Discussion Paper

Adjusting measurement bias in sequential mixed-mode surveys using re-interview data

2015 | 23

Thomas Klausch
Barry Schouten
Bart Buelens
Jan van den Brakel

In mixed-mode surveys, mode-differences in measurement bias, also called measurement effects or mode effects, continue to pose a problem to survey practitioners. In this paper, we discuss statistical adjustment of measurement bias to the level of a measurement benchmark mode during inference from mixed-mode data. In doing so, statistical methodology requires auxiliary information which we suggest to collect in a re-interview administered to a sub-set of respondents to the first stage of a sequential mixed-mode survey. In the re-interview, relevant questions from the main survey are repeated. After introducing the design and presenting relevant statistical theory, this paper evaluates the performance of a set of six candidate estimators that exploit re-interview information in a Monto Carlo simulation. In the simulation, a large number of parameters is systematically varied, which define the size and type of measurement and selection effects between modes in the mixed-mode design. Our results indicate that the performance of the estimators strongly depends on the true measurement error model. However, one estimator, called the inverse regression estimator, performs particularly well under all considered scenarios. Our results suggest that the re-interview method is a useful approach to adjust measurement effects in the presence of non-ignorable selectivity between modes in mixed-mode data.

1 Introduction

Sequential mixed-mode surveys combine multiple modes of data collection in sequential order to maximize on survey response while optimizing on data collection costs (De Leeuw, 2005; Groves et al., 2010; Lynn, 2013). Usually, a sequential design starts with a cost efficient mode (e.g., web data collection) and, subsequently, non-respondents to the first stage are approached by another mode (e.g., face-to-face). This second stage, typically, strongly improves survey response. When a face-to-face follow up is used, for example, the combined mixed-mode design often reaches response rates comparable to those of single-mode face-to-face survey designs, but at lower costs (Klausch, Hox, & Schouten, 2015). Sequential designs are not limited to two modes and in practice many designs use three or even four modes.

The increase in survey response may be an indication for a reduction in survey non-response bias and may lead to more balanced response samples (Klausch, Hox, & Schouten, 2015; Schouten, Cobben, & Bethlehem, 2009). However, any mode has particular measurement error properties, which turn certain modes more or less suitable for the measurement of specific target variables (Klausch, Hox, & Schouten, 2013). It is often noted, for example, that socially desirable answering can introduce systematic measurement error (bias) to estimates of statistics of sensitive characteristics (Kreuter, Presser, & Tourangeau, 2008). This behaviour is typically stronger in interviewer administered than in self-administered modes. A mixed-mode design combining self- and interviewer administration can, therefore, increase the measurement bias of linear estimates, such as means and totals.

Mode differences in random and systematic measurement error, which are also called ‘measurement effects’ (or mode effects), are considered a key problem of mixed-mode

surveys (Jäckle, Roberts, & Lynn, 2010; Klausch et al., 2013). Besides increases in bias of mixed-mode estimates, they can lead to instability of time series, because the relative sizes of mode-specific response samples in repeated mixed-mode surveys often vary over time (Buelens & van den Brakel, 2014). Such deficits may outweigh the increase in response.

The present paper contributes to the growing body of literature that discusses statistical adjustment of measurement effects (Kolenikov & Kennedy, 2014; Suzer Gurtekin, 2013; Vannieuwenhuyze, 2015). Statistical adjustment seeks to convert measurements obtained under different modes to the level of a common measurement benchmark mode (Klausch, Schouten, & Hox, 2015). The measurement benchmark mode is assumed to be the desirable way (combination of mode and question format) to measure a target variable. For example, if the systematic measurement error differences between modes were known for a continuous target variable, the difference in means could be used in a mixed-mode survey as a fixed adjustment to all responses that were not observed under the benchmark mode.

However, the primary difficulty in estimating measurement effects, is a common confounding with so-called selection effects in mixed-mode data (Vannieuwenhuyze, 2014, 2015; Vannieuwenhuyze & Loosveldt, 2013; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010). A selection effect denotes a difference in the true score distributions of a target variable between mode-specific response samples. Practically, this situation suggests that different people participate in different modes. This, frankly, is the objective of any mixed-mode survey and in the absence of selection effects using a sequential mixed-mode survey has only very limited practical advantages.

Disentangling measurement and selection effects requires additional auxiliary data, which are typically unavailable to analysts. Previous literature has often applied relatively weak auxiliary data for estimating measurement effects, such as socio-demographic sampling frame information, leading to potential bias of unknown size in effect estimates (Vannieuwenhuyze, 2015; Vannieuwenhuyze & Loosveldt, 2013). In particular, estimates will be biased when mode-specific non-response does not occur at random in the mixed-mode design (Little & Rubin, 2002), as may be indicated by weak relations of auxiliary information and response mechanisms.

