Evaluating Bias of Sequential Mixed-Mode Designs against Benchmark Surveys

Paper to appear in Sociological Methods & Research

Corresponding Author:
Thomas Klausch
Utrecht University
Faculty for Social and Behavioural Sciences
Department of Methodology and Statistics
PO Box 80140, 3508 TC Utrecht
The Netherlands
Phone: +31 30 253 9075
Email: *l.t.klausch@uu.nl*

Co-Authors:
Barry Schouten
Utrecht University, Department of Methodology and Statistics
and
Statistics Netherlands
Methodology Department
P.O. Box 24500
2490 HA The Hague
The Netherlands

Joop J. Hox
Utrecht University
Faculty for Social and Behavioural Sciences
Department of Methodology and Statistics
PO Box 80140, 3508 TC Utrecht
The Netherlands

Abstract

This study evaluated three types of bias – total, measurement, and selection bias – in three sequential mixed-mode designs of the Dutch Crime Victimization Survey: telephone, mail, and web, where nonrespondents were followed up face-to-face. In the absence of true scores, all biases were estimated as mode effects against two different types of benchmarks. In the single-mode benchmark (SMB), effects were evaluated against a face-to-face reference survey. In an alternative analysis, a 'hybrid-mode benchmark' (HMB) was used, where effects were evaluated against a mix of the measurements of a web survey and the selection bias of a face-to-face survey. A special re-interview design made available additional auxiliary data exploited in estimation for a range of survey variables. Depending on the SMB and HMB perspectives, a telephone, mail, or web design with a face-to-face follow-up (SMB) or a design involving only mail and/or web but not a face-to-face follow-up (HMB) is recommended based on the empirical findings.

1 Introduction

Sequential mixed-mode designs have become an important alternative in international survey research. Sequential designs provide nonrespondents or non-covered persons in a single-mode survey (e.g., web) with another opportunity to respond by offering at least one other mode as a response option (e.g., face-to-face). This follow-up can increase response and coverage while achieving cost efficiency by sequencing inexpensive before expensive modes (Dillman, Smyth, and Christian 2009; De Leeuw 2005; Lynn 2013).

Since there is a direct tradeoff of mode-specific errors in the total survey error (TSE) of sequential mixed-mode designs, mixed-mode surveys may offer an optimal error balance (Bethlehem and Biffignandi 2011:235; Groves et al. 2010:176; De Leeuw 2005). Selection errors (i.e., coverage error or nonresponse error) and measurement errors have received particular attention in this context, because they are often systematic (i.e., they cause bias) and differ in size across modes (Biemer and Lyberg 2003:59; Klausch, Hox, and Schouten 2013; Kreuter, Presser, and Tourangeau 2008; De Leeuw 2005, 2008; De Leeuw, Dillman, and Hox 2008). In practice, survey designers need to estimate these biases to monitor survey accuracy and to decide about the impact of changes to fieldwork or questionnaire designs on accuracy. In doing so, two basic research questions (RQs) are of concern:

First (RQ1), *is the mixed-mode survey needed or would a single-mode survey suffice in terms of accuracy?* Given equal budget and time constraints the design featuring lowest survey error is preferred (Biemer 2010; Biemer and Lyberg 2003:44; Groves and Lyberg 2010; Kreuter, Müller, and Trappmann 2010). To evaluate the accuracy of a mixed-mode design, the expected total bias needs to be estimated for different candidate

mixed-mode designs and compared to the error of single-mode surveys without the mixed-mode follow-up. The mixed-mode survey is only needed, if the total bias of the single-mode survey is decreased by the follow-up.

Second (RQ2), *what are the major systematic sources of error in single-mode surveys and how are they impacted by the mixed-mode follow-up?* The primary motivation for sequential mixed-mode surveys is the reduction of selection bias. If the size of selection bias before and after the follow-up is estimated, the actual need and success of reducing selection bias can be evaluated. However, if one mode in a mixed-mode design measures with less accuracy, reductions in selection bias may be offset by increases in measurement bias of the mixed-mode estimate. To prevent this problem, questionnaires of the mode evoking higher measurement bias can be redesigned once the size of bias is known (Dillman et al. 2009:326).

In this paper, we answer these RQs for the case of the Dutch Crime Victimization Survey, a national survey conducted by Statistics Netherlands, based on a large-scale experiment with three mixed-mode designs: web, mail, and telephone followed up by face-to-face, re-analyzing data collected by Schouten et al. (2013). We estimate total bias of the modes before and after the mixed-mode follow-up (RQ 1) and decompose it into selection and measurement bias components (RQ 2). A major problem in the evaluation of survey bias is its estimation. Past research has discussed a series of approaches to this problem. The next section reviews this research and outlines the approach of the present study.

2 Background

Past research has mainly discussed three different approaches to evaluating bias in mixed-mode surveys: the record-check approach, the relative mode-effect approach, and the benchmark survey approach. Exact bias estimation is only possible using the record-check approach, which supposes that true scores are available from an external database (Biemer and Lyberg 2003:289; Groves 2006). A growing number of case studies evaluate mode differences in bias using record checks (Kirchner and Felderer 2013; Körmendi 1988; Kreuter, Müller, and Trappmann 2013; Kreuter et al. 2008; Olson 2006; Sakshaug, Yan, and Tourangeau 2010; Tourangeau, Groves, and Redline 2010) and the approach is also applied to mixed-mode surveys (Fowler et al. 2002; Kreuter et al. 2010; Link and Mokdad 2006; Sakshaug et al. 2010; Voogt and Saris 2005). In practice, however, records are hardly ever available for the survey variables of interest and involve various other practical problems. Records may not coincide with the study time period and may themselves contain errors or be incomplete, survey questions may suffer from specification problems compared to information encoded in records, and there may be problems arising from incomplete matches of respondents to records (Biemer and Lyberg 2003:291; Körmendi 1988; Miller and Groves 1985). Furthermore, records are seldom available to all researchers due to privacy limitations and sometimes can concern very specific sub-populations, such as students (Kreuter et al. 2008).

Vannieuwenhuyze and Loosveldt (2013) suggested evaluating mixed-mode surveys by 'relative mode effects' (Vannieuwenhuyze, Loosveldt, and Molenberghs 2010, 2014). Relative effects are defined as difference in total bias, measurement bias, and selection bias between mode-specific response groups in a mixed-mode design, referred to as overall, measurement, and selection effects. However, relative mode effects cannot

be interpreted in an absolute sense. For example, a relative selection effect in a sequential mixed-mode design does not suggest that selection bias is reduced or increased by the follow-up, but only that different 'types' of respondents are reached.

This problem is addressed when biases are evaluated against a 'preferred' single-mode survey whose measurements are considered valid and which is also considered optimal on selection bias (Biemer 1988; Biemer and Lyberg 2003:287; Körmendi 1988; De Leeuw 2005; De Leeuw et al. 2008; Vannieuwenhuyze 2014; Vannieuwenhuyze et al. 2010). This mode is called the 'single-mode benchmark' (SMB). A relevant application of SMBs is the redesign of repeated cross-sectional surveys to mixed-mode designs. To assure comparability, the change in bias relative to established SMB time-series is an important concern (addressing RQ 1). Furthermore, preventing shifts after assessing selection and measurement effects against the SMB may be possible (RQ 2).

Using the SMB approach, Schouten et al. (2013) and Klausch, Schouten, and Hox (2014) defined differences in bias between a SMB and other modes as 'single-mode effects'. The primary difference to the relative mode effect approach is that effects are defined between single-mode surveys and a reference survey (the SMB), and not between mode-specific response groups in a mixed-mode design. In an empirical study of the Crime Victimization Survey, the authors used a face-to-face survey as SMB against which telephone, mail, and web modes were compared. However, in evaluating sequential mixed-mode surveys (RQ 1 and 2) against a SMB, the impact of the follow-up mode on single-mode effects needs to be considered as well, which has not been accomplished so far. The bias of a mixed-mode survey against the benchmark is called a 'mixed-mode effect'. This extension is provided in the present study.

A limitation of the SMB approach is that single surveys may not be fully appropriate as benchmark. For the case of face-to-face, for example, we assume that due to generally high response rates selection bias is acceptable, but face-to-face measurements may be considered too erroneous to be used as benchmark (e.g., due to social desirability bias; Tourangeau, Rips, and Rasinski 2000:257; Tourangeau and Yan 2007). In particular, self-administered modes, such as 'web', may be more precise for sensitive questions (Kreuter et al. 2008). In this case, a combined benchmark of the selection bias of face-to-face and measurements of web may thus appear superior to the SMB. We call this combined benchmark the 'hybrid-mode benchmark' (HMB). Even though the SMB has a long-standing tradition in evaluating survey errors (Biemer and Lyberg 2003:291), the HMB may be more immediate to survey practice when assuring comparability to time series is not central (e.g., when questions are sensitive or interviewer effects are strong in the SMB). The present paper is the first to consider a HMB besides the SMB. Previous literature also established the general conditions under which unbiased estimation of measurement and selection effects is possible (Klausch et al. 2014; Vannieuwenhuyze and Loosveldt 2013; Vannieuwenhuyze et al. 2014). The primary difficulty in estimation is the confounding of both effects in the overall mode effect. To disentangle the effects, exogenous auxiliary information needs to be available, conditional on which measurements and the selection mechanism into mode-specific response groups are independent (Imbens 2004; Morgan and Winship 2007; Pearl 2009; Rubin 2005). In mixed-mode surveys, finding variables that allow both unconfoundedness and exogeneity appears difficult, however. To address this problem, Schouten et al. (2013) and Klausch et al. (2014) introduced a re-interview design. In the re-interview, repeated measures of survey target variables as well as other auxiliary information are collected.

Conditional on the repeated measures, the unconfoundedness assumption appears more plausible (cf. section 4). The present study re-analyses this data for the case of sequential mixed-mode surveys using both a SMB and HMB.

3 The Crime Victimization Survey case study

The case study was conducted in the context of the Crime Victimization Survey (CVS), administered by Statistics Netherlands in 2011 in an experiment conducted independently from the regular CVS (Klausch et al. 2013, 2014; Schouten et al. 2013). In this section, we argue for possible choices of benchmarks (SMB and HMB) in the case study and provide details on the fieldwork.

