Statistics Netherlands

Discussion Paper

The use of within-subject experiments for estimating measurement effects in mixed-mode surveys

2014 | 06

Thomas Klausch
Barry Schouten
Joop Hox
12-03-2014

Explanation of symbols

.	Data not available
*	Provisional figure
**	Revised provisional figure (but not definite)
x	Publication prohibited (confidential figure)
–	Nil
–	(Between two figures) inclusive
0 (0.0)	Less than half of unit concerned
empty cell	Not applicable
2013–2014	2013 to 2014 inclusive
2013/2014	Average for 2013 to 2014 inclusive
2013/'14	Crop year, financial year, school year, etc., beginning in 2013 and ending in 2014
2011/'12–2013/'14	Crop year, financial year, etc., 2011/'12 to 2013/'14 inclusive

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

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

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

Where to order
verkoop@cbs.nl
Fax +31 45 570 62 68

60083 201406 X-10

The use of within-subject experiments for estimating measurement effects in mixed-mode surveys

Thomas Klausch, Barry Schouten and Joop Hox

Summary: The estimation of measurement effects (*MEs*) of survey modes in the presence of selection bias poses a great problem to methodologists. In practice, unit nonresponse is often assumed to be ignorable (MAR) conditional on socio-demographic auxiliary data for lack of other sample-level information. Since socio-demographics commonly are only weakly related to target variables or selection mechanisms this assumption often may not hold true.

We present a new method to estimate *MEs* by means of "within-subject designs", in which the same sample is approached by two different modes at two subsequent points in time. This design allows ignoring nonresponse on repeatedly measured target variables implying weaker MAR assumptions, because repeated target variables typically are strongly related. Further assumptions of this design are discussed in detail, in particular (a) time-stability of target variables and response probabilities and (b) the independence of measurement occasions. In extensions of simple within-subject designs, an independent control group, to which the same mode is assigned at both occasions, is useful to test these assumptions and adjust for time-instability.

The decomposition of mode effects into *MEs* and selection biases is illustrated for key statistics from the Dutch Crime Victimization Survey using data from a large-scale within-subject experiment conducted within Statistics Netherlands' project Mode Effects in Social Surveys (abbreviated to MEPS in Dutch).

The method presented in this paper shows similarity to the mode effect decomposition method proposed by Buelens et al (2012) within project MEPS. We will discuss differences in assumptions and estimators.

Keywords: Missing Data; Nonresponse Adjustment; Selection Bias; Causal Inference; Longitudinal Surveys; Mode Effects; Measurement Effects; Mixed-Mode Surveys;

1. Introduction

It is a well-known problem of mixed-mode research that measurement error of questions can differ across modes threatening the accuracy and comparability of data in mixed-mode surveys (Tourangeau, Rips & Rasinski, 2000; Krosnick, 1991, 1999; Dillman et al., 2009). One option to deal with this problem is to minimize any differences in measurement error between modes before they occur, for example by means of 'unified mode questionnaires' (de Leeuw, 2005; Dillman & Christian, 2005; Dillman, Smyth, & Christian, 2009, p. 326). In designing such questionnaires, estimating the average difference in measurement error between two modes, called the average 'measurement effect' of a given question (*ME*), is an essential prerequisite.
In the present paper, we suggest a new method for unbiased estimation of *MEs*. It is widely recognized that this objective is not trivial, because selection bias (i.e., bias caused by mode-specific unit nonresponse) represents an alternative explanation for *MEs* in mixed-mode surveys (Vannieuwenhuyze, Molenberghs, & Loosveldt, 2010; Jäckle, Roberts, & Lynn, 2010; Vannieuwenhuyze & Loosveldt, 2013). Available techniques to estimate *MEs* mainly rely on covariate-based adjustment, such as calibration weighting, regression estimation or matching (Morgan & Winship, 2007; Schafer & Kang, 2008), where, in practice, socio-demographics are the most common type of covariates that is available (Tourangeau & Smith, 1996; Schonlau et al., 2004; Jäckle, Roberts, & Lynn, 2010; Lugtig et al., 2011; Vannieuwenhuyze & Loosveldt, 2013). However, since socio-demographics are often only weakly related to response mechanisms or target variables, it is possible that many *ME* estimates are still biased after adjustment for selection bias. Plausible adjustment therefore is the key challenge of *ME* estimation today.

Finding stronger adjustment covariates is one potential solution to the problem. Buelens et al (2012) and Schouten et al (2013) proposed a 'between-subject' design where the full sample received a follow-up using face-to-face. The follow-up is conducted to create strongly related adjustment covariates by repeating part of the questions. Buelens et al (2012) and Schouten et al (2013) propose to estimate mode effect components by calibrating response to the follow-up wave response on these repeated variables. We call this a between-subject design with follow-up.

In this paper, we propose an alternative design based on a similar data collection; we address the deficiency of adjustment covariates by using 'within-subject' designs with repeated measures (Winship & Morgan, 1999). In these designs, the same sample is again approached at two subsequent points in time by two different modes while posing relevant target variables repeatedly. However, contrary to the between-subject design with follow-up we view the answers at the two time points as repeated measurements and derive measurement effect estimates directly from them.

We note that we use the term 'design' to indicate the combination of data collection design and estimators. Although the within-subject design and between-subject design with follow-up are similar in their data collection design, they employ different estimators. In this paper, we will discuss the resulting differences in assumptions that are made by the two designs. We will, however, not compare differences in actual estimates, but reserve this for a future paper.

We also note that the focus of this paper is on measurement effects. For this reason, we do not distinguish between coverage effects and nonresponse effects, as is done in Buelens et al (2012) and Schouten et al (2013), but consider only the compound of the two effects as selection effects.

The within-subject data collection is fielded independently of any ongoing mixed-mode survey. Its sole purpose is the estimation of particular types of *MEs*, which, as we demonstrate in section 2, enable survey practitioners to design unified mode questionnaires in two common mixed-mode scenarios: (a) single-to-single mode switches during longitudinal surveys and (b) introduction of sequential mixed-mode surveys. In the within-subject experiment, a selection of candidate modes (i.e., modes considered for a mixed mode design) is administered to all relevant population domains before design decisions are taken. This aspect can be considered a strong feature of the method. Earlier approaches often suppose a specific mixed-mode survey is readily available and *MEs* are to be estimated from this data only (Vanniewenhuyze & Loosveldt, 2013). However, in this case, alternative modes can hardly be considered, whereas a separate experiment enables researchers to evaluate different mixed-mode options before taking design decisions.

This paper is structured as follows. We start by defining different types of *MEs* and explain how they can be applied for unimode questionnaire design. Next, we review the available approaches to estimation of measurement effects. Subsequently, we explain how to estimate *MEs* in within-subject designs and offer tests of the method's basic assumptions. Finally, we illustrate the method using a series of practical examples from a large-scale within-subject data collection within the Dutch Crime Victimization Survey.

2. Definition and practical application of MEs

Let Y_t^m denote a continuous or discrete random variable for the outcome on a given question posed under mode *M* at occasion *t* (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010). The distribution of differences in outcomes between two modes, A and B, is denoted by $P(Y_1^b - Y_1^a)$, if *Y* is continuous, and $P(Y_1^b, Y_1^a)$, if *Y* is discrete. These can be considered the distributions of 'individual-level' *MEs*. The average 'marginal *ME*' is then defined as expected value of the continuous distribution of individual-level *MEs*

$$ME = E(Y_1^b) - E(Y_1^a). \quad (1)$$

If *Y* is discrete, (1) denotes the deviation from marginal homogeneity of a category *y* in the contingency table of Y_1^b and Y_1^a.

The deviations of the mean of outcomes $E(Y_t^m)$ from their 'true mean' can be defined as 'measurement error'. Therefore, an *ME* of size zero indicates that a question posed under different modes evokes the same extent of measurement error. This idea will become central when applying *ME* estimates for unified mode design.

First, however, we introduce the situation when a survey is administered using a particular mode. In this situation, the fieldwork differences between modes evoke specific 'response mechanisms', denoted by binary response variable S_t^m, where '1' indicates response and '0' nonresponse. S_t^m can subsume the outcome of all possible reasons for mode-specific non-observation against the full population, such as refusal, non-contact or non-coverage (Groves, 1989; Groves et al., 2010). In principle, measurement error might differ between respondents with higher or lower 'response probability' $P(S_t^m)$ (Fricker & Tourangeau, 2010; Kaminska, McCutcheon, & Billiet, 2010). Therefore, it makes sense to consider *MEs* also conditional on S_t^m, which is called the average 'conditional *ME*'. For two modes, there are two conditional *MEs* for respondents:

$$ME_R^a = E(Y_1^b \mid S_1^a = 1) - E(Y_1^a \mid S_1^a = 1) \quad (2)$$

$$ME_R^b = E(Y_1^b \mid S_1^b = 1) - E(Y_1^a \mid S_1^b = 1) \quad (3)$$

and two conditional *MEs* for nonrespondents, not considered here. For example, ME_R^a indicates the expected difference in measurement error (answers) between modes A and B that can be expected from respondents in mode A. Questionnaire designers often try to minimize absolute measurement error of a survey design against the population value. Since any mode may measure *Y* with error, the presence of a *ME* indicates in most situations that one of the modes measures *Y* with more error than

the other. Furthermore, if it is known which mode evokes less measurement error on *Y*, questionnaires can be optimized towards this 'benchmark mode'. One way to do so is by so-called unified mode questionnaires, which are designed to evoke equal measurement error from two given modes (de Leeuw, 2005; Dillman & Christian, 2005; Dillman, Smyth, & Christian, 2009, p. 326). An optimal unified mode questionnaire minimizes *MEs* and also maintains measurement error of the mode that is known to evoke less error.

For designing such questionnaires *ME* estimates are needed. However, since different mixed-mode designs evoke different response mechanisms, the type of *ME* that has to be estimated differs, too. As we discuss in the following, the two conditional *MEs* (2) and (3) can be used for designing unified mode questionnaires in two different types of mixed-mode situations (as discussed in De Leeuw, 2005; Dillman, Smyth, & Christian, 2009, p. 306-310; Groves et al., 2010, p. 175-177).