In this article, we employ an innovative approach to this problem using a research design called the mixed-mode re-interview (Klausch, Schouten, & Hox, 2015; Schouten, van den Brakel, Buelens, van der Laan, & Klausch, 2013). In the re-interview, respondents to the first stage of the mixed-mode design are re-approached under a second mode, where relevant questions from the main survey are repeated. This additional information is exploited in estimation. Using a simulation study, we evaluate whether the re-interview is a useful approach for adjusting measurement error bias between survey modes in the important case when selection into modes depends on the target variable and thus occurs not at random.

This paper is structured into two parts. First, we provide a formal framework to describe survey errors and statistical adjustment in mixed-mode surveys and the re-interview design (sections 2 and 3). This theory is novel to the field and needed to describe and solve adjustment problems of the type discussed here. Second, we use a statistical

simulation to evaluate the performance of a total of six adjusted estimators, which exploit re-interview information in different ways (section 4). In the simulation, we systematically vary the size of measurement and selection effects and study how the efficiency of the estimators changes across conditions.

2 The sequential mixed-mode re-interview design

There are, in principle, three ways to address mode differences in measurement error (Schouten et al., 2013). First, the occurrence of effects may be prevented by designing questions and questionnaires to evoke the same and correct answer under all modes (a situation also called measurement equivalence; Dillman, Smyth, and Christian, 2009). If achieving equivalence is not possible a second option may be to avoid problematic modes, such as those modes leading to an increase in measurement error. However, avoiding some modes may not be desirable from a response and a cost perspective. The third approach therefore addresses measurement effects during the estimation stage, where we distinguish two approaches. The first approach uses the so-called calibration method. The analyst accepts that measurement error models of modes may in fact differ and instead focusses on calibrating the size of mode-specific response samples to fixed proportions, thus keeping measurement error of the whole design stable across time (Buelens & van den Brakel, 2014). Therefore it is a strong advantage of this procedure that it helps to stabilize time series data from repeated cross-sectional mixed-mode surveys, when the size of mode-specific response groups differs across time as it is often the case in practice. However, the approach does not adjust the measurement error difference between modes. Statistical adjustment of measurement effects is considered in the present paper.

2.1 Problem proposition

A schematic illustration of the available data from a mixed-mode survey is provided in figure 2.1, i (Klausch, 2014; Klausch, Hox, & Schouten, 2015). For ease of exposition, we focus on a mixed-mode design with two modes (extensions to more modes are addressed in the appendix). We distinguish three types of variables in the mixed-mode survey: first, the 'true' scores of a target variable, Y, second, variable Y as measured by mode 1, Y^{m_1}, and third, variable Y as measured by mode 2, Y^{m_2}.

In figure 2.1, i, we depict the missing data pattern typical to a standard sequential mixed-mode survey, where response (available data) is characterised by white areas and unavailable data is characterised by grey areas. It can be seen that the true scores Y are unobserved, which is the primary motivation to conduct a survey, whereas Y^{m_1} and Y^{m_2} are partly observed, because, in a sequential mixed-mode design, nonrespondents to m_1 are followed up in m_2 resulting in some response under either m_1 (field A) or m_2 (field D). The unobserved outcomes can be called 'potential' (following terminology introduced by Rubin, 2005) denoting the hypothetical events that respondents under m_2

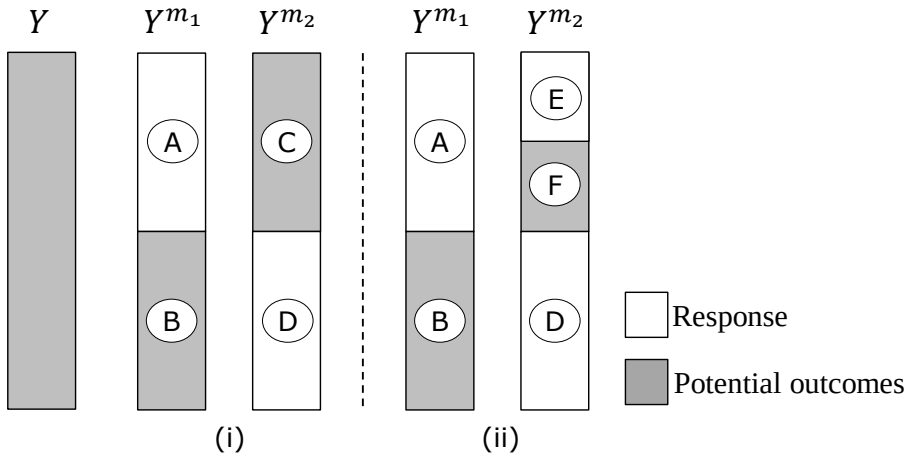

Figure 2.1 ***Schematic illustration of the missing data pattern of two sequential mixed-mode surveys: left (i) a simple sequential design, right (ii) a sequential design with re-interview. The true scores Y are not observed, while respondents in modes 1 and 2 give answers Y^{m_1} (field A) and Y^{m_2} (field D). The additional re-interview data in (ii) create overlap between the m_1 (field A) and m_2 (field E) response distributions.***

had been observed under m_1 (Field B) or, reversely, respondents under m_1 had been observed under m_2 (Field C).