Traditionally, the F2F mode has been considered an ideal mode for the CVS. Whereas the CVS has been a F2F survey upon initialization, in past decades it had to be re-designed multiple times using different mixed-mode protocols. However, predecessors of the CVS were F2F and the measurements from F2F are still by many regarded as accurate or desirable. The social control and additional explanations provided by interviewers may increase validity of measurement, whereas high response rates suggest small selection bias. Arguably, F2F surveys often take on this 'benchmark role' in survey research for these and similar reasons. Setting F2F as SMB appeared plausible from this point of view.

For this reason, a split-ballot mode experiment was conducted, where in parallel a single-mode F2F survey was compared to three single-mode designs (telephone, mail, and web; Table 1). In addition, the experiment entailed a sequential mixed-mode component re-approaching the nonrespondents in telephone, mail, and web by F2F. This procedure may yield mixed-mode estimates that are similar to the F2F benchmark and

provide inexpensive designs compared to F2F alone. Evaluating the success of this procedure is the objective of the present case study (following RQs 1 and 2).

Table 1: Response rates and mixture weights in the CVS mixed-mode experiment (weighted[1])

	F2F (SMB) % (n=1,639)	Telephone[a] % (n=1,658)	Mail % (n=1,760)	Web % (n=1,746)
Single-mode	64.5	48.6	49.3	29.0
Mixed-Mode	-	65.2	67.3	59.7
Prop. Single-mode Resp. (π)	-	74.6	73.3	48.6
Prop. Follow-up Resp. ($1-\pi$)	-	25.4	26.7	51.4
Re-interview (full 2nd wave)	53.9	50.6	52.4	51.4

a. The telephone response rates are taken against the net sample of all sampled units (including persons without known telephone number).

Each mode-specific survey was based on a probability sample drawn from the national register[1] (a person sampling frame). The sample size was chosen such that minimal observable total single-mode bias against F2F is equal to the required precision of the CVS at the national level. The regular CVS is much larger because of detailed publications for a range of subpopulations. The fieldwork was conducted in the time from April until June 2011. First, all respondents received mailed pre-notifications. In the two self-administered conditions, these letters contained a paper questionnaire with return envelope or information on how to access the survey online. In the interviewer modes, the contact attempt by an interviewer on the phone or in person was announced. The fieldwork period of this first wave was four weeks and, subsequently, the nonresponse follow-up in F2F was administered. The F2F single-mode survey remained active for the same period as telephone, mail, and web.

[1] The original sample size was 2,200 persons in each mode-specific condition. However, only approx. 80% of the sample was followed up by F2F for cost-related reasons. In this sub-sample, all persons without registered telephone number were followed up, which lead to a slight over-representation of non-telephone households in the sub-sample. The response rates and subsequent analyses use design weights to yield unbiased estimates. In general, the deviations between weighted and unweighted results were small however.

The F2F survey showed the largest response rates of the four modes (64.5%, Table 1). The response rates in telephone, mail, and web were lower than in F2F, but the F2F follow-up increased the response rates in all modes (up to 65.2, 67.3, and 59.7%, respectively). Since the single-mode response rate in the web mode was lower than in telephone or mail conditions, the relative proportion of F2F respondents in the web mixed-mode sample was substantially higher (51.4% vs. 25.4% and 26.7%, fourth row Table 1). These relative proportions (called π) suggest that the estimates from the mixed-mode web sample may be impacted more strongly by the follow-up than mixed-mode mail or telephone. This parameter is, therefore, relevant in estimation and interpretation of results, discussed in more detail in section 4 and 5.

The high response rate of F2F gives reason for choosing F2F as selection benchmark. In addition, Klausch, Hox, and Schouten (2016) showed that representativeness of F2F on socio-demographic background variables was high in the CVS experiment. Nevertheless, the F2F mode may not always be an 'optimal' mode of measurement. For the case of the CVS it can be argued that the questionnaire contains a large number of attitudinal and sensitive questions (cf. appendix A), which may be particularly susceptible to stronger measurement bias due to the presence of an interviewer (cf. section 2). Measurement in the anonymous situation of self-administration (e.g., web) may be more appropriate. However, web surveys are traditionally looked at critically with regard to their selective properties due to, for example, lower response rates (29.0% in the present study) and incomplete population coverage. Therefore, a web-F2F hybrid-mode benchmark was evaluated as an alternative to the F2F single-mode benchmark, which assumes F2F as the benchmark for selection and web as the benchmark for measurement.

It should be noted that the choice of web as measurement benchmark potentially may be as flawed as choosing F2F or any other mode as measurement benchmark. Whereas web may indeed suffer from less social desirability bias, it may produce other measurement problems. For example, interviewers may act as instance of social control and provide motivation to respondents, keeping satisficing and response effects low (Klausch et al. 2013). Web interviews lack this advantage due to self-administration. Further disadvantages of web surveys, such as respondent ability (e.g., problems with eye sight and literacy), have been described in the literature as sources of measurement error and satisficing (Couper 2000; Couper, Traugott, and Lamias 2001; Fricker et al. 2005; Krosnick 1991). In practice, the choice of benchmark is, therefore, based on a plausibility argument – that is, in the absence of any other information on measurement accuracy of a particular mode and survey (questionnaire) it has to be plausibly argued for or against a particular mode as measurement benchmark. In this case study, we present results for both a F2F measurement benchmark (SMB) as well as a HMB with web as measurement benchmark. We note that for the CVS – a government survey – social desirable responding to authorities may be a substantial problem on a great number of questions. This conjecture lets the web mode appear a logical alternative choice as measurement benchmark. Furthermore, we do not consider questions in the analyses that typically are associated with strong measurement problems in web surveys, such as open ended questions or questions with many answer categories. Nevertheless, the measurement benchmark is an assumption made by the analyst and we therefore compare the results of the SMB and HMB in the case study, discussing differences in conclusions about the optimal design depending on measurement benchmark choice. In the

discussion, we also return to the important aspect of choosing benchmarks for the evaluations of biases.

Finally, as noted in section 2, also the respondents in the mixed-mode design received a follow-up in the face-to-face mode (administered in parallel to the nonresponse follow up in face-to-face). The response rates across the full re-interview (2nd wave in face-to-face) are shown in table 1. How this data is used to estimate measurement and selection bias is discussed in more detail in the next section.

4 Definition and Estimation of Single- and Mixed-Mode Effects

In this section, we explain how single- and mixed-mode bias components required to answer the RQs were estimated using the single-mode (SMB) and hybrid-mode (HMB) benchmark. We first define all biases following the total survey error (TSE) framework, but in a mixed-mode context some extensions of the TSE notations are necessary (Biemer 2010; Biemer and Lyberg 2003; Groves et al. 2010:48; Groves and Lyberg 2010). Subsequently, we explain estimation of the biases against the SMB and HMB.

4.1 Defining single- and mixed-mode bias of the sample mean

To simplify notation, we define the bias of a mixed-mode telephone - F2F survey, but the definitions for the other modes follow likewise. We are interested in the bias of the estimator of the response sample mean of a continuous or discrete survey variable *Y*. Let $\hat{\mu}^{tel}$ be the estimator of the response sample mean in the telephone mode and let Y^{tel} denote the measurement of Y by telephone. Then the expected value of $\hat{\mu}^{tel}$ is

$$E\left(\hat{\mu}^{tel}\right) = E(Y^{tel}|S^{tel} = 1) \coloneqq \mu^{tel},$$

where the binary random variable S^{tel} represents the response mechanism of telephone and $S^{tel} = 1$ indicates the group of respondents. The total bias (TB) of $\hat{\mu}^{tel}$ against the population mean $\mu = E(Y)$ is (addressing RQ 1)

$$B(\hat{\mu}^{tel}) = TB^{tel} = \mu^{tel} - \mu. \quad (1)$$

Of course, μ is not observed without external validation data. When using the SMB and HMB approach, the population mean, therefore, will be substituted by an estimator that is least biased based on the respective benchmark assumptions (section 4.2).

The TB can now be decomposed into its systematic components, where we distinguish selection (SB) and measurement bias (MB). In limiting our elaboration to these biases, we also assume that other sources of bias discussed under the TSE framework, such as specification and data processing error, are negligible or at least equal across modes. Then measurement and selection bias can be said to add up to 'total bias'. Single-mode SB and MB follow as (addressing RQ 2)

$$SB^{tel} = E(Y|S^{tel} = 1) - \mu \quad (2)$$

and

$$MB^{tel} = E(Y^{tel}|S^{tel} = 1) - E(Y|S^{tel} = 1). \quad (3)$$

It can be seen that, following the TSE framework, SB is defined as the difference in sample mean of the true score and the population mean, whereas MB represents the

difference of sample mean of the true score and measured mean answers[2]. MB and SB add up to TB.

In the mixed-mode survey, the answers of follow-up respondents in F2F are added to the single-mode response set. The mean from the pooled mixed-mode survey can then be regarded as the mean of a mixture distribution of Y^{tel} and Y^{F2F},

$$\mu^{MM} = \pi\mu^{tel} + (1-\pi)E(Y^{f2f}\big|S^{tel} = 0, S^{f2f} = 1), \quad (4)$$

where the mixture constant π is defined by the expected proportion of single-mode respondents introduced in section 3 (cf. Table 1 for sample estimates from the case study). The total bias of the mixed-mode mean now can be expressed as (addressing RQ 1)

$$B\big(\hat{\mu}^{MM}\big) = TB^{MM} = \pi TB^{tel} + (1-\pi)TB^{FU}, \quad (5)$$

where

$$TB^{FU} = E(Y^{f2f}\big|S^{tel} = 0, S^{f2f} = 1) - \mu \quad (6)$$

represents the total bias of the follow-up mode. It can be seen that the relative impact of single-mode biases is reduced by the size of the response proportion π, but the sign and size of bias of the follow-up sample is crucial for the overall mixed-mode TB (for a numerical example see Bethlehem and Biffignandi 2011:259). The mixed-mode SB and MB depend on the (weighted) follow-up mode bias in the same manner, i.e. (addressing RQ 2)

$$SB^{MM} = \pi SB^{tel} + (1-\pi)SB^{FU}, \quad (7)$$

where

[2] It should be noted that selection bias treats nonresponse and coverage bias as a compound (Vannieuwenhuyze and Loosveldt 2013), which allows studying errors of modes by two fundamentally different processes (i.e., measurement and selection). Schouten et al. (2013) discuss estimation of single-mode coverage and nonresponse bias against a F2F benchmark.

$$SB^{FU} = E(Y|S^{tel} = 0, S^{f2f} = 1) - \mu, \quad (8)$$

and

$$MB^{MM} = \pi MB^{tel} + (1 - \pi)MB^{FU}, \quad (9)$$

where

$$MB^{FU} = E(Y^{f2f}|S^{tel} = 0, S^{f2f} = 1) - E(Y|S^{tel} = 0, S^{f2f} = 1). \quad (10)$$

In the next section, we discuss how these biases can be estimated. We address estimation against a SMB and HMB in turn.