First, consider the scenario when a survey in mode A is exchanged by a survey in mode B. There are various situations in practice, when this 'single-to-single mode switch' may be required. For example, in repeated cross-sectional surveys using mode A, cost constraints might necessitate switching the survey completely to (cheaper) mode B. In this scenario, measurement error of mode A often needs to be preserved to ensure comparability in time or because it is known to have smaller error. Another situation is represented by mixed-mode panel surveys that apply different modes in different survey waves (e.g., mode A for recruiting respondents and mode B for re-interviewing).

Table 1: Hypotheses about (double-) conditional MEs to find unified mode designs

	Measurement Error desired from (Benchmark Mode)		
Design Scenario	Mode A	Mode B	Unified design only (i.e., A or B)
1. Single: A switched to B	$ME_R^b = 0$	(always provided)	$ME_R^b = 0$
2a. Sequential: B followed by A	$ME_R^b = 0$	$ME_{NR,R}^{b,a} = 0$	$ME_R^b = 0 \wedge ME_{NR,R}^{b,a} = 0$
2b. Sequential: A followed by B	$ME_{NR,R}^{a,b} = 0$	$ME_R^a = 0$	$ME_R^a = 0 \wedge ME_{NR,R}^{a,b} = 0$

Second, consider the scenarios when a survey is conducted in mode A (or B) and the nonrespondents are followed up by alternative mode B (or A). This design is known as a 'sequential mixed-mode' survey. Minimizing measurement error is a problem of considerable concern in this design, too, because the presence of *MEs* implies that one of the modes increases total error of the survey.

Now, three objectives may apply: mode A is known to evoke least error and is taken as the benchmark mode, mode B is known to evoke least error and is taken as the benchmark mode, or it is unknown what mode produces least error. The three scenarios and the three objectives lead to nine situations (Table 1). We discuss each of them.

When it is the objective to nullify *MEs*, zero-constraint hypotheses on different *MEs* are assessed. First, consider the single-to-single mode switch (scenario 1). Suppose it is the goal to optimize measurement error at the level of mode A. Then the relevant test is represented by $H_0 : ME_R^b = 0$ indicating that

$$E(Y_1^b \mid S_1^b = 1) = E(Y_1^a \mid S_1^b = 1),$$

so that measurement error of respondents in the 'new' mode B is equal to error provided under mode A (first row, second column). If $ME_R^b \neq 0$, the question's wording and format need to be reconsidered in order to find a better unified mode design, where, obviously, this might not always be possible in practice. Column three lists the second objective, when measurement error at the level of mode B is the benchmark. In this case, the 'new' survey using mode B obviously evokes measurement error of mode B, so that separate testing is not needed. The test in column four ('unified design only') considers the third objective. In some practical cases, the researcher cannot be certain about the size of measurement error in different modes. In this situation one has to act as if either mode A or mode B evoked least error. This situation thus requires testing $ME_R^b = 0$ as well.

We consider the two sequential mixed-mode designs (scenarios 2a and 2b). Consider first design 2a. The mean of the outcomes provided by respondents at the first step of the sequential design $E(Y_1^b \mid S_1^b = 1)$ exhibits measurement error of mode B. If measurement error of mode A is desired for the mixed-mode sample (second column, table 1), it needs to be evaluated whether $ME_R^b = 0$, likewise argued for design 1. This is sufficient, because respondents at the second step of the sequential design already provide measurement error at the level of mode A. If measurement error of mode B is desired (third column), the respondents at the second occasion provide answers with measurement error of mode A which might differ from mode B. The mean outcome of this group is $E(Y_2^a \mid S_1^b = 0, S_2^a = 1)$, where Y_2^a and S_2^a now reflect that outcomes and response mechanism are observed at a later point in time (occasion two). Since measurement error of $E(Y_2^a \mid S_1^b = 0, S_2^a = 1)$ should not differ from mode B, we require

$$ME_{NR,R}^{b,a} = E(Y_2^b \mid S_1^b = 0, S_2^a = 1) - E(Y_2^a \mid S_1^b = 0, S_2^a = 1) = 0. \qquad (4)$$

$ME^{b,a}_{NR,R}$ is called a 'double-conditional *ME*', because it conditions on two response mechanisms. It is defined for a sequential mixed-mode design, which follows up nonrespondents in mode B by mode A. The hypotheses for design 2b follow the same logic as for 2a and can be taken from table 1 (last row). However, now the double conditional *ME* is defined as:

$$ME^{a,b}_{NR,R} = E(Y_2^b \mid S_2^b = 1, S_1^a = 0) - E(Y_2^a \mid S_2^b = 1, S_1^a = 0) \quad (5)$$

for a sequential design following up mode A by mode B at occasion two. Under the third objective, both the conditional and double-conditional effects need to be assessed simultaneously to guarantee a unified design.

In summary, it is important to realize that marginal *MEs* are irrelevant in all scenarios. Estimation can, therefore, focus only on (double-) conditional instead of marginal *MEs*. The next section discusses the available approaches and assumptions needed for estimating the defined quantities before section 4 presents the new method.

3. Estimating measurement effects in between subject designs

The most common approach to estimating conditional and marginal *MEs* are experiments using so-called between-subject designs (BSD), for which separate independent samples are drawn and assigned to different modes (e.g., Aquilino, 1994; Tourangeau & Smith, 1996; Schonlau et al., 2004; Fricker et al., 2005; Kreuter, Presser, & Tourangeau, 2008; Heerwegh & Loosveldt, 2008; Chang & Krosnick, 2009; Dillman et al., 2009; Heerwegh, 2009; Jäckle, Roberts, & Lynn, 2010).

The key problem in BSD is the presence of missing data, of which we can distinguish two different types illustrated in figure 1. Depicted are the variables Y_1^a, Y_1^b introduced in section 2 and auxiliary information *X*. The black areas represent the observed part of the data (i.e., response, $S_1^a = 1$ and $S_1^b = 1$) and the grey areas missing data due to unit nonresponse ($S_1^a = 0$ and $S_1^b = 0$), the first type of missing data. Furthermore, the moment in time Y_t^m has been observed under a given mode all outcomes in other modes are considered 'potential' (Rubin, 1974, 1976, 1977, 2005; Holland, 1986). Potential outcomes can never be observed in reality, because it is not possible to observe outcomes under two modes at the same point in time. Therefore, the potential outcomes represent the second type of missing data depicted by white color. Finally, the auxiliary information *X* is supposed to be exogenous (i.e., unaffected by any *MEs*; Imbens, 2004) and available for all units.

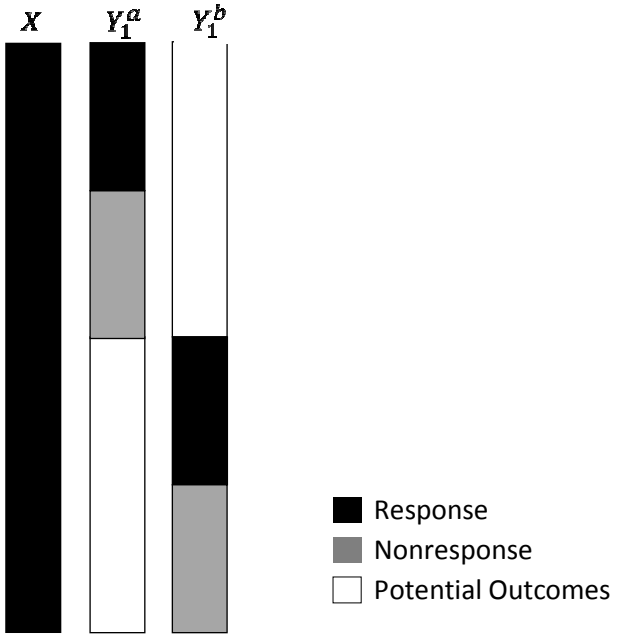

Figure 1: Missing Data Pattern of a Between-Subject Design

In sections 3 and 4, we introduce various assumptions. To distinguish them from other expressions, we label them using a prescript "A". So the first assumption is labeled A1 .

3.1 The naïve unadjusted ME estimator in between-subject designs

A first approach to estimate marginal or conditional *MEs* is represented by the simple difference in means between response samples. The expected value of this 'naïve' estimator, i.e.:

$$E(\hat{ME}_R^{naive}) = E(Y_1^b \mid S_1^b = 1) - E(Y_1^a \mid S_1^a = 1) \tag{6}$$

confounds measurement error with non-observation error. The naïve estimator is then said to suffer from selection bias. Using $\hat{ME}_R^{naive}$ as an estimator for ME_R^b, the selection bias is represented by the difference in non-observation errors on Y_1^a:

$$SE(Y_1^a) = E(Y_1^a \mid S_1^b = 1) - E(Y_1^a \mid S_1^a = 1) \tag{7}$$

and when using $\hat{ME}_R^{naive}$ as an estimator for ME_R^a it is the difference in non-observation error on Y_1^b:

$$SE(Y_1^b) = E(Y_1^b \mid S_1^b = 1) - E(Y_1^b \mid S_1^a = 1). \tag{8}$$

The selection bias of $\hat{ME}_R^{naive}$ against the marginal *ME* is given by the difference in non-observation errors on Y_1^b and Y_1^a:

$$SE = (E(Y_1^b \mid S_1^b = 1) - E(Y_1^b)) - (E(Y_1^a \mid S_1^a = 1) - E(Y_1^a)). \tag{9}$$

In all cases, $\hat{ME}_R^{naive}$ is an estimator for a 'net effect', confounding selection bias with *MEs*, e.g. for conditional *MEs*:

$$E(\hat{ME}_R^{naive}) = SE(Y_1^a) + ME_R^b = SE(Y_1^b) + ME_R^a. \tag{10}$$

Obviously, any technique leading to unbiased estimates of conditional or marginal *MEs* 'disentangles' *MEs* from their selection biases, also referred to as 'selection effects' in this context (Vannieuwenhuyze & Loosveldt, 2013). The available approaches to unbiased estimation and their assumptions are discussed next.