In the following, it is our objective to estimate the true mixed-mode response mean of Y, denoted $\bar{Y}_{r_{mm}}$. This is the true mean of the response sample. It should be noted that an estimator of $\bar{Y}_{r_{mm}}$ may additionally suffer from non-response bias against the true population mean. In this paper, we focus on adjusting measurement error bias of response mean estimators. We refer to literature on non-response adjustment for correcting the remaining non-response bias of the measurement error-adjusted response mean (Bethlehem, 2002; Bethlehem, Cobben, & Schouten, 2011; Cochran, 1977; Little & Rubin, 2002; Särndal & Lundström, 2005; Särndal, Swensson, & Wretman, 1992).

Since, observed variables Y^m are employed in estimation, the measurement error of both variables may bias an unadjusted estimator of $\bar{Y}_{r_{mm}}$ that results when simply pooling the mixed-mode data by taking the sample mean across observed data (Kolenikov & Kennedy, 2014). We seek to find an estimator whose mean squared error is lower compared to the unadjusted estimator.

In doing so, we make an important assumption, called the measurement benchmark assumption. In the absence of true scores Y, it is impossible to correct the measurement error bias contributed by both modes to the estimator. Instead, we focus on the situation when one of the modes is assumed as a measurement benchmark, setting the observed scores of this mode equal to Y. The choice of benchmark mode is made by the analyst and guided by the practical consideration which combination of mode and question format can be assumed to evoke the least or no measurement error.

A major complication in this endeavour is the occurrence of non-random selection

effects between modes, in the sense that the response mean of true scores Y under m_1 is not equivalent to the response mean under m_2. Such effects are desired in mixed-mode surveys, because they reflect that different modes reach different respondents. If both modes reached the same respondents, the second mode in the mixed-mode design would not have any additional value except for increasing the overall response rate. Unfortunately, selection effects are confounded with mode differences in measurement error (measurement effects. A difference in response means on Y^{m_1} and Y^{m_2} thus can denote a measurement or a selection effect or a combination. In a simple sequential design (Figure 2.1, i), the analyst has insufficient information to determine the size of these effects (Vannieuwenhuyze & Loosveldt, 2013). In adjusting measurement error it is therefore necessary to control for selection effects between modes.

2.2 Design and use of a re-interview extension

Any estimator of $\bar{Y}_{r_{mm}}$ that seeks to be superior to an unadjusted estimator necessarily needs to employ additional data in estimation. The re-interview method is one approach for collecting such data. It may be an approach to use register (sampling frame) information for modelling response mechanisms and target variables. We consider these variables weak auxiliary information in many cases and for this reason evaluate the role of re-interview data as potentially stronger auxiliary data.

The re-interview design consists of a standard sequential mixed-mode survey, where in addition a subset of respondents in m_1 is followed up in m_2. The additional information is used during estimation of the mixed-mode response mean. Figure 2.1, ii, illustrates the missing data pattern of this design. It can be seen that the re-interview data create overlap between the partly observed response vectors Y^{m_1} and Y^{m_2}, so that for this subset of respondents the outcomes are observed in both modes (Field E). This overlap is essential in all estimation techniques discussed bellow.

Three further aspects of the re-interview design are worth highlighting. First, the introduction of a re-interview next to an ongoing mixed-mode design does not impact the standard fieldwork of the sequential mixed-mode survey. Since m_1 respondents are only re-interviewed after their Y^{m_1} answers have been recorded, the additional measurement occasion cannot 'bias' the regular measurement process.

Second, the re-interview fieldwork normally suggests additional costs. However, it is not required to approach every m_1 respondent again. Instead, only a smaller sub-set of m_1 respondents needs to be approached for a re-interview. This measure makes sure that the mixed-mode re-interview design still offers the advantage of cost savings compared to a standard design fielded in m_2 alone. Intuitively, the smaller the re-interview sample size will be, the less precise will be an adjustment basing on the re-interview data. In this paper, we focus on evaluating the precision of estimators in the large sample scenario. In future research we will assess the sensitivity of the best estimators across different re-interview sample sizes.

Third, when adding the re-interview measurement to a sequential mixed-mode design, the repeated measurements in m_2 potentially may be influenced by the earlier measurement occasion. In this situation, Y^{m_2} in the re-interview would not follow the

same measurement model as standard responses in m_2. In the present paper, we assume that the measurement error models in the re-interview and the regular m_2 model are identical. We call this assumption *measurement equivalence*.