4.2 Estimation of single- and mixed-mode effects against the SMB

Setting a benchmark implies choosing two components. First, a mode is chosen that substitutes the true scores of *Y* by the measurements of the benchmark mode. These observed scores are close or equal to the true scores (cf. section 2). Second, a mode is chosen that evokes acceptable selection bias. While acknowledging that the selection benchmark may not be free of selection bias (because itself suffers from unit nonresponse), it is considered the mode with best selection properties for the variable at hand. In principle, the measurement and selection benchmark modes may differ, in which case the combined benchmark is called hybrid. The SMB is the special case when both components are taken from the same mode.

In the present study, the F2F survey represents the SMB. In doing so, we set the F2F answers as the true score to assess comparability of single- and mixed-mode estimates of telephone (or the other modes, RQ 1) as well as reasons for incomparability (RQ 2). Thus, the population parameter μ is substituted by the F2F

mean $E(Y^{f2f}|S^{f2f}=1)$. From the split-ballot mixed-mode design of the case study, the single- and mixed-mode total biases can now be estimated as (addressing RQ 1):

$$T\hat{B}_{SMB}^{tel} = \hat{E}(Y^{tel}|S^{tel}=1) - \hat{E}\left(Y_1^{f2f}\middle|S_1^{f2f}=1\right), \quad (11)$$

and

$$T\hat{B}_{SMB}^{MM} = \pi T\hat{B}_{SMB}^{tel} + (1-\pi)\left(\hat{E}\left(Y_2^{f2f}\middle|S^{tel}=0, S_2^{f2f}=1\right) - \hat{E}\left(Y_1^{f2f}\middle|S_1^{f2f}=1\right)\right). \quad (12)$$

Here, we use the operator $\hat{E}$ to denote an estimator of the response means using the respective mode-specific measurements *Y* and response mechanisms (groups) *S*, so that, for example, $\hat{E}(Y^{tel}|S^{tel}=1) = \hat{\mu}^{tel}$ denotes the simple telephone sample response mean ($S^{tel}=1$) over telephone answers (Y^{tel}). This notation is necessary to denote clearly which information is used to estimate the mean components of the single-mode total bias (formula 1) and mixed-mode total bias (formula 5) against the benchmark. In addition, we added the indices '1' and '2' for the F2F measurements and response mechanisms to distinguish the single-mode F2F comparison sample and the follow-up in F2F to nonrespondents in telephone. Thus, $\hat{E}\left(Y_2^{f2f}\middle|S^{tel}=0, S_2^{f2f}=1\right)$ denotes the response mean of follow-up respondents ($S^{tel}=0, S_2^{f2f}=1$) on F2F answers (Y_2^{f2f}), whereas $\hat{E}\left(Y_1^{f2f}\middle|S_1^{f2f}=1\right)$ denotes the single-mode F2F response mean.

The bias estimates (11) and (12) can also be called total mode effects (alternatively, 'overall mode effects' or 'mode system effects'; Biemer 1988; De Leeuw 2005; Schouten et al. 2013; Vannieuwenhuyze and Loosveldt 2013) and in answering RQ 1 it is important to distinguish between the single- (11) and the mixed-mode effect (12). The primary difficulty in estimating the single-mode selection and measurement effect com-

ponents of the total effect (addressing RQ 2) is that after substitution of Y_1^{f2f} as 'true score' Y in (2) and (3) the outcome

$$E\left(Y_1^{f2f} \middle| S^{tel} = 1\right)$$

is not observed in a simple comparative mixed-mode design. This can be seen when illustrating the missing data pattern for the SMB and mixed-mode samples (Figure 1, left hand).

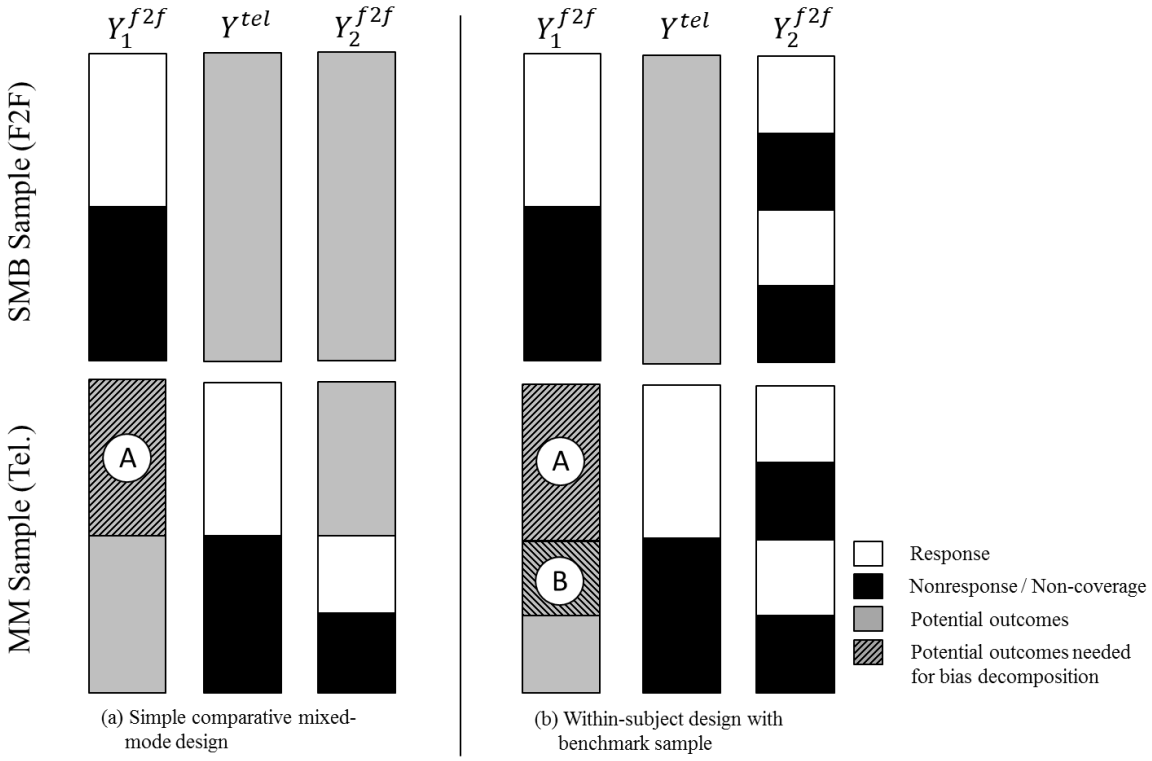

Figure 1: Missing data pattern of a simple comparative mixed-mode design (a) and a within-subject design (b) for estimating bias components against a SMB (Example of a telephone-F2F mixed-mode design with F2F benchmark)

In the SMB sample, respondents provide answers on Y_1^{f2f} (white area) and unit nonresponse (black area). Also in the mixed-mode sample, the first part of the mixed-mode survey leads to telephone answers Y^{tel} and nonresponse, but additionally, the F2F follow-up of nonrespondents in telephone leads to responses Y_2^{f2f}. The grey areas indi-

cate the part of missing information that is not observed by design. These outcomes can be called 'potential' following terminology introduced by Rubin (Klausch et al. 2014; Rubin 1978, 2005; Vannieuwenhuyze and Loosveldt 2013).

$E\left(Y_1^{f2f} \middle| S^{tel} = 1\right)$ represents the mean of a subset of potential outcomes (i.e. telephone respondents), indicated by shaded area A. In the simple comparative design (left hand, Figure 1), additional auxiliary information is needed to estimate this mean (Vannieuwenhuyze et al., 2010, 2013). In the case study, we, therefore, applied a re-interview design called within-subject design (right hand; Schouten et al., 2013; Klausch et al., 2014; cf. re-interview response rates in table 1). In the design, questions from the first occasion were repeated in the mixed-mode and the SMB samples. Using this data, we estimated $E\left(Y_1^{f2f} \middle| S^{tel} = 1\right)$ by multiply imputing potential comes Y_1^{f2f} followed by taking the mean of imputed Y_1^{f2f} over the telephone respondents (Little and Rubin 2002; Rubin 1987; Schafer and Graham 2002). Peytchev (2012) discusses advantages of imputation for estimation of nonresponse and measurement bias in similar survey designs.

In doing so, we assume that potential outcomes and unit nonresponse at the re-interview are missing at random[3] (MAR) given observed outcomes Y^{tel} and Y_2^{f2f}. The repeated measures play an important role in making this assumption plausible (Schouten et al., 2013). The telephone measurements Y^{tel} are replications of Y_1^{f2f} given a measurement effect, but the partial correlation of both variables is unknown (cf. Figure 1,

[3] For the benchmark mode, we assume
$P\left(Y_1^{f2f} \middle| {}^{obs}Y_1^{tel}, {}^{obs}Y_2^{f2f}, S_1^{f2f} = 1\right) = P\left(Y_1^{f2f} \middle| {}^{obs}Y_1^{tel}, {}^{obs}Y_2^{f2f}, S^{tel} = 1\right) =$
$P\left(Y_1^{f2f} \middle| {}^{obs}Y_1^{tel}, {}^{obs}Y_2^{f2f}, S^{tel} = 0, S_2^{f2f} = 1\right)$. This assumption is equivalent to requiring the response mechanism of F2F response relative to mixed-mode telephone-F2F response to be conditionally unconfounded with Y_1^{f2f} (Imbens 2004).

right side). However, from the re-interview data, Y_2^{f2f}, partial correlations to both Y^{tel} and Y_1^{f2f} can be estimated[4]. In addition, Y_2^{f2f} are closely related to these variables, which is an important criterion for the adequacy of auxiliary variables in causal inference and nonresponse adjustment (Bethlehem, Cobben, and Schouten 2011:249; Schafer and Kang 2008). The details of the practical implementation of the imputation procedure are provided in section 4.4.

