. These findings reflect the empirical results of Groves (Groves and Peytcheva, 2008; Groves, 2006) regarding the weak relationship of RRs and non-response bias. The fact that the Web mode, in the face of its low RR (28.7%), performed slightly better than the other modes is particularly remarkable in this respect.

Second, the FTF follow-up survey to telephone, mail and Web non-respondents turned out to be able to reduce the selection error of the SM designs, when the error in the SM designs was larger than in the FTF design (RQ 2). However, when the error in the SM design was smaller than in the FTF, the error was enlarged. Across all variables a reduction in selection error beyond SM FTF was not possible, but the selection error of the MM samples was aligned towards an FTF SM sample.

The research design of the present study offered several advantages over prior empirical approaches. Random sampling from a high quality person sampling frame allows generalizing findings to the Dutch population. Furthermore, the set of sociodemographic and CVS benchmark variables was exceptionally rich and of high quality, because we partly used register data (for sociodemographics) and partly an SM reinterview design (CVS target variables) to control for mode-specific confounding of selection and measurement error. In spite of these strong aspects of the design, some limitations need to be considered. It can be argued that assessing different survey topics might alter conclusions about target variables, although this threat is likely to be weaker for sociodemographics. In addition, population inference for target variables is based on an assumption of missingness at random used to adjust non-response in wave 2.

Moreover, results might be affected, if this study is repeated in populations with different survey-taking climates and different coverage by Internet or telephone access. In the Netherlands, Internet access reached approximately 90% of households in 2011 (Eurostat, 2012). The good performance of the Web mode might therefore be a particularity of the Netherlands. This statement also applies to countries with different registered telephone coverage (71.5% in the present study). Non-registered phones, of which a considerable part is mobile phones, were not included in the present study and doing so might change the results.

Because of these limitations, replicating our findings for different survey topics and populations is strongly desirable, but if our results are generalizable they will have important implications for survey practice. Selection error would be a less important factor in the choice of survey mode than is commonly assumed and RRs should gain less prominence in arguing for or against particular modes. Furthermore, it would become doubtful whether non-response follow-ups provide any strong additional benefit over any of the SM designs, because the selec-

tion error would not differ substantially between SMs of administration and the FTF follow-up could not increase representativeness strongly. In conclusion, our findings challenge the current practice of using sequential MM surveys as a tool for controlling alleged differences in selection error between SM designs.

Acknowledgement

The authors acknowledge the comments by two reviewers, which greatly helped to improve the manuscript.

References

Agresti, A. (2002) *Categorical Data Analysis*, 2nd edn. Hoboken: Wiley-Interscience.
Bakker, B. F. M. (2012) Estimating the validity of administrative variables. *Statist. Neerland.*, **66**, 8–17.
Bälter, K. A., Bälter, O., Fondell, E. and Lagerros, Y. T. (2005) Web-based and mailed questionnaires. *Epidemiology*, **16**, 577–579.
Bethlehem, J. (1988) Reduction of nonresponse bias through regression estimation. *J. Off. Statist.*, **4**, 251–260.
Biemer, P. P. and Lyberg, L. E. (2003) *Introduction to Survey Quality*. Hoboken: Wiley.
van Buuren, S. (2012) *Flexible Imputation of Missing Data*. Boca Raton: CRC Press.
van Buuren, S. and Groothuis-Oudshoorn, K. (2011) mice: multivariate imputation by chained equations in R. *J. Statist. Softwr.*, **45**, 1–67.
Cohen, C. (1977) *Statistical Power Analysis for the Behvioral Sciences*. New York: Academic Press.
De Leeuw, E. (2005) To mix or not to mix data collection modes in surveys. *J. Off. Statist.*, **21**, 233–255.
De Leeuw, E., Dillman, D. A. and Hox, J. J. (2008) Mixed mode surveys: When and Why. In *International Handbook of Survey Methodology* (eds E. De Leeuw, D. A. Dillman and J. J. Hox), pp. 299–316. New York: Erlbaum.
De Leeuw, E. and Hox, J. J. (2011) Internet surveys as part of a mixed-mode design. In *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (eds M. Das, P. Ester and L. Kaczmirek), pp. 45–76. New York: Routledge.
Dillman, D. A. and Christian, L. M. (2005) Survey mode as a source of instability in responses across surveys. *Fld Meth.*, **17**, 30–52.
Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. and Messer, B. L. (2009a) Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Socl Sci. Res.*, **38**, 1–18.
Dillman, D. A., Smyth, J. D. and Christian, L. M. (2009b) *Internet, Mail, and Mixed-mode Surveys: the Tailored Design Method*. Hoboken: Wiley.
Eurostat (2012) Households having access to the Internet, by type of connection. Eurostat, Luxembourg. (Available from `http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&language=en&pcode=tin00073&plugin=1`.)
Eva, G., Loosveldt, G., Lynn, P., Martin, P., Revilla, M., Saris, W. and Vannieuwenhuyze, J. (2010) Assessing the cost-effectiveness of different modes for ESS data collection. City University, London.
Fowler, F. J., Gallagher, P. M., Stringfellow, V. L., Zaslavsky, A. M., Thompson, J. W. and Cleary, P. D. (2002) Using telephone interviews to reduce nonresponse bias to mail surveys of health plan members. *Med. Care*, **40**, 190–200.
Greene, J., Speizer, H. and Wiitala, W. (2008) Telephone and Web: mixed-mode challenge. *Hlth Serv. Res.*, **43**, 230–248.
Groves, R. M. (1989) *Survey Errors and Survey Costs*. New York: Wiley.
Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Publ. Opin. Q.*, **70**, 646–675.
Groves, R. M. and Peytcheva, E. (2008) The impact of nonresponse rates on nonresponse bias. *Publ. Opin. Q.*, **72**, 167–189.
Hox, J. J. and De Leeuw, E. (1994) A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Qual. Quant.*, **28**, 329–344.
Jäckle, A., Roberts, C. and Lynn, P. (2010) Assessing the effect of data collection mode on measurement. *Int. Statist. Rev.*, **78**, 3–20.
Klausch, T., Hox, J. J. and Schouten, B. (2013) Assessing the mode-dependency of sample selectivity across the survey response process. *Discussion Paper 2013-03*. Statistics Netherlands, The Hague. (Available from `http://www.cbs.nl/NR/rdonlyres/D285D803-D201-437D-99F6-3FB7C5DA9C11/0/201303x10pub.pdf`.)