Measurement equivalence is likely to occur in many practical situations. For in-equivalence to occur, respondents first need to recall answers given at the first occasion (m_1) when the question is repeated under m_2. In addition, respondents need to be motivated to reproduce the answer they recall from the first occasion. The time lag between occasions, which in practice usually lies in the range of several weeks, plays a relevant role, because longer time lags increase chances of answers being forgotten. Even if answers given earlier in m_1 are recalled, it is doubtful whether this causes the re-interview respondent to answer consistently with this answer in the response situation under m_2. Measurement equivalence, nevertheless, is an important assumption for the re-interview method. Whereas we assume equivalence, we point to the need for experimental evaluation of possible in-equivalence between regular m_2 and re-interview measurements in practice. In addition, it may be possible to relax the assumption in further development of the suggested re-interview methodology. We re-emphasize these aspects in the discussion section.

3 Bias and adjustment of the mixed-mode response mean

In this section, we, first, present a statistical model for the data generating process of mixed-mode re-interview surveys. Second, we derive the bias of the unadjusted mixed-mode estimator. Third, we suggest a set of six potentially superior candidate estimators that exploit re-interview data in estimation.

3.1 Selection model

For the simple sequential mixed-mode survey (Figure 2.1, i) we assume a fixed response model that separates all units i in a population of size N for a given mixed-mode design D into two response strata (units participating in either m_1 or m_2) and a non-response stratum (Cochran, 1977). Since we focus on the response mean in this paper (cf. section 2), the non-response stratum is ignored in the following discussion. Let fixed indicator variables r_1 and r_2 identify these groups, so that $N_{r_j} = \sum_i r_{ji}$ is the population size of the response stratum of mode $j = \{1, 2\}$. Let

$$P_j = \frac{N_{r_j}}{(N_{r_1} + N_{r_2})} \tag{1}$$

denote the relative size of the strata, and let $\bar{Y}_{r_j} = N_{r_j}^{-1} \sum r_{ij} y_i$ be the stratum mean, where y_i is the true score of unit i on continuous target variable y. The mixed-mode response mean is then given by

$$\bar{Y}_{r_{mm}} = P_1\bar{Y}_{r_1} + P_2\bar{Y}_{r_2}. \quad (2)$$

The contrast

$$SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = \bar{Y}_{r_1} - \bar{Y}_{r_{mm}} = P_2(\bar{Y}_{r_1} - \bar{Y}_{r_2}) \quad (3)$$

denotes the selection effect of $\bar{Y}_{r_1}$ relative to $\bar{Y}_{r_{mm}}$. It can be seen that the relative selection effect between modes, i.e. $SE(\bar{Y}_{r_1}, \bar{Y}_{r_2}) = \bar{Y}_{r_1} - \bar{Y}_{r_2}$, is dependent on $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$ and if $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) \neq 0$, $SE(\bar{Y}_{r_1}, \bar{Y}_{r_2}) \neq 0$ follows. Relative selection effects between modes are a major motivation for conducting mixed-mode surveys, because, if $\bar{Y}_{r_1} = \bar{Y}_{r_2}$ (or equivalently $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = 0$), the second stage survey in mode m_2 does not contribute to a reduction of selection bias of mixed-mode estimators against the population mean and thus would not be needed.

When a re-interview is added to the sequential mixed-mode design (Figure 2.1, ii), the r_1 response stratum is split up into a re-interview response (field E) and re-interview non-response (field F) stratum. Let r_{re} denote the response indicator indicating whether a respondent in m_1 responds in the re-interview (if assigned) and let $P_{re} = N_{re}/N_{r_1}$ denote the size of the re-interview response sample relative to the number of m_1 respondents and let $P_{nre} = 1 - P_{re}$. Furthermore, let $\bar{Y}_{r_{re}}$ indicate the response mean of this stratum and $\bar{Y}_{r_{nre}}$ the non-response mean. The contrast

$$SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) = \bar{Y}_{r_{re}} - \bar{Y}_{r_1} = P_{nre}(\bar{Y}_{r_{re}} - \bar{Y}_{r_{nre}}) \quad (4)$$

denotes the re-interview selection effect, which occurs in the situation when systematically different respondents participate in the re-interview than in m_1. Such effects have practical relevance, because response in the re-interview may be selective relative to m_1 since a different mode is offered (m_2) and for reasons of response burden due to repeated participation. In the present paper, we explicitly allow for the possibility of re-interview selection effects by including them in the simulation study. As we show after introducing the candidate estimators, the size of re-interview selection effect has a direct implication for the bias of the estimators.

Under the fixed-response model, the population response distribution of Y is a mixture distribution of the stratum distributions from the sets defined by $r_1 = 1$ and $r_2 = 1$, where r_1 consists of two sub-strata defined by $r_{re} = 1$ and $r_{re} = 0$. Let σ_Y^2 denote the population response variance of Y and let $\sigma^2_{Y_{r_{re}}}$, $\sigma^2_{Y_{r_{nre}}}$, and $\sigma^2_{Y_{r_2}}$ denote the population variances within the strata, respectively. Furthermore, let $\sigma^2_{Y_r} = P_1(P_{re}\sigma^2_{Y_{r_{re}}} + P_{nre}\sigma^2_{Y_{r_{nre}}}) + P_2\sigma^2_{Y_{r_2}}$ give the within-stratum variance pooled across strata. It follows that the total variance σ_Y^2 is equal to the sum of between-stratum and pooled within-stratum variances, where the between-stratum variance is determined by the size of selection effects:

$$\begin{aligned}\sigma_Y^2 &= \sigma_{Y_r}^2 + P_1\big(P_{re}(\bar{Y}_{r_{re}} - \bar{Y}_{r_{mm}})^2 + P_{nre}(\bar{Y}_{r_{nre}} - \bar{Y}_{r_{mm}})^2\big) + P_2(\bar{Y}_{r_2} - \bar{Y}_{r_{mm}})^2 \\ &= \sigma_{Y_r}^2 + P_1\big(\frac{P_{re}}{P_{nre}} SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})^2 + SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})^2\big) + P_2\big(-\frac{P_1}{P_2} SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})^2\big) \\ &= \sigma_{Y_r}^2 + \frac{P_1 P_{re}}{P_{nre}} SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})^2. \end{aligned} \quad (5)$$

3.2 Measurement model

Assume that each mode is associated with a question-specific measurement error model that describes the relation of true scores y_i to the observed outcomes in m_j, denoted $y_i^{m_j}$, as (Alwin, 2007; Biemer & Stokes, 1991; Lord & Norvick, 1968)

$$y_i^{m_j} = \mu^{m_j} + \lambda^{m_j}(y_i + u_i^{m_j}) \quad \forall \quad i, \quad (6)$$

where λ^{m_j} is a scale parameter that is equal to 1 if m_j measures on the scale of the true score, and $u_i^{m_j}$ is an independently and identically distributed measurement error term with

$$u_i^{m_j} \sim iid(0, (\sigma_u^{m_j})^2) \quad \forall \quad i. \quad (7)$$

μ^{m_j} is referred to as systematic measurement error common to all units, whereas $(\sigma_u^{m_j})^2$ denotes the variance of measurement errors in the population and is called the random measurement error component. We assume independence of true scores y_i and measurement errors $u_i^{m_j}$ for all i and j.

Let $c_j = cor(Y^{m_j}, Y)$ be the population correlation between Y^{m_j} and Y. Random error $(\sigma_u^{m_j})^2$ is dependent on σ_Y^2 and c_j:

$$(\sigma_u^{m_j})^2 = \frac{1 - c_j^2}{c_j^2}\sigma_Y^2. \quad (8)$$

c_j is sometimes referred to as validity or reliability coefficient (Biemer & Stokes, 1991). In the simulation we use different levels of c_j to scale the error variance of Y^m, without specifically arguing on the source of random error.

When adjusting measurement bias, the analyst chooses a measurement benchmark mode (cf. section 2). We denote the chosen benchmark by letter b and denote the alternative mode as focal mode j. Henceforth, b may be chosen as $b = 1$ (thus $j = 2$), in which case the first mode (m_1) in the sequential design is the measurement benchmark (e.g., web) and the second mode (e.g., face-to-face) is the focal mode, or $b = 2$ (thus

$j = 1$) denoting the reverse situation. For the measurement benchmark mode it is assumed that

$$y_i^{m_b} = \mu^{m_b} + \lambda^{m_b}(y_i + u_i^{m_b}) = y_i \quad \forall \quad i \tag{9}$$

implying $\mu^{m_b} = 0$, $\lambda^{m_b} = 1$, and $u_i^{m_b} = 0$ for all i (thus $c_b = 1$).

Furthermore, we assume measurement equivalence between re-interview and m_2 (cf. section 2). Let $y_i^{m_{re}}$ denote potential re-interview outcomes. The measurement equivalence assumption states that

$$y_i^{m_{re}} = y_i^{m_2} \quad \forall \quad i. \tag{10}$$

3.3 Bias of unadjusted mean estimators

Consider now the unadjusted estimator of the mixed-mode response mean which simply pools the observed mixed-mode data without applying any correction in the estimator,

$$\hat{\bar{Y}}_{r_{mm}}^{unadj} = \frac{1}{\hat{N}_{r_1} + \hat{N}_{r_2}} \sum_{i=1}^{N} I_i d_i \left(r_{1i} y_i^{m_1} + r_{2i} y_i^{m_2}\right), \tag{11}$$

where d_i denotes design weights determined by the sampling design D as inverse of inclusion probability of unit i and I denotes the indicator for the outcome of random sampling, where $E_D(I_i) = d_i^{-1}$, and $\hat{N}_{r_j} = \sum I_i d_i r_{ij}$. The bias of the unadjusted mean estimator over sampling design D and measurement model M is given by

$$\begin{aligned} B(\hat{\bar{Y}}_{r_{mm}}^{unadj}) &= P_1((\lambda^{m_1} - 1)\bar{Y}_{r_1} + \mu^{m_1}) + P_2((\lambda^{m_2} - 1)\bar{Y}_{r_2} + \mu^{m_2}) \\ &= P_1 B_{me}^{m_1} + P_2 B_{me}^{m_2}, \end{aligned} \tag{12}$$

where $B_{me}^{m_1}$ and $B_{me}^{m_2}$ denote measurement bias contributed by modes 1 and 2. If m_1 or m_2 represent a measurement benchmark, one of the measurement bias terms is zero. If $b = 1$, it follows $B(\hat{\bar{Y}}_{r_{mm}}) = P_2 B_{me}^{m_2}$, and $B(\hat{\bar{Y}}_{r_{mm}}) = P_1 B_{me}^{m_1}$ in the reverse situation ($b = 2$).