After imputation, the potential outcomes $E\left(Y_1^{f2f}\middle|S^{tel}=1\right)$ were estimated by the mean of imputed F2F outcomes for the telephone response sample (formulas 2 and 3). Mixed-mode selection and measurement effects against the SMB were estimated in an analogous way. This task required estimating the mean of benchmark answers of follow-up respondents (formulas 8 and 10) using the second potential outcome

$$E\left(Y_1^{f2f}\middle|S^{tel}=0,S_2^{f2f}=1\right),$$

which is indicated by shaded area B in figure 1. These potential outcomes were imputed as part of the estimation procedure under the MAR assumption discussed above.

4.3 Estimation of single- and mixed-mode effects against the HMB

As discussed in section 3, we chose to use the web mode as a measurement benchmark for the HMB case, while keeping F2F as selection benchmark. Likewise the SMB, now web measurements substitute the true score, but the selection mechanism of F2F is still deemed optimal, so that the population parameter μ is substituted by the potential outcome $E(Y^{web}|S^{f2f}=1)$. From the case study data, web measurements were availa-

[4] If only Y_1^{tel} was available for imputing potential outcomes, unbiased estimation was only possible, if the response mechanism to F2F or mixed-mode telephone was strongly ignorable (Y_1^{f2f} was missing completely at random), because Y_1^{f2f} and Y_1^{tel} do not overlap (i.e. their partial correlation is unknown).

ble in the mixed-mode sample. Figure 2 demonstrates the location of these potential outcomes as shaded area C (the figure omits nonrespondents in the SMB sample and double-nonrespondents in the mixed-mode samples).

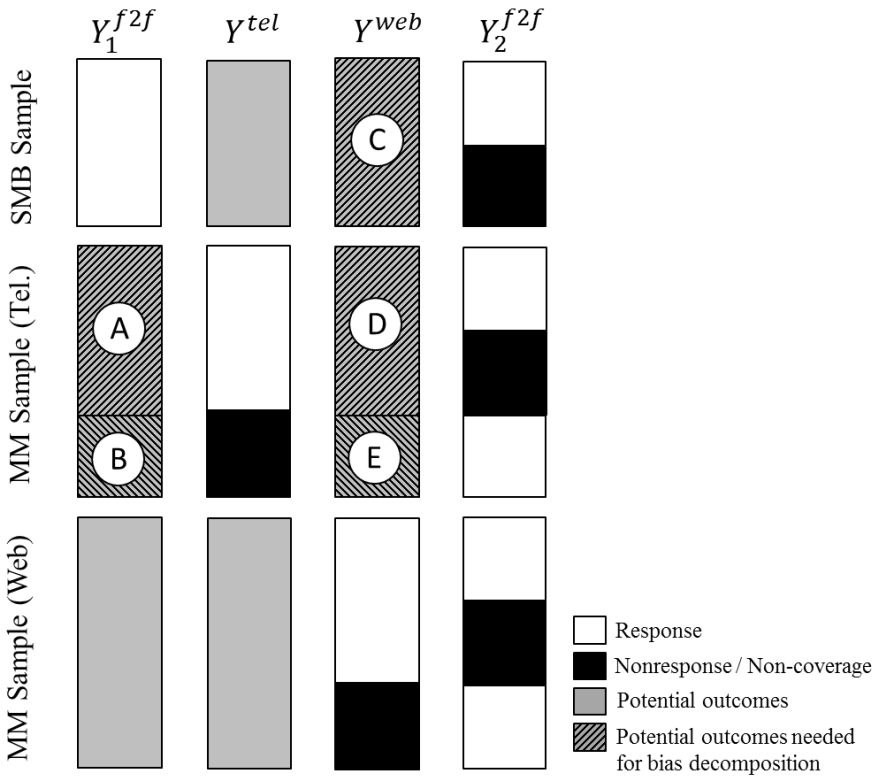

Figure 2: Missing data pattern of an extended within-subject design with two different mixed-mode samples (telephone and web) for use in the HMB case (nonrespondents in the SMB sample and double-nonrespondents in the mixed-mode samples omitted)

To estimate single-mode selection- and measurement effects of the telephone sample against the HMB, the potential outcomes $E(Y^{web}|S^{tel}=1)$ are required (shaded area D). Figure 2 demonstrates that these outcomes are counterpart to the potential outcomes $E\left(Y_1^{f2f}\middle|S^{tel}=1\right)$ required for the SMB case (area A). Finally, to estimate the mixed-mode effects the benchmark mean for follow-up respondents needs to be estimated (area E) analogous to area B in the SMB case. As in the SMB case, multiple im-

putation was applied to estimate these potential outcomes using the case study data. The details of this procedure are provided next.

4.4 Practical implementation of the multiple imputation procedure

In the context of the missing data pattern shown in Figure 2, auxiliary information Y_2^{f2f} is used to impute potential outcomes, but Y_2^{f2f} itself suffered from unit nonresponse (cf. re-interview response rates, Table 1). Both missing data problems may be solved simultaneously by multiple imputation under the MAR assumption discussed in section 4.2 (Rubin 1987; Schafer 1997; Schafer and Graham 2002). A sequential regression approach was applied, called multiple imputation by chained equations, using the software '*MICE*' (van Buuren 2012:109; van Buuren et al. 2006; van Buuren and Groothuis-Oudshoorn 2011; Raghunathan et al. 2001; Rubin 2003). In total, thirty different variables were imputed per mode concerning different aspects of the CVS, in particular the social quality and problems of the neighborhood, frequency of insecurity feelings, police contact and evaluation, and victimization[5] (see overview in appendix A).

In *MICE*, prediction models with an appropriate link function for each imputed variable were specified. Since the CVS variables were measured on polytomous, dichotomous, or interval scales, we applied multinomial, logistic and normal regression models, respectively[6]. A crucial element of the multivariate imputation is the selection of predic-

[5] Victimization is surveyed on multiple variables and aggregated to two summary scores (past victimization in the last year [yes / no] and a count of victimization incidences).

[6] For the ordered polytomous scales (e.g., rating scales) proportional odds models could be used instead. These often led to convergence problems of the model fitting algorithm. In this case, *MICE* internally applies multinomial regression instead (van Buuren 2012:76). Since the problem occurred regularly in our data, we chose to model all variables directly by multinomial regression models.

tor variables. Given the large number of possible predictors in the data set, cautioning of over-specified models was important. To restrict the number of predictors, we applied the following procedure:

- Mode-specific potential outcomes of *Y* at occasion 1 were predicted by their repeated measure Y_2^{f2f}.
- Y_2^{f2f} were predicted by their four counterparts in the modes at the first occasion.
- Any Y_2^{f2f} variable from the follow-up exceeding a medium bivariate association (Cramér's *V*>.30) with a *Y* variable at the first occasion (e.g. Y^{tel}) was included. It should be emphasized that conditioning on these additional Y_2^{f2f} variables can further strengthen the MAR assumption.
- Eight socio-demographic background characteristics were available as additional predictors from the national population register for all units[7]. Any of these background characteristics exceeding a small association of *V*>.15 was included as predictor for any *Y* variable.

Fifty data sets were multiply imputed. The proportion of missing data was high in the present study, due to the fact that potential outcomes were imputed across four samples. However, the fraction of missing information, a model-based estimate of missingness (van Buuren 2012:41; Rubin 1987), was below 50% in most cases. This fraction suggests that when using fifty multiply imputed datasets, estimates of total variance are

[7] I.e., gender, age, income, household-size, civil status, nationality, urbanization of living area, and living in one of three large Dutch cities. The income variable shows a small amount of missing cases, which were treated as category in the present analyses.

precise[8]. To estimate the within-imputation variances, all effects were bootstrapped by 1,000 iterated draws within each imputed data set. The within- and between imputation variances were pooled to the total variance using Rubin's rules (Schafer 1997:109–110). Two sided significance tests were executed using t-tests with adjusted degrees of freedom.

Another relevant issue in estimation was the treatment of item nonresponse due to "don't know" (DK) or refused answering. In principle, there is not a unique way of handling this problem, because DK answers may be either regarded as missing information or substantial answer (i.e., as an answering category). The results presented are based on imputed item nonresponse as part of the missing data correction. Results turned out robust, because treating DK as answering category yielded very similar findings to the results presented here.

5 Results

In this section, we answer RQs 1 and 2 for the case of the CVS experiment, discussing the F2F SMB and the web/F2F HMB separately. We note again that these analyses are based on different benchmark assumptions and research interests, as exposed in the previous sections. After presenting results, we discuss implications for the CVS under both perspectives.

[8] The fraction of missing information determines the strength of inflation of total variance estimates by the equation: $T_m = (1 + \frac{\gamma_0}{m})T_\infty$, where T_∞ is the ideal variance under infinite imputations, m is the number of imputations and $\gamma_0$is the fraction of missing information (van Buuren 2012:49; Rubin 1987:114). For a fraction of missing information of 50%, 50 imputations lead to approx. 0.5% higher standard error estimates than under the ideal situation (1% higher variance). The maximum fraction of missing information across all variables, modes, and effects was .771 for the case of the mixed-mode measurement effect of mail in the HMB case. Also in this case the total variance estimate is still precise (approx. 0.77% higher standard error estimates).

5.1 Effects against the F2F single-mode benchmark

RQ 1 (*Is the mixed-mode survey needed or would a single-mode survey suffice in terms of accuracy?*) requires estimating the total effect of using a different focal mode (web, mail, telephone) than the F2F benchmark and assessing the impact of the mixed-mode design on the single-mode effect. Twenty of the thirty included CVS variables were measured on polytomous answering scales and dichotomized similar to reporting by Statistics Netherlands (cf. appendix A for an overview). The dichotomized target statistic is a proportion (e.g., the proportion "agree or completely agree"). Four further variables are summary scales measured on interval level, and one represents a count of victimizations in the past year.