3.2 Missing data adjustment in between-subject designs

Given the missing data problem illustrated in figure 1, missing data adjustment is the primary approach to unbiased *ME* estimation. Various techniques have been discussed in the context of the causal inference literature focusing on the estimation of potential outcomes (Imbens, 2004; Rubin, 2005; Kang & Schafer, 2007; Schafer & Kang, 2008)

and nonresponse in sample surveys focusing on the adjustment of non-observation error (Särndal & Lundström, 2005). Techniques include calibration weighting, (robust) regression estimation (Cochran, 1977; Bethlehem, 1988, 2002; Särndal & Lundström, 2005; Kang & Schafer, 2007), matching (Rosenbaum, 2002), and multiple imputation (Rubin, 1988; Schafer, 1997; van Buuren, 2012). When estimating marginal *MEs*, all of these techniques are based on the assumption that nonresponse in both samples is missing at random (MAR) given auxiliary information *X* implying (Rubin, 1976; Little & Rubin, 2002):

$$Y_1^m \perp S_1^m \mid X \text{ for } m = \{a,b\} \qquad \text{(A1)}$$

where $\perp$ denotes independence of Y_1^m and S_1^m (here condtional on *X*). To the contrary, when estimating conditional *MEs*, the conditional means of potential outcomes, i.e. $E(Y_1^b \mid S_1^a = 1)$ and $E(Y_1^a \mid S_1^b = 1)$, are needed. It is never possible to observe these quantities in reality, but they can be estimated assuming MAR as:

$$Y_1^b \perp S_1^a \mid X, S_1^b = 1 \text{ to estimate } E(Y_1^b \mid S_1^a = 1) \qquad \text{(A2a)}$$

$$Y_1^a \perp S_1^b \mid X, S_1^a = 1 \text{ to estimate } E(Y_1^a \mid S_1^b = 1). \qquad \text{(A2b)}$$

The assumptions are also known as 'unconfoundedness' in the causal inference literature[1] (Imbens, 2004; Kang & Schafer, 2007; Schafer & Kang, 2008). A2a/b imply that *X* explains the distributional differences between response samples on Y_1^a and Y_1^b. For example, assuming (A2a) one may predict the potential outcomes $E(Y_1^b \mid X, S_1^a = 1)$ from a regression model fitted on the observed data $E(Y_1^b \mid X, S_1^b = 1)$. The mean of potential outcomes over *X* then serves as an unbiased estimate of $E(Y_1^b \mid S_1^a = 1)$.

Assumptions (A1) and (A2a/b) are more likely to hold in practice, when *X* strongly relates to the response mechanism and Y_1^a and Y_1^b. Usually socio-demographic variables are available as auxiliary data (*X*) in survey research (e.g., sampling frame information). Socio-demographics have been applied as 'control covariates' in regression models (Tourangeau & Smith, 1996; Heerwegh & Loosveldt, 2008), as weighting variables (Holbrook, Green, & Krosnick, 2003; Schonlau et al., 2004; Fricker et al., 2005; Chang & Krosnick, 2009; Jäckle, Roberts, & Lynn, 2010; Klausch, Hox, & Schouten, 2013b), or for matching (Lugtig et al., 2011). Survey practice has shown,

[1] It can be shown that MAR assumptions, when conditioned on a second response mechanism as in A2a/b, are equivalent to the unconfoundedness assumption typically made in causal inference theory (Rubin, 1974). In this literature 'treatment assignment' is normally indicated by one selection mechanism only, say $M = \{a, b\}$ (in words, 'response is through mode A or mode B'), whereas our notation resorts to more than one selection mechanism, because otherwise mode-specific unit nonresponse cannot be described. Unconfoundedness says that $Y_1^a, Y_1^b \perp M \mid X$ which is equivalent to A2a/b where $S_1^a = 1 \Leftrightarrow M = a$ and $S_1^b = 1 \Leftrightarrow M = b$.

however, that socio-demographics are seldom strongly related to response mechanisms or many survey target variables (Couper et al., 2007; Nicoletti & Peracchi, 2005) and mode differences in socio-demographics between modes are often small (Klausch, Hox, & Schouten, 2013a). Adjusted survey estimates also often do not differ greatly from unadjusted estimates (Schonlau et al., 2009). It is, therefore, possible that assumptions (A1) or (A2a/b) do not hold and estimates of marginal or conditional *MEs* may still be biased when adjusting for socio-demographic differences between response samples.

3.3 The between-subject design with follow-up

An extension to the between-subject design is a design where an intensive follow-up is employed to the full sample to obtain strong auxiliary variables for adjustment. Buelens et al (2012) and Schouten et al. (2013) present a between-subject design in which the same sample of respondents is re-interviewed face-to-face in a second wave after some time has elapsed. They weight the first occasion using repeated target variables from the second survey as additional auxiliary data, while assuming (A2a/b) holds conditional on this new information. Since repeatedly measured target variables are correlated more strongly than target variables and socio-demographics, assumptions (A2a/b) are more plausible in this design. If mode A is used as the re-interview mode, this approach estimates ME_R^a.

Buelens et al (2012) discuss various estimation methods to disentangle the mode effects given the follow-up data in their case study on the Crime Victimisation Survey (CVS) within Statistics Netherlands project Mode Effects in Social Surveys (MEPS in Dutch). In the case study, mode-specific selection bias turned out to be relatively small so that the estimation methods produced very similar estimates.

We label this design as a between-subject design with follow-up (BSFU) and will compare it to the within-subject designs presented in this paper. Buelens, Van der Laan and Schouten (2012) provide detailed mode effect estimates for Crime Victimisation (CVS) and Labour Force Survey (LFS) core variables based on the BSFU.

3.4 Other approaches

Other approaches to common missing data adjustment have recently become available. Vannieuwenhuyze, Loosveldt, and Molenberghs (2013) suggested that a *ME* might be explainable by a mediating factor, the so-called 'frontdoor variable' (e.g., 'survey enjoyment'). If the frontdoor variable is not affected by selection bias itself and fully mediates the *ME* between mode and target variable, conditional *MEs* can be estimated using a method described by Pearl (2009, p. 81-85; Morgan & Winship, 2007, p. 224-230). However, currently, there is not any known set of variables that would plausibly fulfill the criteria required from frontdoor variables.

As both missing data and frontdoor adjustment generally lack useful covariates in practice, Vannieuwenhuyze, Loosveldt, and Molenberghs (2010, 2012) suggested a

second alternative, which does not presuppose auxiliary data. The method supposes, however, that data from a mixed-mode survey are readily available. To illustrate, suppose that this is a sequential mixed-mode survey, where nonrespondents to a survey in mode B are offered to reply by alternative mode A (cf. design 2a in section 2). Now a 'comparison sample' is surveyed additionally in mode A only. Again, this design can be considered a between-subject design, where now one of the samples is made up by a sequential mixed-mode survey. To estimate the conditional ME_R^b and the double-conditional $ME_{NR,R}^{b,a}$, it is then assumed that:

$$E(Y_1^a \mid S_2^a = 1 \cup S_1^b = 1) = E(Y_1^a \mid S_1^a = 1) \qquad \text{(A3)}$$

This 'representativeness assumption' says that error due to non-observation on Y_1^a is equivalent for the sample realized by the mixed-mode survey and the single-mode survey in mode A. To the contrary, mixed-mode surveys are often regarded as a solution to reduce non-observation error below that of single-mode designs (de Leeuw, 2005). However, in this case (A3) would not hold. Despite the theoretical importance of the method, its practical applicability seems somewhat limited for this reason. Although it can be argued that some single mode surveys (e.g., face-to-face) achieve about equivalent response rates as mixed-mode surveys (e.g., a sequential design involving face-to-face), response rates provide insufficient indication about (A3) for any given target variable. Furthermore, this argument limits the number of potential mixed-mode designs to which the method applies (i.e., to those with equivalent response rates as a comparison sample). Therefore, the authors call for designs to collect better adjustment covariates (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2013). This suggestion was first followed by Schouten et al. (2013) presenting the between-subject design with follow-up.

However, whereas all approaches in this section merely exploit information for a re-interview as additional adjustment covariates, an alternative is estimating *MEs* directly using repeated measures data from different modes. Then the design, however, is changed to a 'within-subject' structure. There are several advantages connected to this approach, elaborated in the next section. Prominently, the within-subject method allows estimating all (double-) conditional *MEs* defined in section 2.

4. The within-subject design

In essence, the within-subject design allows estimating conditional and double-conditional *MEs* needed to take design decisions reviewed in section 2 under weaker MAR assumptions than earlier approaches reviewed in section 3. The new method does not allow estimating marginal *MEs*, however. As outlined in section 2, this is, fortunately, also not a necessity for effective unimode questionnaire design.

The exposition of the method is structured as follows. We start by providing an outline of the missing data pattern encountered in within-designs and two basic assumptions made during estimation: first, time-stability of answer distributions as well as response probabilities, and second, independence of occasions (section 4.1). We then explain estimation under these basic assumptions (section 4.2). Subsequently, we discuss some of the practical circumstances under which these assumptions are rendered more plausible (section 4.3). Next, we compare assumption in within-subject designs to between-subject designs with follow-up (section 4.4). Finally, we present statistical tests of the basic assumptions and discuss ways to adjust for time-instability (section 4.5). Whereas this exposition is conceptual, section 5 presents an empirical application.

4.1 Outline and basic assumptions

Contrary to a between-subject design (BSD), the within-subject design (WSD) requires surveying the full sample only by a single mode at a first occasion (assume B). After some time has elapsed, this sample is approached again at a second occasion, but then in a different mode (assume A). Any WSD leads to the missing data pattern shown in figure 2. The first occasion provides observations from respondents ($S_1^b = 1$) on Y_1^b (black area) and unit nonresponse (grey area, $S_1^b = 0$). At the second occasion the survey in mode A leads to observations of Y_2^a.

Estimation of *MEs* now makes the central assumption that the observed outcomes on Y_2^a and S_2^a can be used as substitutes for the potential outcomes Y_1^a and S_1^a which are not observed by design at occasion one (white area, figure 2). This substitution defines two basic assumptions of the within-subject design.

First, answer distributions may differ between occasions due to factors related to the progression of time, such as seasonal change of true-scores. In this case Y_2^a is biased against Y_1^a. The means of answer distributions are called *time-stable* if:

$$E(Y_2^a - Y_1^a) = 0 \qquad \text{and} \quad E(Y_2^b - Y_1^b) = 0. \tag{A4}$$

Furthermore, response mechanism S_1^a may change across time, if respondents have different propensities to participate with respect to Y_1^a and Y_1^b. If we have:

$$E(Y_1^a \mid S_2^a = 1) - E(Y_1^a \mid S_1^a = 1) = 0 \quad \text{(A5a)}$$

$$E(Y_1^b \mid S_2^a = 1) - E(Y_1^b \mid S_1^a = 1) = 0 \quad \text{(A5b)}$$

a change in response probabilities does not affect the means of non-observation errors on Y_1^a and Y_1^b, so that these errors may be called time-stable.