It could be argued that an alternative unadjusted estimator uses only m_1 responses, but not the follow up, for estimating $\bar{Y}_{r_{mm}}$. This estimator pictures the situation when a single-mode survey would be used instead of the mixed-mode design, for example, to save the costs for administering m_2. The estimator is

$$\hat{\bar{Y}}_{r_{mm}}^{unadj2} = \frac{1}{\hat{N}_{r_1}} \sum_{i=1}^{N} I_i d_i r_{1i} y_i^{m_1} \tag{13}$$

and it has bias

$$B(\hat{\bar{Y}}_{r_{mm}}^{unadj2}) = (\lambda^{m_1} - 1)\bar{Y}_{r_1} + \mu^{m_1} + SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = B_{me}^{m_1} + SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}). \tag{14}$$

It can be seen that when m_1 is the benchmark the estimator suffers from a selection effect against $\bar{Y}_{r_{mm}}$, and when the follow up mode m_2 is the benchmark, it suffers from additional measurement bias $B_{me}^{m_1}$.

3.4 Candidate estimators using re-interview data

In the following, we employ the auxiliary information collected in the re-interview in a set of six adjusted candidate mean estimators. Let indicator $s_{re,i}$ determine whether unit i is selected for a re-interview. Furthermore, let $P_s = \sum s_{re,i} / \sum r_1$ denote the proportion of re-interviewed respondents. $P_s = 1$ denotes the situation when all m_1 respondents are approached for a re-interview, but, as we noted, in practice choices of $P_s < 1$ make the design cost efficient. Depending on which mode represents the measurement benchmark, we now seek to estimate the mean of either Y^{m_1} or Y^{m_2} over the full response sample. This objective presents us with the missing data problem illustrated in figure 2.1, ii. In doing so, we employ auxiliary data obtained from the sub-set of re-interview respondents, that is all i for whom $I_i s_{re,i} r_{re,i} = 1$ holds.

We consider two classes of estimators referred to as π-estimators and y-estimators, respectively (Kang & Schafer, 2007; Särndal & Lundström, 2005). Further estimators, including combined $\pi-$ and $y-$ estimators ('double-robust'), have been suggested in the literature but we do not discuss them in the present paper (Bang & Robins, 2005). π-estimators estimate the propensity of respondents to reply under the benchmark mode and apply it for calibrating a selective sub-group (i.e. response sample in benchmark mode) to a reference group (i.e. the mixed-mode response sample). Let the propensity for unit i to be observed in the benchmark mode be denoted as π_i (Rosenbaum & Rubin, 1983, 1985). The propensity can be estimated using a model f_π:

$$\begin{aligned} \pi_i &= P(r_{ib} = 1 | y_i^{m_j}, r_{ij} = 1 \cup r_{i,re} = 1) \\ &= f_\pi(y_i^{m_j}, r_{ij}, r_{i,re} = 1, \theta) \quad \forall \quad i \; ; \; b, j = 1, 2 \; ; \; b \neq j. \end{aligned} \tag{15}$$

A typical choice for function f_π is the logistic or probit regression model with unknown parameter vector θ. The propensity score can be applied in a variety of ways in estimation, such as matching, stratification or weighting. In this paper, we examine performance of a weighting estimator using the inverse propensity:

$$\hat{\hat{Y}}_{r_{mm}}^{\pi} = \frac{1}{\hat{N}_1 + \hat{N}_2} \sum_{i=1}^{N} I_i d_i y_i^{m_b} \left(r_{i,re} s_{re,i} \frac{(1 - \hat{\pi}_i)}{\hat{\pi}_i} + 1 \right) \quad ; \; b = 1, 2. \tag{16}$$

This estimator first estimates the total of the observed benchmark outcomes $y_i^{m_b}$ from the response in benchmark mode b. It then adds an estimate of the total of benchmark outcomes in the focal mode using benchmark outcomes from the re-interview response sample calibrated by weight $(1 - \hat{\pi}_i)/\hat{\pi}_i$.