Table 2 presents a summary of the significant single- and mixed-mode effects on these variables. A majority of variables showed total effects against F2F in web (20) or mail (16), but fewer variables were affected in telephone (7). The counts beneath these numbers inform about the impact of the mixed-mode follow-up. We distinguish four possibilities: a mixed-mode effect is insignificant after the follow-up ($TB^{MM} = 0$) or it is still significant and, if it is significant, it may decrease significantly[9] ($|TB^{SM}| > |TB^{MM}|$), may stay equal ($|TB^{SM}| = |TB^{MM}|$) or may increase ($|TB^{SM}| < |TB^{MM}|$). It can be seen that for web and mail the follow-up was very effective, decreasing total bias against F2F in 18 (of 20) and 12 (of 16) cases, respectively. Half of these cases even reached insignificant level after the follow-up.

[9] To evaluate significance of change between single and mixed-mode total effects, we bootstrapped the statistic $\Delta = TB^{SM} - TB^{MM}$ and tested its difference from zero. If the test was significant, we inspected, whether bias was increased or decreased relative to the benchmark, by evaluating $\left|T\hat{B}^{SM}\right| < \left|T\hat{B}^{MM}\right|$ against $\left|T\hat{B}^{SM}\right| > \left|T\hat{B}^{MM}\right|$ for the final effect estimates.

Table 2: Count of significant and non-significant single-mode total effects and the change induced by the F2F follow-up (mixed-mode effects) for the SMB and the HMB case (significance tests on p<.05)

		F2F Benchmark (SMB)			Web/F2F Benchmark (HMB)			
Single-Mode	Mixed-Mode	Web	Mail	Tel.	Web	Mail	Tel.	F2F
$TB^{SM} = 0$		10	14	23	30	20	9	7
	$TB^{MM} = 0$	9	13	22	20	18	9	-
	$TB^{MM} \neq 0$	1	1	1	10	2	0	-
$TB^{SM} \neq 0$		20	16	7	0	10	21	23
	$TB^{MM} = 0$	9	4	1	0	2	1	-
	$\|TB^{SM}\| > \|TB^{MM}\| \neq 0$	9	8	0	0	0	3	-
	$\|TB^{SM}\| = \|TB^{MM}\| \neq 0$	2	4	6	0	4	17	-
	$\|TB^{SM}\| < \|TB^{MM}\| \neq 0$	0	0	0	0	4	0	-
Total		30	30	30	30	30	30	30

Table 2 does not allow an assessment of the size of effects, however. For this purpose, and answering RQ 2 below, we employ three scatterplots of single-mode against mixed-mode effects showing estimates for the 25 variables available on dichotomous scale (Figure 3, upper row). The lower row of scatterplots presents t-statistics for each variable and may be evaluated against critical values from the t-distribution. Critical values for a two-sided test (p<.05) are provided by horizontal (single-mode) and vertical (mixed-mode) dashed lines[10]. The diagonal line in all plots has slope one (i.e., it is not a regression line). Deviations from the line, therefore, imply change in effects by the follow-up.

Figure 3 about here

[10] T-tests of a zero-effect null hypotheses based on multiply imputed data require a variable-specific adjustment of degrees of freedom that is based on the between and within-imputation variance (Rubin 1987; Schafer 1997:110). For this reason, all tests involve variable-specific critical values. For the present data the maximum critical value across all variables and modes is depicted as dashed line (approx. +/-1.975), which, however, did only deviate marginally from critical values of large sample t-tests (i.e. normal distribution critical values, +/-1.965, p<.05).

Consider first the plot of single- against mixed-mode total effects (upper left hand). Single-mode total effects of web and mail are substantially larger than for telephone in many cases. Moreover, the seven significant single-mode total effects for telephone (Table 2) are found to be of smaller magnitude than for mail and web. Secondly, the impact of the mixed-mode follow-up is apparent for both web and mail, but not telephone, as estimates are moved towards zero mixed-mode effects (i.e., the horizontal axis). This effect is particularly pronounced for web suggesting that the mode profits more strongly from the F2F follow-up in reducing total effects.

RQ 2 asks about the sources of the total effects we identified (*What are the major systematic sources of error in single-mode surveys and how are they impacted by the mixed-mode follow-up?*). This question is addressed by the middle (selection effects) and right plots (measurement effects). It is immediately clear that selection effects were very small, and that measurement effects were the dominant component in creating effects between the SMB and the three focal modes. It is important to emphasize that the reduction in measurement effects by the F2F follow-up seemed to be effective, because the follow-up is conducted in the benchmark mode. The follow-up mode measurement effect (formula 10) was indeed small in all cases (not shown here). The reduction in single-mode measurement effects is, therefore, dominated by the term $\hat{\pi}M\hat{B}^{SM}$ (formula 9). The web mixed-mode sample showed a substantially higher amount (51.4%) of F2F follow-up respondents suggesting that single-mode measurement effects were, roughly, reduced by this factor, whereas mail and telephone were impacted less strongly (26.7%, 25.4%; cf. Table 1).

5.2 Effects against the hybrid-mode benchmark

The HMB takes web measurements as benchmark while allowing for the selection mechanism of F2F. Significance tests of the total effects against the HMB are provided on the right side of table 2. In addition to the three mixed-mode designs, effects for the single-mode F2F survey are shown against the HMB. It is apparent that for telephone and F2F, a large number of variables (21 and 23, respectively) showed significant total mode effects, whereas for mail fewer variables (10) reached significant level. There were no significant total effects for web. With respect to web, it should be noted that only single-mode selection effects could have caused a total effect, given that web measurements are used as benchmark.

The F2F follow-up to the three modes was mainly ineffective or harmful. For web, it was even very harmful, increasing total effects in 10 cases. Similarly in mail, it increased total effects on 6 variables, while only reducing it on 2. It was mainly ineffective to reduce the telephone total effect against the hybrid web-F2F benchmark.

For a more detailed picture, we consider the scatterplots of the three single- and mixed-mode effects for the HMB case (Figure 4). Since the selection benchmark did not change, selection effects against F2F were small and insignificant, likewise the SMB case. However, telephone showed large measurement effects against the web measurement benchmark and the F2F follow-up was ineffective. Mail showed smaller single-mode measurement effects against web or no effects. However, it can be seen that the F2F follow-up increased measurement and total effects on many mail variables (cf. points below the diagonal line) reflecting that the F2F follow-up was not beneficial to the mail single-mode bias.

Figure 4 about here

In addition, it can be seen that web, as measurement benchmark, does not exhibit single-mode measurement effects. Measurement effects are caused by the F2F follow-up, however, reflecting that F2F measurements showed a bias against the web benchmark. For this reason, the mixed-mode total effect of web is determined by the follow-up measurement effect of F2F against the HMB (i.e., $(1-\hat{\pi})M\hat{B}^{FU} = .514M\hat{B}^{FU}$).

5.3 Evaluation of effects by variable groups

Finally, we consider measurement and total effects by the type of variables and statistics reported in the above analyses (Table 3). Significance of measurement and total effects is evaluated separately. It can be seen that for some cases a total effect did not imply a measurement effect (e.g., 11 of 17 significant web total effects imply a measurement effect). For these cases, a clear conclusion about the source of total bias against the benchmark cannot be drawn (i.e., a selection effect may represent an alternative explanation). However, for the majority of variables, a total effect did imply a measurement effect, reflecting the observations from figures 3 and 4. Also considering the clear measurement effect pattern of both figures, it is plausible that measurement effects underlay also the insignificant cases, where a total effect is observed but remains within observable differences.

We found that, regardless of mode and benchmark, all CVS variables may have been subject to a measurement effect, regardless of type of variable and answering scale. In particular the two groups of questions on the social quality and problems of the neighborhood appeared susceptible to measurement effects. However, not all questions of any given group show clear measurement (or total) effects. We may conclude that

measurement effects appear to be a general, but still a question-dependent phenomenon in the CVS. However, the strong presence of Web and Mail measurement effects against F2F, and telephone and F2F effects against the HMB (web measurements) is again evident.

Table 3: Count of significant single-mode measurement / total effect estimates against the SMB and HMB by variable types (significance tests on p<.05)

		F2F Benchmark (SMB) (Sig. MB^{SM} / TB^{SM})			Web/F2F Benchmark (HMB) (Sig. MB^{SM} / TB^{SM})			
	No. of items	Web	Mail	Tel.	Web	Mail	Tel.	F2F
Proportion statistics								
Social quality[a]	9	5 / 6	8/ 8	2 / 2	0 / 0	2 / 2	7 / 6	7 / 7
Neighbourhood problems[b]	8	6 / 7	4/ 4	1 / 2	0 / 0	5 / 6	8 / 8	7 / 7
Insecurity feelings[c]	4	0 / 2	1/ 1	0 / 0	0 / 0	0 / 0	3 / 2	2 / 2
Police contact (yes / no)	1	0 / 1	0 / 0	0 / 1	0 / 0	0 / 0	0 / 0	0 / 0
Police evaluations[d]	2	0 / 1	0 / 1	0 / 1	0 / 0	0 / 0	1 / 1	1 / 1
Victim of crime[e]	1	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Total	25	11 / 17	13 / 14	3 / 6	0 / 0	7 / 8	19 / 17	17 / 17
Mean statistics								
No. of victimizations[f]	1	0/ 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Quality of life rating[g]	1	0/ 0	1 / 0	0 / 0	0 / 0	0 / 0	0 / 1	0 / 0
Neighbourhood scales[h]	3	2 / 3	2 / 2	1 / 1	0 / 0	1 / 2	3 / 3	3 / 3
Total (incl. dichotomous)	30	13 / 20	16 / 16	4 / 7	0 / 0	8 / 10	22 / 21	20 / 20

For full details on all items see Appendix A.
a. Likert scale items: % (completely) agree (5 answering categories)
b. Frequency scale items: % (frequently or sometimes) (3 answering categories)
c. Insecurity feelings: 2 items (% yes), 2 items (% frequently or sometimes)
d. % very satisfied or satisfied
e. % victim in the past 12 months (aggregated across multiple items)
f. Count of victimization past 12 months (aggregated across multiple items)
g. Score on 10-point scale from very low (1) to very high (10)
h. Aggregated summary indices based on multiple social quality and neighbourhood problems items.

The only characteristics which were not affected by measurement across modes were two 'victimization' variables (victim of crime past 12 months / count of victimization), which represent key statistics in the CVS. The insensitivity to effects is generally a positive result. However, it should be noted that not all standard victimization questions could be included in the re-interview design and the two index variables are based

on shortened versions of the standard statistic. For this reason, the present findings about victimization should be interpreted with care.