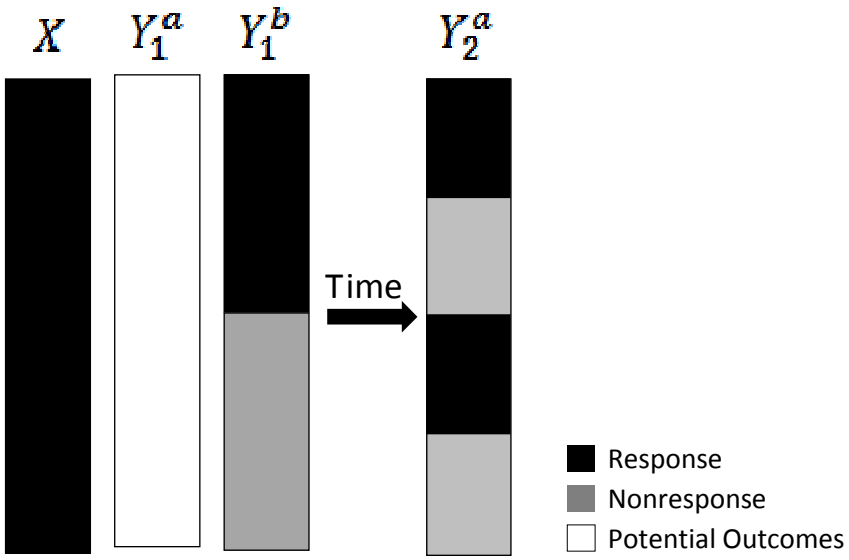

Figure 2: Missing Data Pattern of a Within-Subject Design

Second, being a respondent or a nonrespondent at occasion one might influence the levels of Y_2^a or response probabilities defined by S_2^a. For example, respondents at both occasions might reproduce answers from the first occasion. In this situation, bias of unknown size would be created on Y_2^a. The assumption that respondents at occasion two answer as if the survey at occasion one had not taken place is called '*measurement independence*' (A6a). Similarly, the assumption that response samples are realized as if the survey at occasion one had not taken place is called '*response independence*' (A6b).

It is important to note that if modifications are made to the wave 1 questionnaire in wave 2, i.e. repeating only part of the questions or a slight rewording of the questions themselves, then questionnaire effects may enter and affect wave 2 answers. Absence of such effects would then be seen as part of measurement independence (A6a). They play a role when comparing the WSD to the between-subject design with follow-up (BSFU).

In the next section, we outline estimation assuming both basic assumptions hold true, before discussing them in detail in section 4.3.

4.2 Estimating conditional and double-conditional MEs

Likewise BSDs, WSDs also encounter unit nonresponse under both modes (grey areas, figure 2). To deal with this problem, estimation in within-designs considers unit nonresponse ignorable given the observed outcomes on repeatedly measured *Y*. Formally:

$$Y_2^a \perp S_2^a \mid X, Y_1^b, S_1^b = 1 \quad \textit{(Forward -directed MAR)} \qquad \text{(A7a)}$$

$$Y_1^b \perp S_1^b \mid X, Y_2^a, S_2^a = 1 \quad \textit{(Backward-directed MAR).} \qquad \text{(A7b)}$$

These assumptions imply that observations at occasion one and two are missing at random (MAR) (Rubin, 1976; Little & Rubin, 2002). (A7a) is called forward-directed MAR, because missing data at a later point in time are ignorable on their earlier measurements. Similarly, (A7b) is called backward-directed MAR, because earlier missing data are ignorable on later observations. The assumptions are extensions of the 'unconfoundedness' assumptions in between-designs, (A2a/b), by information from the within-subject design. However, since the partial correlation of Y_1^b and Y_2^a is probably strong in most cases, (A7a/b) appear much more plausible than (A2a/b) when using only socio-demographics as auxiliary data. Likewise (A2a/b), we do not assume, however, that (A7a/b) would hold marginally, i.e. for all individuals, because outcomes are not observed for the group of individuals who are nonrespondents under both modes. (A7a/b) therefore allow estimating (double-) conditional *MEs*, but not marginal *MEs*. Estimating marginal *MEs* would still require ignorable nonresponse on *X* observed for all respondents as stated, for example, in (A1).

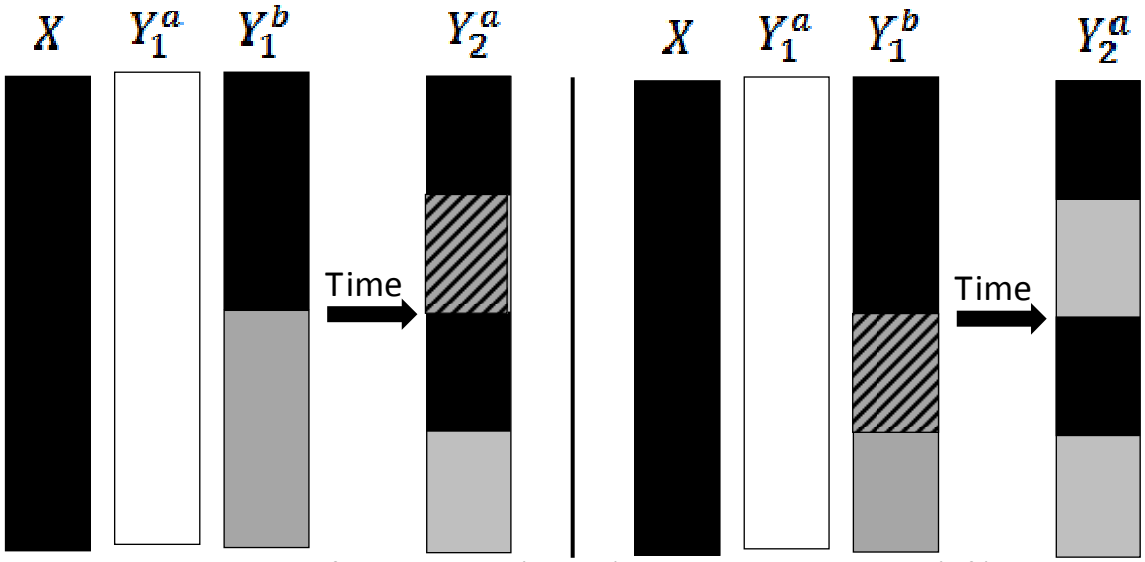

Figure 3: Ignorable part of missing data (dashed) in the Forward Method (left) and Backward Method (right)

4.2.1 The forward method for estimating ME_R^b

The two mean components of ME_R^b, $E(Y_1^b \mid S_1^b = 1)$ and $E(Y_1^a \mid S_1^b = 1)$, are estimated separately. First, $E(Y_1^b \mid S_1^b = 1)$ is estimated directly from the data observed at the first occasion. Second, the mean of potential outcomes $E(Y_1^a \mid S_1^b = 1)$ is not observable by design. Since we assume that answer distributions are time-stable, (A4), we can employ the observations under mode A at occasion two, Y_2^a, as a surrogate to Y_1^a, while using only the information from respondents under mode B at occasion one (i.e., $E(Y_2^a \mid S_1^b = 1)$).

Obviously, some information to estimate $E(Y_2^a \mid S_1^b = 1)$ is missing due to unit nonresponse at occasion two. This problem is illustrated in figure 3 (left), where the missing information is highlighted as a dashed area.

Assuming now forward-directed MAR, (A7a), these missing data are ignorable given the earlier observation of Y_1^b. Consequently, the missing data problem can be solved using techniques such as regression estimation, weighting, matching or imputation.

4.2.2 The backward method for estimating ME_R^a

To estimate ME_R^a, again two conditional means are needed. First, the conditional mean $E(Y_1^a \mid S_1^a = 1)$ is estimated from the observed data under mode A at the second occasion, i.e. $E(Y_2^a \mid S_2^a = 1)$. Under time-stability, (A4) and (A5a), these means are equal.

Second, estimating the mean $E(Y_1^b \mid S_1^a = 1)$ exploits answers from respondents at the first occasion Y_1^b under mode B, but the answers of respondents to mode A are actually needed. However, the within-design facilitates observing the response mechanism A at the second occasion, so that we estimate the conditional mean $E(Y_1^b \mid S_2^a = 1)$ instead. Since we assume that non-observation error on Y_1^b is stable across time, (A5b), $E(Y_1^b \mid S_2^a = 1)$ equals $E(Y_1^b \mid S_1^a = 1)$.

Again we face a missing data problem due to nonrespondents under mode B at occasion one, who are respondents at occasion two. This missing part of the data is illustrated by the dashed area in figure 3 (right). However, assuming data are backward-directed MAR (A7b), the conditional mean can be estimated.

4.2.3 Estimating double-conditional MEs

Double conditional *MEs* are defined for sequential mixed-mode designs (cf. section 2), and conditioned on two response mechanisms. In particular, $ME_{NR,R}^{b,a}$ (formula 4) is defined for respondents to mode A at a second occasion who were nonrespondents under mode B (i.e., $\{S_1^b = 0, S_2^a = 1\}$). The mean $E(Y_2^a \mid S_1^b = 0, S_2^a = 1)$ is observed in the within-subject design. In fact, sequential mixed-mode designs can be regarded as a variant of within-subject designs, in which respondents at the first occasion are not followed up in an alternative mode.

Assuming time-stability we now only need to estimate $E(Y_1^b \mid S_1^b = 0, S_2^a = 1)$, for which observations are missing due to nonresponse under mode B at occasion one. This missing part of the data is illustrated by the dashed area in figure 3 (right). Assuming backward-directed MAR, (A7b), this mean can be estimated, since then $E(Y_1^b \mid S_1^b = 0, S_2^a = 1) = E(Y_1^b \mid S_1^b = 1, S_2^a = 1)$.

The second double-conditional $ME_{NR,R}^{a,b}$ exchanges the order of modes. Now nonrespondents under mode A, who are respondents under mode B are of interest. Since independence of occasions is assumed (A6a/b), the order in which modes are presented is irrelevant for $ME_{NR,R}^{a,b}$. Estimation is then based on the Forward Method.

4.3 Design considerations about time-stability and independence

Crucial to the forward and the backward method are the assumptions of time-stable answer distributions (A4), time-stable response probabilities (A5a/b), as well as measurement and response independence (A6a/b).