On the other hand, y-estimators seek to find accurate predictions of the potential benchmark outcomes y^{m_b} using a suitable model for y^{m_b} and finally sum over the joint vector of observed and predicted scores (Schafer & Kang, 2008). A general form of y-estimator can be written as:

$$\hat{\hat{Y}}_{r_{mm}}^{yest} = \frac{1}{\hat{N}_1 + \hat{N}_2} \sum_{i=1}^{N} I_i d_i \left(r_{ib} y_i^{m_b} + r_{ij} \hat{y}_i^{m_b} \right) \quad ; \quad b, j = 1, 2 \; ; \; b \neq j, \tag{17}$$

where $\hat{y}_i^{m_b}$ represent the estimated potential (unobserved) benchmark outcomes for respondents in the focal mode j. The y-estimator requires specifying a y-model that describes the relation of benchmark to alternative mode outcomes. It is then assumed that the model also holds in the response stratum to mode j and can be used to transform observed y^{m_j} to y^{m_b}. Three simple y-models lead to the fixed-effect estimator

$$\hat{\hat{Y}}_{r_{mm}}^{fe} = \frac{1}{\hat{N}_1 + \hat{N}_2} \sum_{i=1}^{N} I_i d_i \left(r_{ib} y_i^{m_b} + r_{ij} \left(y_i^{m_j} - (\hat{\hat{Y}}_{re}^{m_j} - \hat{\hat{Y}}_{re}^{m_b}) \right) \right) \quad ; \quad b, j = 1, 2 \; ; \; b \neq j, \tag{18}$$

where

$$\hat{\hat{Y}}_{re}^{m_j} = \frac{1}{N} \sum_{i=1}^{N} I_i d_i r_{re,i} s_{re,i} y_i^{m_j} \tag{19}$$

is the sample mean of the re-interview stratum (and $\hat{\hat{Y}}_{re}^{m_b}$ defined analogously), the ratio estimator

$$\hat{\hat{Y}}_{r_{mm}}^{ratio} = \frac{1}{\hat{N}_1 + \hat{N}_2} \sum_{i=1}^{N} I_i d_i \left(r_{ib} y_i^{m_b} + r_{ij} y_i^{m_j} \frac{\hat{\hat{Y}}_{re}^{m_b}}{\hat{\hat{Y}}_{re}^{m_j}} \right) \quad ; \quad b, j = 1, 2 \; ; \; b \neq j, \tag{20}$$

and the regression (GREG) estimator

$$\hat{\bar{Y}}_{r_{mm}}^{greg} = \frac{1}{\hat{N}_1 + \hat{N}_2} \sum_{i=1}^{N} I_i d_i \left(r_{ib} y_i^{m_b} + r_{ij} \left(\hat{\bar{Y}}_{re}^{m_b} - \hat{\beta}_{re} (\hat{\bar{Y}}_{re}^{m_j} - y_i^{m_j}) \right) \right) \quad ; \quad b, j = 1, 2 \; ; \; b \neq j, \tag{21}$$

where β_{re} denotes the (population) 'slope' of the linear regression of Y^{m_b} on Y^{m_j} in the re-interview stratum, given by the ratio of population covariance of Y^{m_b} and Y^{m_j} over the variance of Y^{m_j} in the re-interview (Bethlehem, 1988; Särndal & Lundström, 2005).

It can be seen that the fixed-effect estimator is a special case of the GREG estimator, where the slope is equal to 1. Furthermore, it can be shown that the ratio estimator is the special case of the GREG estimator, where the intercept is fixed at zero. In practice, the parameters β_{re} as well as $\bar{Y}_{r_{re}}^{m_j}$ and $\bar{Y}_{r_{re}}^{m_b}$, are estimated by their sample analogues in the re-interview response stratum.

The IPW, regression and ratio estimators are standard approaches of survey statisticians for estimation with missing outcomes. The fixed-effect estimator is useful, because the contrast between $\bar{Y}_{r_{re}}^{m_j}$ and $\bar{Y}_{r_{re}}^{m_b}$ may form a good estimator of measurement bias μ^{m_j}. In addition to the standard estimators, we consider two further estimators in this paper. First, we use an inverse version of a regression estimator (IREG). The idea of IREG is to use benchmark measurements Y^{m_b} instead of Y^{m_j} as auxiliary data for modelling focal mode outcomes Y^{m_j} in

$$y_i^{m_j} = \nu_0 + \nu_{re} y_i^{m_b} + \epsilon_i, \tag{22}$$

estimated by ordinary least squares from re-interview response, and then use the inverse of ν_{re} in a GREG-type estimator leading to

$$\hat{\bar{Y}}_{r_{mm}}^{ireg} = \frac{1}{\hat{N}_1 + \hat{N}_2} \sum_{i=1}^{N} I_i d_i \left(r_{ib} y_i^{m_b} + r_{ij} \left(\hat{\bar{Y}}_{re}^{m_b} - \frac{1}{\nu_{re}} (\hat{\bar{Y}}_{re}^{m_j} - y_i^{m_j}) \right) \right) \quad ; \quad b, j = 1, 2 \; ; \; b \neq j. \tag{23}$$

In the next section, we explain why IREG may be a superior estimator to the standard approaches.