5.4 Conclusions for the CVS

In drawing conclusions for the CVS, it is important to recall the objectives of the SMB and HMB. In the SMB case, both measurement and selection mechanisms are taken from the same mode (F2F, in this study). F2F can be considered a historical benchmark for the CVS, which was a F2F survey upon first introduction. This study revealed that when using the web or the mail mode instead of F2F, it is impossible to avoid a strong change in statistics for a large number of CVS variables (RQ 1). The reason for these effects was an increase in measurement bias in web and mail relative to F2F. For telephone, measurement effects were also present, but on much smaller scale and in smaller number. In conclusion, when F2F is the desired benchmark estimate, use of single-mode web and mail should be avoided. Use of telephone may be viable when accepting some smaller systematic changes.

The classical motivation of a sequential F2F follow-up is reducing single-mode selection bias (RQ 2). Selection effects against F2F, however, were not identified on a statistically significant level in this study. Still, estimates from the web and mail mixed-mode surveys were often closer to the SMB than the single-mode estimate alone, because the F2F follow-up provided measurements that were very similar to the single-mode F2F benchmark, as can be expected. To achieve this effect, a F2F follow-up would be chiefly desirable for both modes, but not for reducing single-mode selection effects. However, there remain a number of relatively large total effects suggesting that this procedure cannot fully compensate the measurement bias created by web and mail.

The response proportion π determines the strength of follow-up mode impact. In future research it is important to evaluate the mix of biases and the role of π further.

The objective of the HMB is to optimize a benchmark with respect to both measurements and selection mechanisms. Under the assumption that web measurements are superior to F2F, e.g. due to anonymous answering, we used web as measurement benchmark instead of F2F. This change strongly affects conclusions about the CVS. Since we did not identify any selection effects against F2F, the optimal mode would now be a single-mode web survey. Furthermore, using telephone or F2F would suggest increasing measurement bias and should be avoided. However, the mail mode showed only smaller effects against the HMB. These were primarily limited to a single group of questions ('neighborhood problems', Table 4). In many cases, mail may, therefore, evoke similar estimates as web. However, for both mail and web, the F2F follow-up would only introduce measurement bias. Therefore, a mixed-mode design involving F2F should be avoided. Moreover, given these findings web and mail may be compatible for sequential mixed-mode surveys themselves. In the absence of selection effects, this may seem unnecessary. However, web yielded a small response rate (29.0%), which may be raised by including a mail follow-up, for example (Millar and Dillman 2011). The mode effects in this design could not be evaluated in the present study.

In sum, these results reflect theoretical and empirical arguments in the literature that self-administered and interviewer modes often form a dichotomy with respect to measurement bias (Klausch et al. 2013; De Leeuw 1992, 2008), but also suggest that there may be exceptions to the rule and mode effects remain question dependent phenomena. A question-specific evaluation of effects is, therefore, necessary for any

mixed-mode survey. In the CVS, it was eventually decided for a non-interviewer mode only redesign based on these results and practical considerations.

6 Discussion

Evaluating total bias, measurement bias, and selection bias, of mixed-mode designs before their introduction or redesign is a problem of great concern (RQ 1 and 2). In the absence of true scores, we suggested using measurements and selection mechanisms as benchmark, which are defined as optimal yielding either single-mode (SMB) or hybrid-mode benchmarks (HMB). It is important to distinguish between single-mode and mixed-mode effects against the SMB and HMB. Evaluating single-mode effects is relevant to assess the need for mixed-mode designs to reduce bias, while mixed-mode effects inform about the success of the procedure. In doing so, selection and measurement effects indicate the sources of the observed total mode effects.

A crucial first step in the evaluation of mode effects is the choice of measurement and selection benchmarks, which lead to either a SMB, if both elements are taken from the same mode, or a HMB, if elements are taken from different modes. In section 3, we provided some discussion how choice of benchmarks was motivated in the present case study. However, in general terms, the choice strongly depends on the particular survey, questionnaires, and properties of the population, and thus may be different in other contexts. In the absence of any further information on mode-specific quality of measurement or selection, a choice of benchmark needs to be based on heuristic arguments, such as the degree of sensitivity of questions (e.g., choose self-administered mode as measurement benchmarks if sensitivity is high) and response rates (e.g., choose modes with high response rates as selection benchmark). In addition, comparing conclusions under

different choices of benchmarks, as in the present study, may give analysts an idea on the importance and implication of benchmark choices.

Given the large number of potential sources of measurement and selection bias, however, heuristic arguments alone may be insufficient. A relevant path for future research is, therefore, the development of methodology for choosing selection and measurement benchmarks. For example, alternative indicators of representativity, such as 'R-indicators' and the fraction of missing information (Schouten, Cobben, and Bethlehem 2009; Wagner 2012), and measurement quality, such as answering behaviors and satisficing (Krosnick 1991) or complex measurement models (Klausch et al. 2013; Revilla 2010; Saris and Gallhofer 2007), may be useful to support benchmark mode choice.

A second step in the evaluation of mode effects is estimation of total effects and their disentanglement into selection and measurement components. In the SMB case, total mode effects may be estimated from simple comparative designs. However, evaluating measurement and selection effects, as well as effects against a HMB, requires estimating potential outcomes (cf. section 4.2). This objective necessitates additional auxiliary data. We suggested using re-interview data for this purpose. Our design is related to other re-interview methods for bias estimation including the basic question approach (Kersten and Bethlehem 1984), the call-back approach (Elliott, Little, and Lewitzky 2000; Hansen and Hurwitz 1946; Keeter et al. 2000), and test-retest designs (Biemer and Lyberg 2003:291), but also features important differences. The basic question and call-back approach involve re-interviews of nonrespondents to estimate nonresponse bias. Test-retest designs normally aim at estimating measurement error. In doing so, measurement independence is often assumed. Our design allows estimating both measurement and selection effects against benchmarks. Furthermore, by modeling potential

outcomes at the first occasion, we explicitly allowed for the possibility that re-interview measurements and selection mechanisms can change between initial interview and follow-up. Change may occur, if follow-up F2F respondents provide different answers than in the benchmark mode (e.g., due to experienced response burden). An alternative explanation is substantial change in statistics across time, but it is often likely to be small in the study period of sequential mixed-mode surveys (two months, in the present study).

An advantage of the design is that it is tailored for use in parallel to sequential mixed-mode surveys and can be implemented even for ongoing surveys without affecting the standard fieldwork (i.e., the benchmark sample is independent and the re-interview does not impact the standard mixed-mode fieldwork). Furthermore, our design does not require the follow-up to be conducted in the measurement and selection benchmark mode. Although in the current case study F2F served as, both, the SMB and follow-up mode, estimation would still be possible with a different follow-up or benchmark mode. For example, a design using telephone as follow-up to web, while F2F remains the SMB, can also be evaluated, if the re-interview is conducted in telephone.

Another advantage of our design is that it allows for a structured view on design decisions in mixed-mode surveys. We demonstrated how single- and mixed-mode designs can be evaluated against SMB and HMB and how a top-down evaluation can be performed (i.e., from total mode effects to its selection and measurement components). Such a top-down approach supports data collection and questionnaire designers. A decisive role is played by the size of the experimental samples and effect sizes that are required to be observable. Future research should evaluate minimum mode-specific sample size requirements, also considering costs of the experimental design.

Our results and design should be judged against a number of limitations which show up paths for further research. Firstly, the size of effects should be evaluated against costs and budgets. For example, F2F is the most expensive mode in data collection, whereas web is inexpensive. The web-F2F mixed-mode design could reduce bias compared to single-mode web (SMB case), but the F2F follow-up response (cf. Table 1) may cause substantial additional costs over single-mode web. Further research, therefore, needs to weigh off mode effects against budget constraints (Vannieuwenhuyze 2014).

Secondly, the re-interview measurements, Y_2^{f2f}, are assumed to be observations from the same response distribution regardless of sample assigned at the first occasion (e.g., F2F or mixed-mode telephone; Schouten et al. 2013) or being respondent or non-respondent. The first part of this assumption was checked by comparing response distributions across assignment conditions (Klausch et al. 2014). We did not find any strong evidence for dependence in the present study[11]. The second part suggests that follow-up respondents in the mixed-mode design provide equivalent Y_2^{f2f} measurements as re-interviewed respondents. In the present study, a relatively long period (6-8 weeks) lay in between first interview and re-interview, which may turn this assumption plausible. However, developing approaches for evaluating this assumption is necessary in the future, especially for use in designs which allow less time between first and second stage of the sequential mixed-mode design.

[11] We conducted chi-square tests of independence between the mode assignment indicator and the polytomous items of the re-interview. For the scaled variables, log-likelihood ratio tests of an empty linear regression model against a model including the mode assignment indicator as predictor were employed. Rao-Scott adjusted tests accounting for sampling weights were applied (Lumley 2010). We found that 3 out of 30 variables showed significant differences across sample assignment under $p<.05$. This number of significant tests may occur by chance due to multiple testing. After Bonferroni adjustment for multiple testing these differences disappeared. Furthermore, empirical differences in distributions were very small.

Thirdly, our results may depend on the specific estimation procedure (multiple imputation) and related modeling decisions. For example, we included Y_2^{f2f} as predictors of potential outcomes and it might be an option to extend the models to observed information at the first interview (e.g., Y^{tel}). We started to evaluate this possibility but including such predictors led to convergence problems of the Gibbs sampler in *MICE*. We suspect this may be related to either the fact that the partial correlation between response distributions at the first occasion is not observed or to the discrete level of measurement of many variables leading to highly parameterized models. Future research could, therefore, evaluate more general imputation and modeling strategies for the missing data pattern of the design for both continuous and discrete data and the use of alternative imputation techniques (e.g., predictive mean matching; van Buuren 2012:68–74).

The evaluation of bias in mixed-mode surveys will continue to be of concern for methodologists and practitioners. In this respect, further development of our method is desirable. If the strong prevalence of measurement effects should prove a problem in many mixed-mode surveys, developing adjustment methodology for measurement effects is necessary. Such methodology could use Bayesian approaches that yield multiple imputations of potential outcomes in mixed-mode surveys. Developing and evaluating these methods appears urgent in face of our empirical results.

Bibliography

Bethlehem, Jelke and Silvia Biffignandi. 2011. *Handbook of Web Surveys*. John Wiley & Sons, Inc.

Bethlehem, Jelke, Fannie Cobben, and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys.* New Jersey: Wiley.