On the one hand, time-stability mainly depends on exogenous factors related to the progression of time, such as seasonal change of true scores (A4) or survey climates in societies (A5a/b). The extent of change thus also depends on the type of target variables and populations at hand. In general, however, shorter time lags between occasions let time stability appear more plausible regardless of particular variables and populations.

On the other hand, measurement and response independence depends more heavily on design inherent factors and thus might be controllable to greater extents by fieldwork design. For example, respondents' ability and motivation to recall answers from the first occasion affects measurement independence (A6a). Cognitively salient questions take longer time to be forgotten and thus require longer time lags between occasions. More generally, measurement independence is a rather common assumption in social research. It is of importance in the MTMM literature (Saris, Satorra, & Coenders, 2004), for example, or during psychometric testing (e.g., intelligence, mathematical abilities), which normally assumes independence of answers to similar items.

Response dependence (A6b) can be caused by heavy response burden during the first occasion. Several fieldwork precautions can reduce this risk, however. First, interview length of the first survey should be short. Second, during the recruitment of the second survey it should be made clear to all individuals why (repeated) participation is necessary (e.g., to ask additional questions on same topic). Third, if interviewers are employed at the second occasion, they require special training to deal with respondents' questions about the first occasion. Fourth, in order not to influence response probabilities at the first occasion, individuals should be kept unaware at occasion one about the follow up survey. Finally, longer time lags between occasions

also turn response independence much more likely, as response probabilities can normalize across time.

Curiously, time-stability is more likely to hold for shorter, whereas both independence assumptions are more likely for longer time lags. Hence, there is a tradeoff in terms of the timing of occasions. We suggest that prior knowledge about target variables and response probabilities (e.g., from time-series data) can give an additional indication about the time frame in which stability can be assumed. If time-stability can be assured, choosing longer time lags is always advisable to guarantee both independence assumptions. Another option is thorough pre-study that evaluates across which time periods important target variables and response probabilities can be considered time stable and independent.

Not always can a thorough design reduce the risk that some basic assumptions do not hold. Therefore, separate testing of the assumptions offers additional reassurance. In the next section several tests are presented in detail.

4.4 Differences to the between-subject design with follow-up

The BSFU and the WSD rely differently on the variables measured in the follow-up wave. For the BSFU the wave 2 variables are merely adjustment covariates, while for the WSD they are repeated measurements and are employed also to derive estimates for double-conditional *MEs*. Contrary to WSD, BSD and BSFU are not designed to estimate double-conditional *MEs* or the single mode ME_R^b. The BSD and BSFU are designed to estimate ME_R^a only, where A is the follow-up mode, so that they can only be compared to WSD with the backward method.

Essentially, BSFU and WSD differ only in two assumptions: the measurement independence assumption (A6a) and the unconfoundedness assumptions (A2a and b). Whereas WSD has to make the measurement independence assumption in order to view the follow-up wave as a repeated measurement, in BSFU this assumption does not have to be made. However, that does not imply that BSFU estimation is not affected by any measurement dependence. In WSD, the unconfoundedness assumptions only have to be valid for differences in answers between modes for the same respondent. For BSFU the unconfoundedness assumptions apply to the answers themselves, which is (intuitively) a stronger assumption. Any measurement dependence threatens the unconfoundedness assumptions as the adjustment covariates are likely to become less informative. This threat is anticipated to be larger for BSFU than for WSD. By and large, the two designs may impose similar demands on data collection, but as mentioned, WSD explicitly includes estimation of double-conditional *MEs* and also allows for the estimation of ME_R^b, which is interesting when redesigning a single mode face-to-face survey to a sequential design with face-to-face as follow-up mode.

A real difference between BSFU and WSD, however, arises when the questionnaire is considerably revised or shortened for the follow-up wave. Any questionnaire effects are then combined with measurement dependence and it is likely that BSFU is more

robust. As an extreme case one may consider a follow-up wave with repeated measures rather than with repeated measurements, i.e. in the follow-up wave different items are asked that are anticipated to load strongly on the same latent variables as the original items; WSD would then not be possible.

4.5 Testing and adjusting for time-instability and dependence of occasions

Estimates from the WSD can be biased, if the time-stability (A4 and A5a/b) and independence assumptions (A6a/b) do not hold. In this section, we present design extensions that help to assess the potential for violations of the basic assumptions in practice. It is possible to adjust for bias due to time-instability, furthermore.

Tests for time-stability are based on an independent, parallel sample, in which the mode is kept constant across occasions to avoid confounding of *MEs* with time-related change. This sample is then referred to as a control group and the resulting design is called a 'within-subject control-group' design (WSCG). Control groups represent a common approach to exclude time related change as an alternative explanation to a treatment in interrupted time-series designs (Winship & Morgan, 1999).

If mode A is used for the control group, the missing data pattern of the resulting WSCG design is shown in figure 4. In this case, WSCG designs are equivalent to between-subject designs extended for a second measurement occasion. Based on this WSCG design it is possible to estimate and adjust for time-related bias assuming that change observed in the control group is equivalent in the sample initially interviewed in mode B ('treatment').

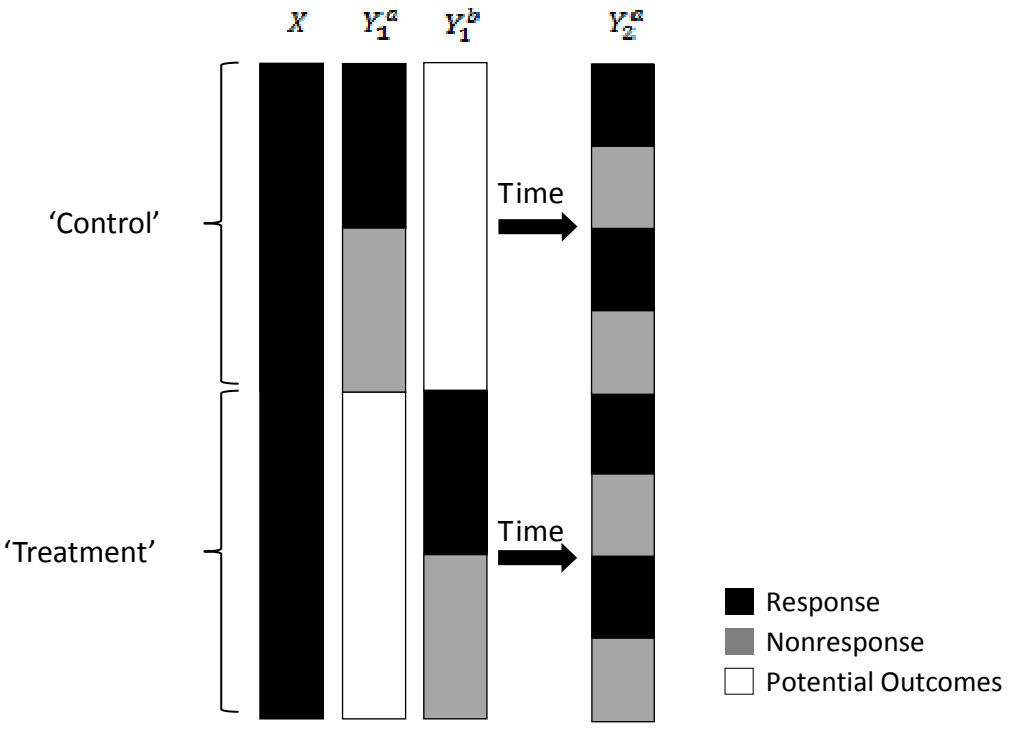

Figure 4: Missing Data Pattern of a Within-Subject Control-Group Design

4.5.1 Testing and adjusting time-related bias of the forward method

The Forward Method might introduce bias if Y_1^a changes across time, i.e.,

$$\Delta_1 = E(Y_1^a \mid S_1^b = 1) - E(Y_2^a \mid S_1^b = 1), \quad (11)$$

because observations from the second occasion are used instead of the potential outcomes at the first. Data from the control group design is used estimate this bias, assess its size, and adjust the estimator appropriately.

However, whereas the control group design can be applied to estimate change in Y_t^a for respondents in mode A, the change Δ_1 is defined for respondents in mode B. It, therefore, needs to be assumed that change on Y_t^a is independent of responding in modes A or B.

$$\Delta_1 = E(Y_1^a \mid S_1^a = 1) - E(Y_2^a \mid S_1^a = 1). \quad (A8a)$$

The assumption implies that we allow for change of Y_t^a, but we assume that there is no 'second order effect' on the change (caused by selection bias). This can be considered a weaker assumption than assuming time-stability for Y_t^a globally (A4). Put differently, if a change in the control group (mode A) is insignificant, it becomes more likely that this also holds for the group of respondents under mode B. Still, an estimate of Δ_1 using the control group can also be biased, if respondents in mode B change differently than respondents in mode A.

Estimation of Δ_1 follows the Forward Method, but now applied to the control group. Whereas $E(Y_1^a \mid S_1^a = 1)$ is directly observable from the control group at occasion one, $E(Y_2^a \mid S_1^a = 1)$ needs to be estimated in the presence of missing data from nonrespondents at occasion two, who are respondents at occasion one. This quantity can be estimated using the Forward Method assuming:

$Y_2^a \perp S_2^a \mid X, Y_1^a, S_1^a = 1$ *(Forward-directed MAR in control group)* (A9a)

4.5.2 Testing and adjusting time-related bias of the backward method

Contrary to the Forward Method, the Backward Method can lead to two types of time-related biases in the within-subject design. First, we have:

$$\Delta_2 = E(Y_1^a \mid S_1^a = 1) - E(Y_2^a \mid S_2^a = 1) \quad (12)$$

which is caused by substituting the mean of respondents at occasion one by respondents at occasion two. Since $E(Y_1^a \mid S_1^a = 1)$ can be estimated directly from the

control-group at occasion one, this bias is avoided immediately (put differently, Δ_2 is estimated without bias in the WSCG design).

Second, we might introduce bias:

$$\Delta_3 = E(Y_1^b \mid S_2^a = 1) - E(Y_1^b \mid S_1^a = 1) \tag{13}$$

which is caused by exploiting the response mechanism at the second occasion instead of the first. This bias is equivalent to the change in non-observation error on Y_1^b from occasion one to occasion two.
Using the control group it is again possible to estimate this bias, when we assume that the change in non-observation error on Y_1^b is equivalent to the change on Y_1^a which is observable in the control group:

$$\Delta_3 = E(Y_1^a \mid S_2^a = 1) - E(Y_1^a \mid S_1^a = 1) \tag{A8b}$$

This assumption appears to be weaker than assuming time-stability of non-observation error (A5b), because when a change in non-observation error on Y_1^a is insignificant, it appears more plausible to assume this change is also small on Y_1^b. Put differently, we assume that there is no second-order effect of the mode of measurement on non-observation error.