Klausch, T., Schouten, B. and Hox, J. J. (2014) The use of within-subject experiments for estimating measurement effects in mixed-mode surveys. *Discussion Paper 2014-06*. Statistics Netherlands, The Hague. (Available from `http://www.cbs.nl/NR/rdonlyres/181793AC-94B8-4748-9C2B-E541DCF9CFB7/0/201406x10pub.pdf`.)
Li, K.-H., Meng, X.-L., Raghunathan, T. E. and Rubin, D. B. (1991) Significance levels from repeated p-values with multiply-imputed data. *Statist. Sin.*, **1**, 65–92.
Link, M. W. and Mokdad, A. H. (2005) Alternative modes for health surveillance surveys: an experiment with web, mail, and telephone. *Epidemiology*, **16**, 701–704.
Link, M. W. and Mokdad, A. H. (2006) Can Web and mail survey modes improve participation in an RDD-based National Health surveillance? *J. Off. Statist.*, **22**, 293–312.
Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. Hoboken: Wiley.
Lynn, P. (2013) Alternative sequential mixed-mode designs: effects on attrition rates, attrition bias, and costs. *J. Surv. Statist. Methodol.*, **1**, 183–205.
Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I. and Vehovar, V. (2008) Web surveys versus other survey modes: a meta-analysis comparing response rates. *Int. J. Markt Res.*, **50**, 79–104.
Millar, M. M. and Dillman, D. A. (2011) Improving response to Web and mixed-mode surveys. *Publ. Opin. Q.*, **75**, 249–269.
Miller, T. I., Kobayashi, M. M., Caldwell, E., Thurston, S. and Collett, B. (2002) Citizen surveys on the Web. *Socl Sci. Comput. Rev.*, **20**, 124–136.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman and Hall–CRC.
Schafer, J. L. and Graham, J. W. (2002) Missing data: Our view of the state of the art. *Psychol. Meth.*, **7**, 147–177.
Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J. and Klausch, T. (2013) Disentangling mode-specific selection and measurement bias in social surveys. *Socl Sci. Res.*, **42**, 1555–1570.
Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the representativeness of survey response. *Surv. Methodol.*, **35**, 101–113.
Schouten, B., Shlomo, N. and Skinner, C. (2011) Indicators for monitoring and improving representativeness of response. *J. Off. Statist.*, **27**, 231–253.
Shih, T.-H. and Fan, X. (2008) Comparing response rates from Web and mail surveys: a meta-analysis. *Fld Meth.*, **20**, 249–271.
Shlomo, N., Skinner, C. and Schouten, B. (2012) Estimation of an indicator of the representativeness of survey response. *J. Statist. Plann Inf.*, **142**, 201–211.
Skalland, B. (2011) An alternative to the response rate for measuring a survey's realization of the target population. *Publ. Opin. Q.*, **75**, 89–98.
US Census Bureau (2010) *Design and Methodology: American Community Survey*. Washington DC: US Census Bureau. (Available from `http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf`.)
Vannieuwenhuyze, J. and Loosveldt, G. (2013) Evaluating relative mode effects in mixed-mode surveys: three methods to disentangle selection and measurement effects. *Sociol. Meth. Res.*, **42**, 82–104.
Voogt, R. J. J. and Saris, W. E. (2005) Mixed mode designs: finding the balance between nonresponse bias and mode effects. *J. Off. Statist.*, **21**, 367–387.
Wagner, J. (2012) A comparison of alternative indicators for the risk of nonresponse bias. *Publ. Opin. Q.*, **76**, 555–575.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplemental online material to "Selection error in single- and mixed-mode surveys of the Dutch general population"'.