Biemer, Paul P. 1988. "Measuring Data Quality." Pp. 341–75 in *Telephone Survey Methodology*, edited by Robert M. Groves et al. New York: Wiley.

Biemer, Paul P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74(5):817–48.

Biemer, Paul P. and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. New Jersey: Wiley.

Van Buuren, Stef. 2012. *Flexible Imputation of Missing Data*. Boca Raton: CRC Press.

Van Buuren, Stef, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and Don B. Rubin. 2006. "Fully Conditional Specification in Multivariate Imputation." *Journal of Statistical Computation and Simulation* 76(12):1049–64.

Van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45(3):1–67.

Couper, Mick P. 2000. "Review: Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64(4):464–94.

Couper, MICK P., MICHAEL W. Traugott, and MARK J. Lamias. 2001. "Web Survey Design and Administration." *Public Opinion Quarterly* 65(2):230–53.

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2009. *Internet, Mail, and Mixed-Mode Surveys. The Tailored Design Method*. New Jersey: Wiles & Sons.

Elliott, Michael R., Roderick J. A. Little, and Steve Lewitzky. 2000. "Subsampling Callbacks to Improve Survey Efficiency." *Journal of the American Statistical Association* 95(451):730–38.

Fowler, Floyd Jackson et al. 2002. "Using Telephone Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members." *Medical Care* 40(3):190–200.

Fricker, Scott, Mirta Galesic, Roger Tourangeau, and Ting Yan. 2005. "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly* 69(3):370–92.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5):646–75.

Groves, Robert M. et al. 2010. *Survey Methodology*. 2nd ed. New Jersey: Wiley.

Groves, Robert M. and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5):849–79.

Hansen, Morris H. and William N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association* 41(236):517–29.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86(1):4–29.

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 64(2):125–48.

Kersten, Hubert M. P. and Jelke Bethlehem. 1984. "Exploring and Reducing the Nonresponse Bias by Asking the Basic Question." *Statistical Journal of the United Nations Economic Commission for Europe* 2(4):369–80.

Kirchner, Antje and Barbara Felderer. 2013. "The Effect of Survey Mode on Nonresponse Bias and Measurement Error: A Validation Approach."

Klausch, Thomas, Joop J. Hox, and Barry Schouten. 2013. "Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions." *Sociological Methods & Research* 42(3):227–63.

Klausch, Thomas, Joop J. Hox, and Barry Schouten. 2016. "Selection Error in Single- and Mixed-Mode Surveys of the Dutch General Population." *Forthcoming in Journal of the Royal Statistical Society: Series A*.

Klausch, Thomas, Barry Schouten, and Joop J. Hox. 2014. *The Use of Within-Subject Experiments for Estimating Measurement Effects in Mixed-Mode Surveys*. The Hague, The Netherlands: Statistics Netherlands. Retrieved January 4, 2014 (http://www.cbs.nl/NR/rdonlyres/181793AC-94B8-4748-9C2B-E541DCF9CFB7/0/201406x10pub.pdf).

Körmendi, E. 1988. "The Quality of Income Information in Telephone and Face to Face Surveys." Pp. 341–75 in *Telephone Survey Methodology*, edited by Robert M. Groves et al. New York: Wiley.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74(5):880–906.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann. 2013. "A Note on Mechanisms Leading to Lower Data Quality of Late or Reluctant Respondents." *Sociological Methods & Research*. Retrieved April 8, 2014 (http://smr.sagepub.com.proxy.library.uu.nl/content/early/2013/10/28/0049124113508094).

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity." *Public Opinion Quarterly* 72(5):847–65.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3):213–36.

De Leeuw, Edith. 1992. *Data Quality in Mail, Telephone, and Face to Face Surveys*. Amsterdam: TT-Publicaties.

De Leeuw, Edith. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21(2):233–55.

De Leeuw, Edith. 2008. "Choosing the Method of Data Collection." Pp. 113–35 in *International Handbook of Survey Methodology*, edited by Edith De Leeuw, Joop J. Hox, and Don A. Dillman. New York: Taylor & Francis.

De Leeuw, Edith, Don A. Dillman, and Joop J. Hox. 2008. "Mixed Mode Surveys: When and Why." Pp. 299–316 in *International Handbook of Survey Methodology*, edited by Edith De Leeuw, Don A. Dillman, and Joop J. Hox. New York: Lawrence Erlbaum.

Link, Michael W. and Ali H. Mokdad. 2006. "Can Web and Mail Survey Modes Improve Participation in an RDD-Based National Health Surveillance?" *Journal of Official Statistics* 22(2):293–312.

Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken: Wiley.

Lumley, Thomas S. 2010. *Complex Surveys: A Guide to Analysis Using R*. Hoboken, New Jersey: Wiley.

Lynn, Peter. 2013. "Alternative Sequential Mixed-Mode Designs: Effects on Attrition Rates, Attrition Bias, and Costs." *Journal of Survey Statistics and Methodology* 1(2):183–205.

Millar, Morgan M. and Don A. Dillman. 2011. "Improving Response to Web and Mixed-Mode Surveys." *Public Opinion Quarterly* 75(2):249–69.

Miller, Peter V. and Robert M. Groves. 1985. "Matching Survey Responses to Official Records: An Exploration of Validity in Victimization Reporting." *Public Opinion Quarterly* 49(3):366–80.

Morgan, Stephen L. and Chirstopher Winship. 2007. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Olson, Kristen. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias." *Public Opinion Quarterly* 70(5):737–58.

Pearl, Judea. 2009. *Causality. Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press.

Peytchev, Andy. 2012. "Multiple Imputation for Unit Nonresponse and Measurement Error." *Public Opinion Quarterly* 76(2):214–37.

Raghunathan, Trivellore E., James Lepkowski, John van Hoewyk, and Peter Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27(1):85–95.

Revilla, Melanie. 2010. "Quality in Unimode and Mixed-Mode Designs: A Multitrait-Multimethod Approach." *Survey Research Methods* 4(3):151–64.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6(1):34–58.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, Donald B. 2003. "Nested Multiple Imputation of NMES via Partially Incompatible MCMC." *Statistica Neerlandica* 57(1):3–18.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes." *Journal of the American Statistical Association* 100(469):322–31.

Sakshaug, Joseph W., Ting Yan, and Roger Tourangeau. 2010. "Nonresponse Error, Measurement Error, And Mode Of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-Sensitive Items." *Public Opinion Quarterly* 74(5):907–33.

Saris, Willem E. and Irmtraud Gallhofer. 2007. "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions." *Survey Research Methods* 1(1):29–43.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall/CRC.

Schafer, Joseph L. and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7(2):147–77.

Schafer, Joseph L. and Joseph Kang. 2008. "Average Causal Effects from Nonrandomized Studies: A Practical Guide and Simulated Example." *Psychological Methods* 13(4):279–313.

Schouten, Barry, Jan van den Brakel, Bart Buelens, Jan van der Laan, and Thomas Klausch. 2013. "Disentangling Mode-Specific Selection and Measurement Bias in Social Surveys." *Social Science Research* 42(6):1555–70.

Schouten, Barry, Fannie Cobben, and Jelke Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35(1):101–13.

Tourangeau, Roger, Robert M. Groves, and Cleo D. Redline. 2010. “Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error.” *Public Opinion Quarterly* 74(3):413–32.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tourangeau, Roger and Ting Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133(5):859–83.

Vannieuwenhuyze, Jorre. 2014. “On the Relative Advantage of Mixed-Mode versus Single-Mode Surveys.” *Survey Research Methods* 8(1):31–42.

Vannieuwenhuyze, Jorre and Geert Loosveldt. 2013. “Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects.” *Sociological Methods & Research* 42(1):82–104.

Vannieuwenhuyze, Jorre, Geert Loosveldt, and Geert Molenberghs. 2010. “A Method for Evaluating Mode Effects in Mixed-Mode Surveys.” *Public Opinion Quarterly* 74(5):1027–45.

Vannieuwenhuyze, Jorre, Geert Loosveldt, and Geert Molenberghs. 2014. “Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models.” *Journal of Official Statistics* 30(1):1–21.

Voogt, Robert J. J. and Willem E. Saris. 2005. “Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects.” *Journal of Official Statistics* 21(3):367–87.

Wagner, James. 2012. “A Comparison of Alternative Indicators for the Risk of Nonresponse Bias.” *Public Opinion Quarterly* 76(3):555–75.