Estimating Δ_3 is straightforward exploiting the Backward Method to find the conditional mean $E(Y_1^a \mid S_2^a = 1)$, assuming:

$Y_1^a \perp S_1^a \mid X, Y_2^a, S_2^a = 1$ (Backward-directed MAR in control group). (A9b)

4.5.3 Testing for measurement and response independence

Our ability to test for dependence of occasions (A6a/b) based on a WSCG design is somewhat more limited. In a WSCG design it can be assessed, whether the assignment to modes A and B at the first occasion (M_1), which is fully random, has an impact on $P(Y_2^a \mid S_2^a = 1)$. If:

$$\Delta_4 = E(Y_2^a \mid S_2^a = 1, M_1 = a) - E(Y_2^a \mid S_2^a = 1, M_1 = b) \neq 0 \tag{14}$$

A6a/b do not hold. This test, however, is not exact, because dependence is tested as a compound (jointly A6a/b). Furthermore dependence can only be detected, if its effects on $E(Y_2^a \mid S_2^a = 1, M_1)$ vary across modes assigned at the first occasion of the WSCG design. If:

$$E(Y_2^a \mid S_2^a = 1, M_1 = a) = E(Y_2^a \mid S_2^a = 1, M_1 = b) \neq E(Y_2^a \mid S_2^a = 1) \tag{15}$$

dependence remains hidden. To avoid this problem it is possible to draw an additional independent sample surveyed only at the second occasion in mode A. Since this sample is fully independent from the first occasion, assessing:

$$\Delta_5 = E(Y_2^a \mid S_2^a = 1) - E(Y_2^a \mid S_2^a = 1, M_1 = b) \tag{16}$$

is superior to assessing significance of Δ_4.

Tests for independence can also be conducted based on exogenous information *X*. However, since *X* might often only be weakly related to survey variables, these tests are probably insufficient in practice.

5. Illustration

In 2011, within Statistics Netherlands' project Mode Effects in Social Surveys (MEPS in Dutch) a large-scale mixed-mode experiment was conducted (Klausch et al. 2013a, 2013b; Schouten et al., 2013). Three separate WSCG designs were administered in parallel with a sample size of 6,803. Mode B was represented by telephone (n=1,658), mail (n=1,760), and web (n=1,746), respectively. Mode A was chosen to be F2F, where a sample of n=1,639 served as control. The survey topic was the national Crime Victimization Survey (CVS).

On average, 6 weeks lay between the first and second interview. From earlier implementations of the CVS, administered on yearly basis, it was known that many statistics only changed slowly or were stagnating, so that time-stability appeared plausible for this time frame. To reassure about this assumption, however, the control group was included.

Additionally, several precautions were taken in advance to assure independence of occasions. First, no reference to the second occasion was made at the first occasion including no mentioning of the possibility to reply later in a F2F mode. Interviewers, furthermore, were only vaguely informed about the second occasion while administering occasion one. This avoided that some individuals did not reply at the first occasion, because they preferred the forthcoming F2F mode at occasion 2. Second, the fieldwork at occasion two could hardly be distinguished from occasion one. No separate advance letter was sent at occasion two and instead interviewers contacted all individuals directly. F2F interviewers were instructed to recruit nonrespondents from occasion one as for a regular F2F survey and they were trained to explain respondents at occasion one the need for repeated participation for answering additional questions from the CVS.

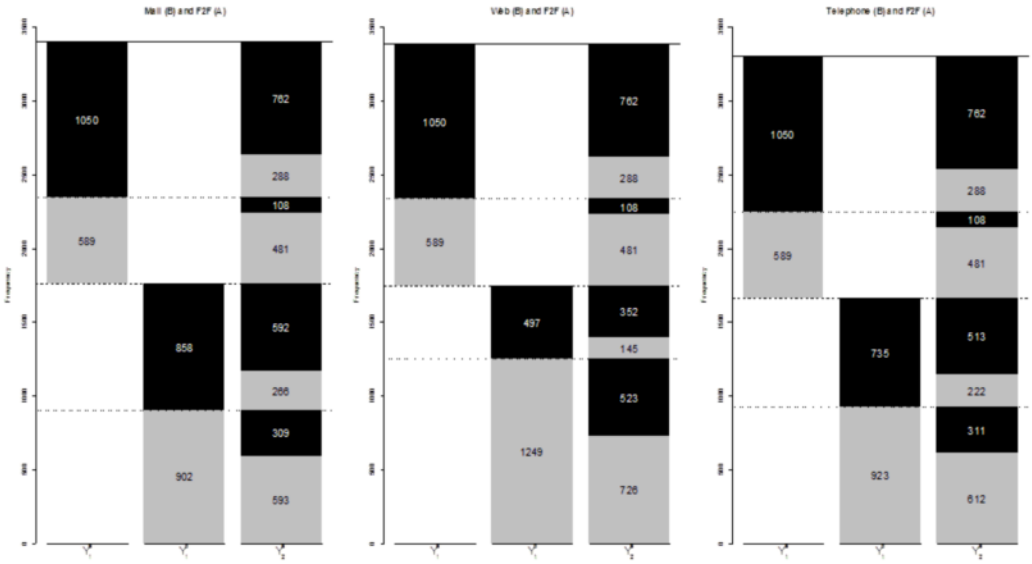

Figure 5: Real-world missing data patterns of three within-subject control-group designs using face-to-face as a reference mode (A)

Figure 5 shows the missing data patterns of the three WSCG designs in analogy to the schematic illustration introduced in section 4.4 (Figure 4). The upper part of the three plots is represented by the F2F control group (mode A; the same sample in the three designs, respectively). The lower parts show the nonresponse and response proportions of telephone, mail, and web (modes B, respectively). Clearly, mail (left) and telephone (right) achieved higher response (n=858 and n=735, respectively) than web (middle, n=497). Response was highest in F2F (n=1050).

The BSFU method has been applied to the same CVS data set by Buelens, Van der Laan and Schouten (2012) and Schouten et al (2013). In this paper, we limit ourselves to the exposition of the WSD and WSCG methods and leave comparisons to the BSFU method to future papers.

5.1 A two-mode comparison

We now illustrate estimation of (double-) conditional *MEs* and their selection biases (7) and (8) against the naïve *ME* estimator (6), which could be estimated in between-subject designs. We start by comparing only two modes, mail and F2F (left missing data pattern in figure 5). The variable selected for this example is an index called the 'Social Quality of the Neighborhood' (variable label *A-sockwal*). It is based on multiple rating scale questions about social cohesions in the neighborhood forming a summary score ranging from one to ten. It is a regularly reported statistic from the CVS.

We estimated the (double-) conditional *MEs* and the coefficients Δ using generalized regression estimation (e.g., Imbens, 2004; Schafer & Kang, 2008). In addition to the target variables, eight socio-demographic indicators (X) were available[2]. The model was built using forward inclusion of those covariates that minimized the model AIC[3], where the target variable was included in the first step. All standard errors were estimated using the bootstrap with 10,000 replications. At each replication a new adjustment model was fitted.

Table 2 gives an overview on point estimates, confidence intervals and significance levels of two sided tests against zero. The left column presents estimates from the within-subject (i.e., pretending the control group was not available), the right column from the WSCG design. Consider first the within-subject design, in which time-stability is assumed and cannot be tested or adjusted. The naïve *ME* estimator specified in (6) is based on the response means at occasions one (mail) and two (F2F), respectively. We find that there is a significant naïve *ME* (-.771). This effect can be decomposed into conditional *MEs* -.435 and -.445 and selection biases -.336 and -.326, respectively. The conditional *MEs* are significantly smaller than the naïve estimator would suggest (by roughly 40%) indicated by significant selection biases. The double-conditional *MEs* can be taken from the last row of table 2 and are equal in magnitude to the conditional

[2] In particular: gender, age, income, civil status, nationality, household size, urbanity, and inhabitation of a large city in The Netherlands.

[3] Akaike Information Criterion

MEs. The right column presents all adjusted estimates from the WSCG design as well as the adjustment coefficients Δ_1 to Δ_3 estimated from the F2F control group. Only Δ_3 is significant, suggesting a small change in non-observation error from the first to the second occasion (formula 13). The adjusted estimate of ME_R^b does not differ greatly from the WSCG design, however, because the sign of Δ_2 counteracts Δ_3.

Table 2: Overview on ME estimates from the within-subject and WSCG design (mode A: F2F, mode B: mail; Variable: 'Social Quality Index', A_sockwal)

	Within-Subject Design			WSCG Design		
	Est.	95% CI	p[a]	(Adj.) Est.	95% CI	p[a]
ME_R^{Naive}	-.771	[-.929,-.611]	<.001	-.731	[-.928,-.534]	<.001
ME_R^a	-.435	[-.559,-.313]	<.001	-.488	[-.679,-.296]	<.001
ME_R^b	-.445	[-.556,-.334]	<.001	-.390	[-.526,-.254]	<.001
$SE(Y_1^b)$	-.336	[-.414,-.073]	<.001	-.243	[-.414,-.073]	.006
$SE(Y_1^a)$	-.326	[-.546,-.138]	<.001	-.341	[-.546,-.138]	.001
Δ_1	-	-	-	-.055	[-.135,.023]	.175
Δ_2	-	-	-	-.040	[-.203,.125]	.634
Δ_3	-	-	-	.094	[.009,.181]	.032
Δ_4	-	-	-	-.083	[-.241,.075]	.291
$ME_{NR,R}^{b,a}$	-.414	[-.588,-.252]	<.001	-.414	[-.588,-.252]	<.001
$ME_{NR,R}^{a,b}$	-.465	[-.609,-.319]	<.001	-.465	[-.609,-.319]	<.001

a: bootstrapped p-value (10^4 draws) of a two-sided test against zero

Furthermore, an approximate test for independence is available (Δ_4 , formula 15). We find an insignificant Δ_4 suggesting that the type of mode offered at occasion one (mail or F2F) did not have any impact on the conditional distribution of the Social Quality index at occasion 2. This is reassuring about the independence of occasions. We advise to only report *ME* estimates from WSCG designs when independence tests are insignificant. Note that Δ_5 (16) cannot be calculated in this design, because an independent sample at the second occasion is not available.

The conditional *ME* estimates suggest that respondents under mail and F2F answer with differing extents of measurement error. In the CVS, F2F was the standard mode of administration for a long time. In the hypothetical situation that the design would be switched completely to mail, for example, a break in this time series could be expected. If F2F is assumed to produce less measurement error, furthermore, the mail mode

would evoke higher extents of error. A sequential design combining mail and F2F is not advisable, because the (double-) conditional *MEs* imply that respondents under both modes provide different answers. Further research could evaluate, whether a unified mode design of the questions underlying the Social Quality Index makes it possible to reduce the (double-) conditional *MEs*.

5.2 Visualization of measurement effects

Another option to interpret the decomposition of naïve *MEs* into conditional *MEs* and selection bias is plotting all conditional means in an interaction diagram, as illustrated for three examples in figure 6.

The first example (upper plot) is based on the (adjusted) estimates presented in table 2. Each line represents a group of respondents in mode A or B. The estimates on the left side, furthermore, represent the conditional means of answers under mode A (Y_1^a), and the estimates on the right side the means of answers under mode B (Y_1^b). Now, the 'slope' of the two lines provides the conditional *MEs*, respectively, whereas the vertical distance between lines represents the selection bias. Selection bias and conditional *MEs* add up to the naive *ME*, respectively. All effects are indicated by labeled arrows.

An advantage of interaction plots is that the scale of all variables can be taken from the ordinate, whereas it cannot be taken from table 2. Interaction plots also give a quick overview on the size and type of conditional *MEs*. We illustrate this idea using two further variables. In the middle plot another index, the extent of 'neighborhood problems', is presented. Mode B is the web mode in this example. It can be seen that the naïve *ME* is very small (and testing demonstrates it is insignificant). However, adjusting for selection bias, conditional *MEs* are uncovered (significant), which are suppressed by counter directional selection biases ('suppression effect').

In the last example, the variable 'safety perception of the neighborhood' is presented. It is measured on dichotomous level (yes/no), so that proportion estimates are depicted here. Contrary to the two previous plots, the two lines slightly intersect. This is the situation of an 'interaction effect' between *MEs* and selection mechanisms, which causes the conditional *MEs* to be unequal (whereas ME_R^b is negligible in the plot, ME_R^a is more substantial). In this case, conclusions drawn from ME_R^b and ME_R^a would differ.

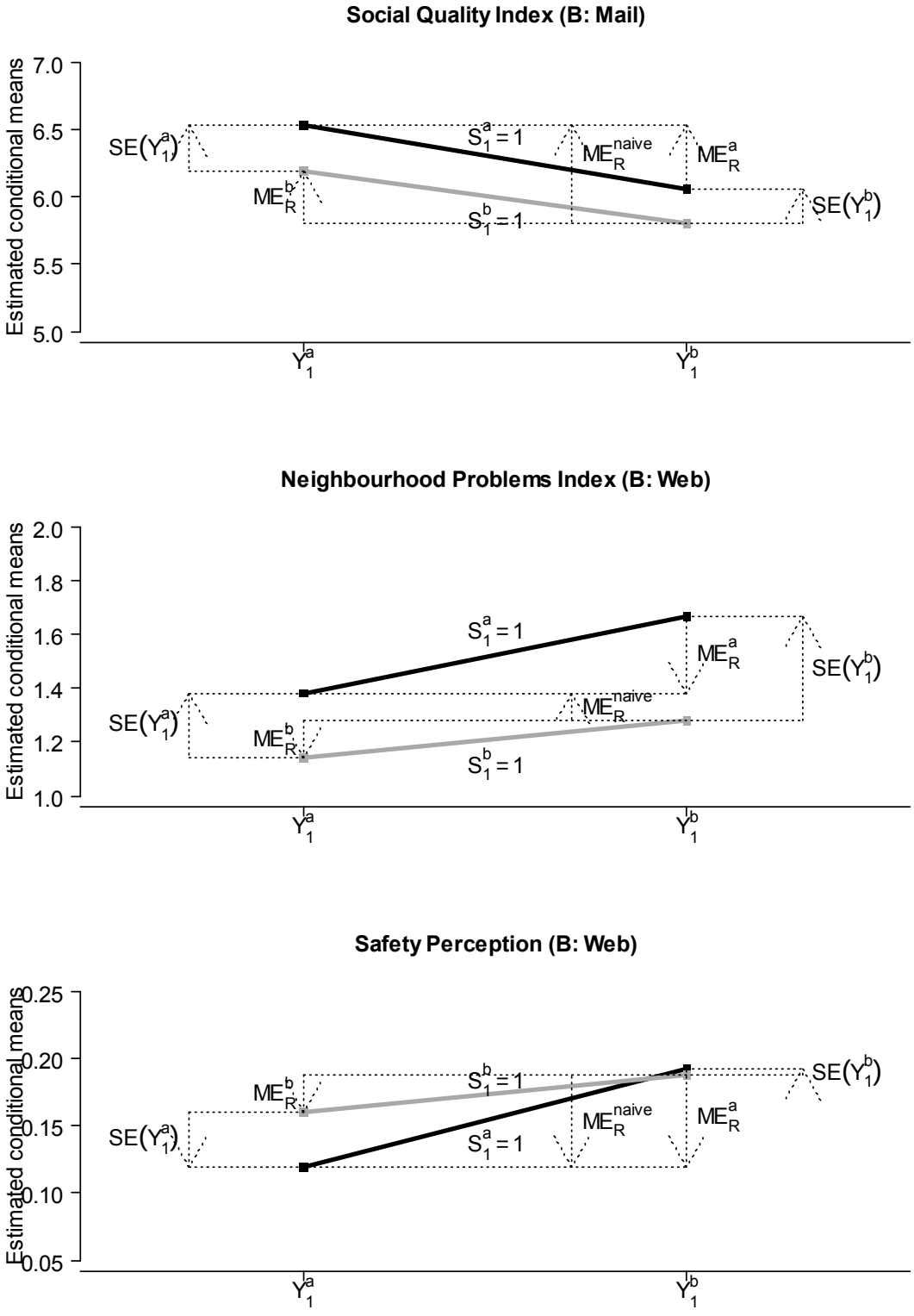

Figure 6: Example of three different types of MEs (based on within-subject control-group designs) visualized by interaction diagrams (different variables and modes indicated in the figures). Variables depicted are "A_sockwal"," overlast" and "onveilig".

5.3 A multi-mode comparison

Comparisons of conditional *MEs* across multiple modes are useful to assess the feasibility of different mode combinations in mixed-mode designs. Figure 7 presents interaction plots for all modes included in the experiment, where F2F serves as a reference mode (mode A), respectively. Since the reference mode does not change across figures, relative comparisons of *MEs* between all modes are possible. The variable used for the plots is again the 'Social Quality Index'. Therefore, the first (left) plot is identical to the upper plot in figure 6 and estimates in table 2. Furthermore, the middle plot illustrates the comparison of web and F2F, and the right plot telephone and F2F.

Clearly, the three plots differ. Comparing the plots of web and mail it can be noticed that the adjusted conditional *MEs* against F2F are equal in size. However, for the web mode no selection bias against the naïve *ME* estimate is identified, whereas for mail it is significant (cf. table 2). Concluding only on the basis of the naïve *ME*, a researcher would thus have incorrectly assumed that the mail mode evokes stronger *MEs* than the web mode when compared against F2F. However, there are no *MEs* between web and mail (the *MEs* against F2F are equal). Considering now the comparison of telephone and F2F (right plot) we find a 'collapsed' graph suggesting that no conditional *MEs* are present.

In summary, these graphs show a typical situation, in which interviewer and self-administered modes exhibit significant *MEs*, but telephone and F2F as well as mail and web, respectively, produce virtually no differences in measurement error (e.g., de Leeuw, 1992). Sequential mixed-mode designs and single-to-single mode switches involving only interviewer modes or only self-administered modes therefore seem feasible using the current questionnaire design, but an interview mode should not be combined with a self-administered mode. Frankly, these conclusions are based on a single variable and should be reproduced for others before taking design decisions.

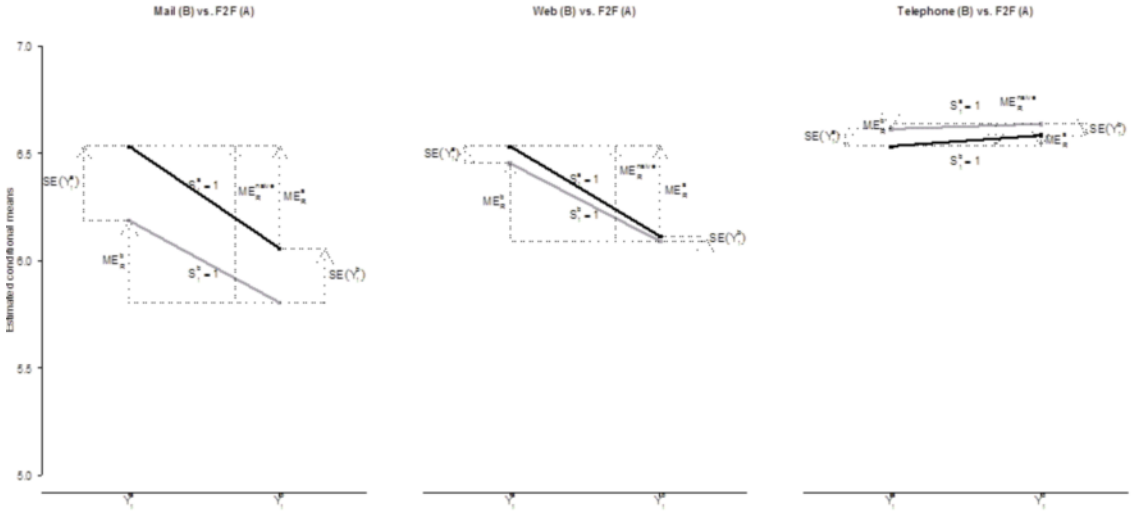

Figure 7: Multi-mode comparison of mail, web, and telephone (mode B, respectively) against F2F (reference mode A) for the 'Social Quality Index' from the Crime Victimization Survey (estimates based on within-subject control-group designs)

6. Discussion

We presented a new approach to estimate conditional and double-conditional *MEs* using within-subject designs. A key advantage of our method is its grounding on weaker MAR assumptions than earlier approaches commonly make in practice. In particular, we argued that it is more plausible to ignore nonresponse conditional on repeatedly measured target variables than ignoring it on socio-demographic variables only, which is common practice in lack of other exogenous covariates or useful frontdoor variables (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2013). Furthermore, we argued that the representativeness assumption of the instrumental variable method probably does not hold for many mixed-mode designs (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010, 2012). It also always requires a mixed-mode design to be administered before *MEs* are known. To the contrary, the within-subject method allows estimating *MEs* before designing a sequential mixed-mode survey or deciding about a single-to-single mode switch.