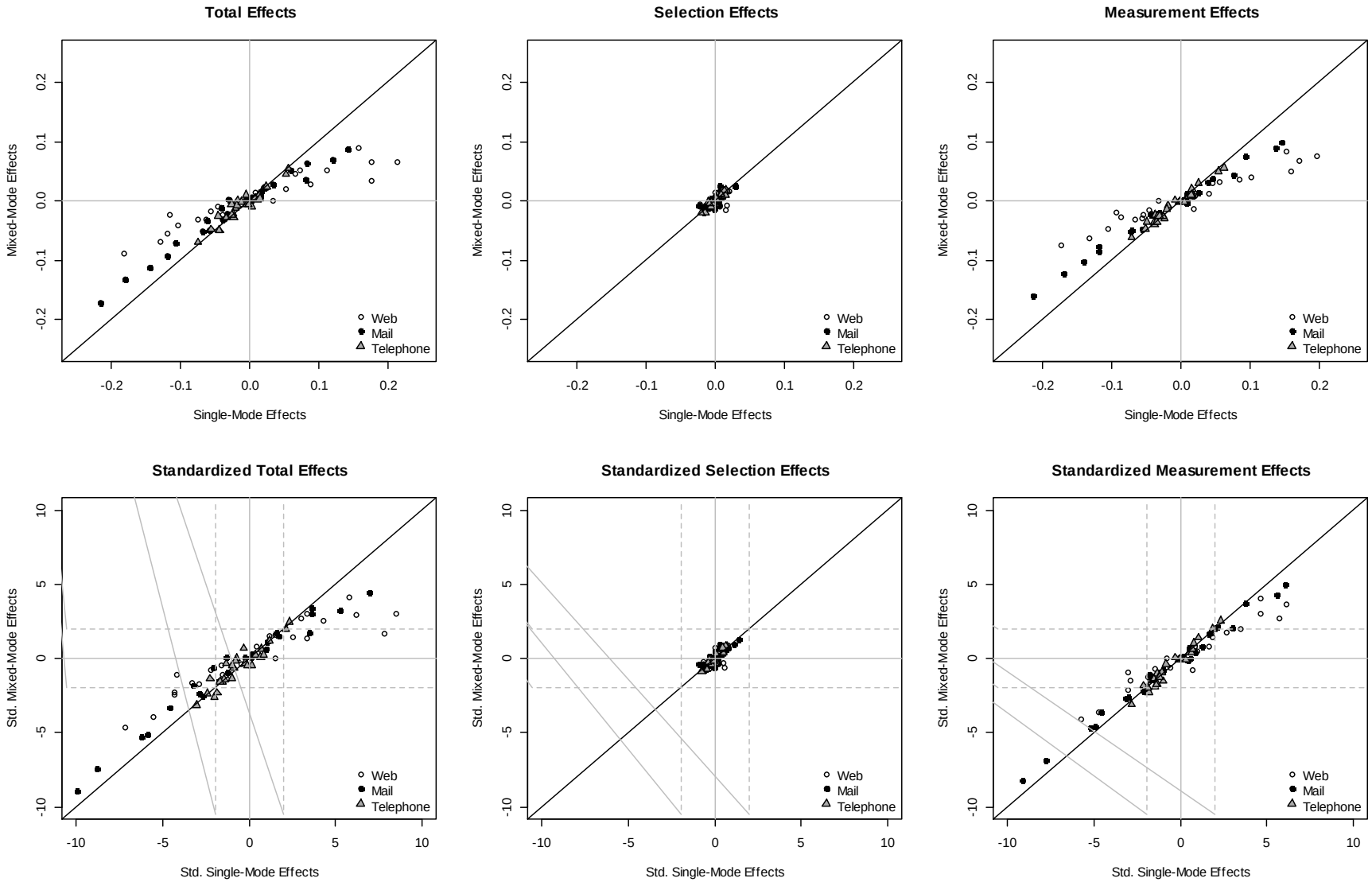

Figure 3: Scatterplots of single-mode against mixed-mode effects for the SMB case (upper row: unstandardized effects; lower row: standardized effects, where dashed lines indicate critical values [p<.05])

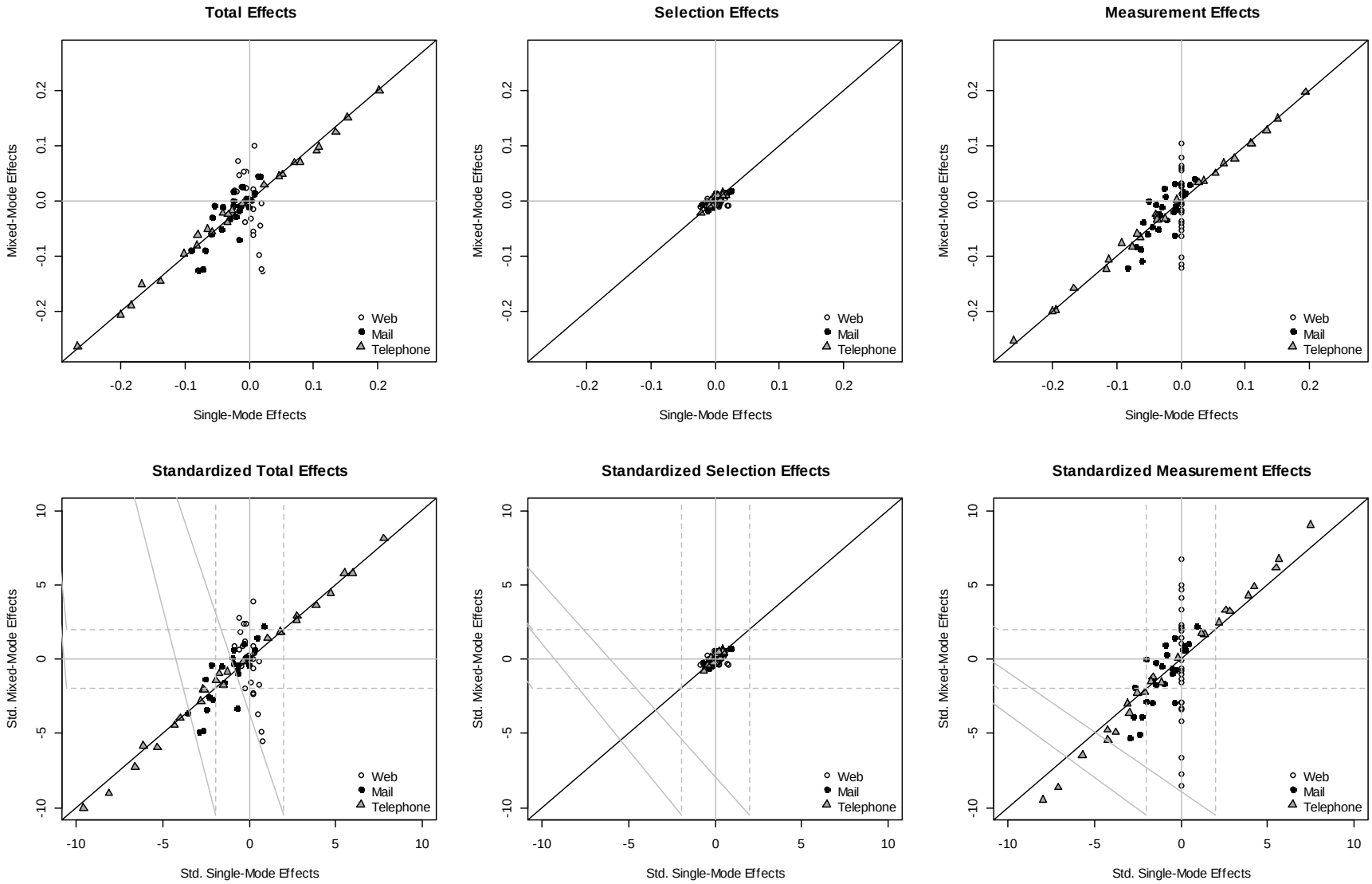

Figure 4: Scatterplots of single-mode against mixed-mode effects for the HMB case (upper row: unstandardized effects; lower row: standardized effects, where dashed lines indicate critical values [p<.05])

Appendix A

Table A1: Overview on CVS items with benchmark statistics (item nonresponse imputed) with proportion of item nonresponse by wave 1 modes

	Benchmark Statistic		% item nonresponse (DK / refusal)			
	SMB (F2F)	Hybrid (F2F-Web)	F2F (n=1,048)	Telephone (n=735)	Mail (n=857)	Web (n=497)
'Social quality neighbourhood'[a]*:*						
State of roads, walkways, squares	70.4	65.9	0.0	0.0	4.3	0.6
Good playgrounds for children	61.6	57.3	4.8	3.1	8.8	4.6
Good provisions for the young	25.5	20.4	9.3	7.2	15.2	9.9
People know each other well	23.7	23.4	0.6	0.4	5.6	1.4
People treat each other well	83.7	71.3	0.5	0.1	5.1	0.6
Nice neighbourhood with solidarity	53.6	43.9	1.0	0.1	5.8	0.6
Feel at home with people	78.5	59.6	0.0	0.1	4.7	1.0
Have a lot of contact with people	48.7	38.6	0.0	0.0	4.2	0.4
Satisfied with population compos.	81.7	72.3	0.3	0.1	4.9	0.8
'Neighbourhood problems'[b]*:*						
Plastering on walls and/or buildings	29.2	37.4	0.3	0.4	8.5	5.2
Harassment by groups of young	34.8	50.0	0.4	0.3	4.6	2.4
Drunken people on the streets	27.9	37.5	0.7	0.7	7.4	6.2
Unpleasant people on the streets	11.2	17.2	1.5	1.5	15.2	12.5
Junk on the streets	52.4	71.8	0.3	0.1	2.9	0.4
Dog excrements on the streets	65.6	81.4	0.1	0.3	3.2	1.0
Destruction of telephone cells, etc.	33.4	49.6	4.5	4.2	17.9	14.9
Drug problems	17.5	20.7	2.0	1.4	21.2	16.9
Summary ratings neighbourhood:						
Insecurity feeling in general[c]	21.6	27.1	0.0	0.1	4.6	2.8
Frequency insecurity feeling[b]	17.1	18.0	0.0	0.0	2.9	0.0
Insecurity feeling in neighbourhood[c]	11.5	17.5	0.0	0.1	4.1	2.2
Frequency insecurity feeling[b]	9.8	12.0	0.0	0.0	0.7	0.0
Quality of Life Rating[d]	7.54	7.47	0.3	0.5	3.9	1.8
Police evaluation						
Police contact (past 12 mth.)[c]	35.1	30.8	0.1	0.0	5.3	2.2
Evaluation Police Contact[e]	23.1	17.4	0.0	0.5	1.9	2.1
Evaluation Police Performance[e]	48.5	45.9	7.6	3.9	20.5	11.9
Victimization (past 12 mth.)[f]						
Mean no. of victimizations	0.158	0.206	0.0	0.0	0.0	0.0
% victim of crime	10.9	14.2	0.0	0.0	0.0	0.0
Summary Indices (scales)[g]						
Social cohesion	6.67	6.33	1.7	0.8	7.9	2.2
Neighbourhood Nuisances	1.44	1.69	4.0	2.9	27.3	23.7
Neighbourhood Deterioration	3.21	4.04	4.9	4.6	21.9	17.7

a. Five rating scale answer categories from 'Completely disagree' to 'Completely Agree'. Target statistic: % (completely) agree
b. Three rating scale answer categories: 'Happens Frequently', 'Sometimes', and 'Rarely or never'. Target statistic: % frequently / sometimes
c. Dichotomous (yes / no). Target statistics: % yes
d. Question 'Do you sometimes feel insecure?' with two answer categories: yes, no
e. Rating of satisfaction from 1 (very satisfied) to 5 (very unsatisfied). Target statistic: % (very) satisfied
f. Victimization variables are aggregated based on multiple questions about victimization in the past year (count / dichotomous). Target statistics: mean no. of victimizations in past 12 months; % victimized in past 12 months
g. Index based on multiple items from the social quality block ('social cohesion') and neighbourhood problems block ('neighbourhood nuisances'; 'neighbourhood deterioration'). Scaled to a range from 1 (very low) to 10 (very high). Target statistic: index mean