However, our approach entails two new types of assumptions: time stability and independence of occasions. Fortunately, both assumptions can be tested using control group extensions of the within-subject design. Time instability can even be adjusted in the WSCG design. Regardless of these tests, within-subject experiments should always be designed to render time-stability and independence most likely. Careful fieldwork design, prior knowledge about target variables, and advance testing are important aspects of this process as discussed in detail in section 4.3 and illustrated by an empirical application in section 5. Future research needs to assess in more general terms, which design aspects can render independence more likely in order to guide the design of within-subject experiments. Importantly, it needs to be evaluated which time lags vis-à-vis salience and centrality of questions are required for measurement and response independence.

The WSD and WSCG designs presented in this paper have a resemblance to the design proposed by Buelens et al (2012) and Schouten et al (2013), that we termed a between-subject design with follow-up. The difference lies in the estimators used and the corresponding assumptions that are made, but, they essentially rely on the same data collection. In future papers, we will compare estimates which are anticipated to be very similar in many cases.

The *MEs* that can be estimated using the new method are useful to take decisions about changes in measurement error caused by single-to-single mode switches as well as in sequential mixed-mode designs (section 2). In this respect, we note an important difference of our approach and earlier definitions suggested by Vannieuwenhuyze & Loosveldt (2013). The authors suggest design-specific *ME* estimands and assume that effects are to be estimated from data provided by an ongoing mixed-mode survey. To the contrary, we defined *MEs* more generally and demonstrated their applicability to two specific mixed-mode scenarios. However, our estimation method is based on an

experiment that is conducted independently of any mixed-mode survey and thus can include any number of candidate modes, before deciding on the final mixed-mode design. The fieldwork design of the modes used in the experiment (e.g., advance letters, interviewer training and questionnaires) should therefore be kept as similar as possible to the design of the final mixed-mode survey to allow valid conclusions. While the sequential mixed-mode designs and single-to-single mode switches are common (re-) design options, conclusions about other types of designs, such as 'concurrent mixed-mode' or complex sequential designs involving more than two modes cannot be drawn, yet. Extensions could be developed in future research, however.

Another important design aspect is represented by a tradeoff between costs, sample size and related power. Clearly, a within-subject design yields approximately equivalent costs as a between-subject design. However, the control group added in the WSCG design can strongly increase costs, especially if based on an expensive mode as chosen in the illustration (F2F). Furthermore, an exact test of independence requires an additional independent sample at the second occasion (cf. formula 16), which increases costs further. In the illustration, a sample size of approx. 1,700 units in the treatment and control conditions, respectively, lead to rather large standard errors relative to the size of *MEs* (see, e.g., table 2). It is therefore important to assess the necessary sample size for acceptable detection power of *MEs* vis-à-vis acceptable costs.

Standard errors of estimates are also influenced by the efficiency of the estimation method. Under correct MAR assumptions any of the major techniques, such as regression estimation, weighting, or imputation yields unbiased estimates, but variance of estimates may differ. Currently, there is ongoing debate beyond this discussion, on the 'best' estimation method for parameters and their standard errors when observations are missing at random (e.g., Little & Rubin, 2002; Imbens, 2004; Schafer & Kang, 2007; Kang & Schafer, 2008).

The new approach suggested in the present paper implies two further important paths for applications and methodological development. First, based on our framework a correction method for *MEs* can be developed. A method could predict or impute plausible values for potential outcomes not observed in a mixed-mode design. As long as it is possible to collect repeated measurements a correct model for individual-level *MEs* can form the basis of adjustments. In any mixed-mode design, repeated measurements could be administered at least for a sub-sample, for example. This path urgently needs to be followed in the future. Second, the method naturally relates to more complex data structures known from panel surveys. Mixed-mode panels combine different modes, often by switching modes between panel waves. Our method then can be applied repeatedly to estimate and adjust *MEs* across waves. Frankly, plausibility of the basic assumptions will remain pivotal in these applications.

References

Aquilino, W. S. (1994). Interview Mode Effects in Surveys of Drug and Alcohol Use: A Field Experiment. *Public Opinion Quarterly, 58*(2), 210–240.

Bethlehem, J. (1988). Reduction of Nonresponse Bias Through Regression Estimation. *Journal of Official Statistics*, *4*(3), 251–260.

Bethlehem, J. (2002). Weighting Nonresponse Adjustment Based on Auxiliary Information. In R. M. Groves, D. A. Dillman, J. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse*. New York: Wiley & Sons.

Biemer, P. P. (2001). Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. *Journal of Official Statistics, 17*(2), 295–320.

Buelens, B., Van der Laan, J., Schouten, B., Van den Brakel, J., Burger, J., Klausch, T. (2012), Disentangling mode-specific selection and measurement bias in social surveys, Discussion paper 201211, CBS, Heerlen.

Buelens, B., Van der Laan, J., Schouten, B. (2012), Decomposition of mode effects for CVS and LFS, Technical report, BPA PPM-2012-02-25-BBUS-DLAN-BSTN, CBS, Den Haag.

Chang, L. & Krosnick, J. A. (2009). National Surveys Via Rdd Telephone Interviewing Versus the Internet. *Public Opinion Quarterly, 73*(4), 641–678.

Cochran, W. G. (1977). *Sampling Techniques* (2nd ed.). New York: Wiley.

Couper, M. P., Kapteyn, A., Schonlau, M. & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research*, *36*(1), 131–148.

De Leeuw, E. (1992). *Data Quality in Mail, Telephone, and Face to Face surveys.* Amsterdam: TT-Publicaties.

De Leeuw, E. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, *21*(2), 233–255.

Dillman, D. A. & Christian, L. M. (2005). Survey Mode as a Source of Instability in Responses across Surveys. *Field Methods*, *17*(1), 30–52.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, *38*(1), 1–18.

Dillman, D. A., Smyth, J. D. & Christian, L. M. (2009). *Internet, Mail, and Mixed-Mode Surveys. The Tailored Design Method.* New Jersey: Wiles & Sons.

Fricker, S., Galesic, M., Tourangeau, R. & Yan, T. (2005). An Experimental Comparison of Web and Telephone Surveys. *Public Opinion Quarterly*, *69*(3), 370–392.

Fricker, S. & Tourangeau, R. (2010). Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys. *Public Opinion Quarterly*, *74*(5), 934–955.

Groves, Robert M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Groves, Robert M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2010). *Survey Methodology* (2nd ed.). New Jersey: Wiley.

Heerwegh, D. (2009). Mode Differences Between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research*, *21*(1), 111 –121.

Heerwegh, D. & Loosveldt, G. (2008). Face-to-Face versus Web Surveying in a High-Internet-Coverage Population. *Public Opinion Quarterly*, *72*(5), 836–846.

Holbrook, A. L., Green, M. C. & Krosnick, J. A. (2003). Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias. *Public Opinion Quarterly*, *67*(1), 79–125.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

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

Jäckle, A., Roberts, C. & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, *78*(1), 3–20.

Kaminska, O., McCutcheon, A. L. & Billiet, J. (2010). Satisficing Among Reluctant Respondents in a Cross-National Context. *Public Opinion Quarterly*, *74*(5), 956–984.

Kang, J. D. Y. & Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, *22*(4), 523–539.

Klausch, T., Hox, J. & Schouten, B. (2013a). *Assessing the mode-dependency of sample selectivity across the survey response process*. Statistics Netherlands Discussion Paper 201303, Den Haag. [online] Available at http://www.cbs.nl/NR/rdonlyres/D285D803-D201-437D-99F6-3FB7C5DA9C11/0/201303x10pub.pdf (Accessed 07-22-2013)

Klausch, T., Hox, J. J., & Schouten, B. (2013b). Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions. *Sociological Methods & Research, 42*(3), 227–263.

Kreuter, F., Presser, S. & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys. *Public Opinion Quarterly*, *72*(5), 847 –865.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology, 50*(1), 537–567.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: Wiley.

Lugtig, P., Lensvelt-Mulderts, G. J. L. M., Frefrichs, R. & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*(5), 669–686.

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Nicoletti, C. & Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(4), 763–781.

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

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). New York: Springer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational and Behavioral Statistics*, *2*(1), 1–26.

Rubin, D. B. (1988). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331.

Saris, W. E., Satorra, A. & Coenders, G. (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology, 34*(1), 311–347.

Särndal, C.-E. & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

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

Schafer, J. L. & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313.

Schonlau, M., van Soest, A., Kapteyn, A. & Couper, M. (2009). Selection Bias in Web Surveys and the Use of Propensity Scores. *Sociological Methods & Research*, *37*(3), 291 –318.

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J. & Berry, S. H. (2004). A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review*, *22*(1), 128 –138.

Schouten, B., Brakel, J. van den, Buelens, B., Laan, J. van der & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, 42, 1555 – 1570

Tourangeau, R., Rips, L. J. & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Tourangeau, R. & Smith, T. W. (1996). Asking Sensitive Questions. The Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly*, *60*(2), 275–304.

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

Vannieuwenhuyze, J., Loosveldt, G. & Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly*, *74*(5), 1027 –1045.

Vannieuwenhuyze, J., Loosveldt, G. & Molenberghs, G. (2012). A Method to Evaluate Mode Effects on the Mean and Variance of a Continuous Variable in Mixed-Mode Surveys. *International Statistical Review*, *80*(2), 306–322.

Vannieuwenhuyze, J., Loosveldt, G. & Molenberghs, G. (2013). Evaluating mode effects in mixed-mode survey data using covariate adjustment models. *Forthcoming in Journal of Official Statistics*.

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

Winship, C. & Morgan, S. L. (1999). The Estimation of Causal Effects from Observational Data. *Annual Review of Sociology*, *25*, 659–706.