Finally, we consider an alternative approach to estimation with missing data using simultaneous multiple imputation (Rubin, 1987). The procedure creates several data sets, in which all missing data are replaced by plausible values. Variance in plausible values for a given unit across imputed data sets reflects the uncertainty in the imputed value. In doing so, we evaluate the performance of sequential regression imputation (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001) using an algorithm called "MICE" with five multiply imputed data sets (multiple imputation by chained equations; van Buuren, 2012; van Buuren, Brand, Groothuis-Oudshoorn, and Rubin, 2006). This algorithm specifies two regression models $P(Y^{m_1}|Y^{m_2}, \theta_1)$ and $P(Y^{m_2}|Y^{m_1}, \theta_2)$.

Starting with the conditional distribution featuring less missing values, a regression model is fit and a draw from the posterior distribution of parameter estimates taken. Based on this draw a predictive model is used to generate imputes. The completed vector of Y^m is then used as a predictor for the second regression model. The procedure is iterative and stops when plausible imputes do not change anymore. It can be shown the method is an approximate Gibbs sampler for the bivariate distribution $P(Y^{m_1}, Y^{m_2})$ (Raghunathan et al., 2001). Depending on benchmark mode choice, $\bar{Y}_{r_{mm}}$ is estimated by taking the sample mean across the completed vector of Y^{m_b} for each imputed data set and then pooling the estimates according to Rubin's rules (Rubin, 1987).

3.5 Bias of standard estimators

Usually, exogenous sampling frame information is available for statistical inference using the standard estimators. This data is assumed to be available for all population units and considered free of random measurement error. Both properties cannot be said to hold in the re-interview design, however. As argued above, response in the re-interview sub-sample may be selective (equation 4). Furthermore, observations in the focal mode may be subject to random error u^{m_j} (equation 6) as well as scaling parameter $\lambda^{m_j} \neq 1$. Since it is unclear, which effect these deviations from standard assumptions have, it is instructive to consider the expected bias of the standard estimators (cf. proof given in the supplemental material),

$$B(\hat{\bar{Y}}^{fe}_{r_{mm}}) = P_j\big((1-\lambda^{m_j})(\bar{Y}_{r_{re}} - \bar{Y}_{r_j})\big), \tag{24}$$

$$B(\hat{\bar{Y}}^{ratio}_{r_{mm}}) \approx P_j\left(\mu^{m_j}\frac{\bar{Y}_{r_{re}} - \bar{Y}_{r_j}}{\lambda^{m_j}\bar{Y}_{r_{re}} + \mu^{m_j}}\right), \tag{25}$$

and

$$B(\hat{\bar{Y}}^{greg}_{r_{mm}}) \approx P_j\Big((1-\lambda^{m_j}\beta_{re})(\bar{Y}_{r_{re}} - \bar{Y}_{r_j})\Big), \tag{26}$$

where

$$\beta_{re} = \frac{\sigma_Y^2}{\lambda^{m_j}(\sigma_Y^2 + (\sigma_u^{m_j})^2)}. \tag{27}$$

Note that the bias of the ratio and GREG estimator are approximated using Taylor linearization (Särndal et al., 1992), where the remainder terms vanish in large samples.

From equations (24) to (26), it can be seen that all biases depend on the contrast $(\bar{Y}_{r_{re}} - \bar{Y}_{r_j})$ which denotes a selection effect between the re-interview and the focal

mode. If m_1 is the benchmark, $(\bar{Y}_{r_{re}} - \bar{Y}_{r_2})$ describes the degree to which m_2 respondents in the re-interview differ from m_2 respondents in the sequential design. If m_2 is the benchmark, $(\bar{Y}_{r_{re}} - \bar{Y}_{r_1})$ describes the degree to which m_2 response in the re-interview is selective relative to m_1 response (i.e. a re-interview selection effect, equation 4). In practical situations it is likely that both contrasts are non-zero introducing bias in all standard y-estimators.

From equation (24) it can be seen that $\hat{\bar{Y}}^{fe}_{r_{mm}}$ is unbiased if $\lambda^{m_j} = 1$, but may be biased in other cases. Put differently, the fixed-effect estimator does correct a systematic error difference between modes, if there is no scale difference λ^{m_j}. Furthermore, it can be seen that the bias $\hat{\bar{Y}}^{fe}_{r_{mm}}$ is equivalent to the GREG estimator, when $\beta_{re} = 1$, which could be expected from the definition of the estimators.

From equation (25) it follows that the bias of $\hat{\bar{Y}}^{ratio}_{r_{mm}}$ is determined by factor μ^{m_j}. The ratio estimator thus corrects a scale difference between modes, if there is no systematic error μ^{m_j}. Therefore it represents a counter-part to $\hat{\bar{Y}}^{fe}_{r_{mm}}$, which is unbiased in the reverse situation.

From equation (26) it can be seen that $\hat{\bar{Y}}^{greg}_{r_{mm}}$ is approximately unbiased if $\lambda^{m_j}\beta_{re} = 1$, and thus

$$\frac{\sigma_Y^2}{\sigma_Y^2 + (\sigma_u^{m_j})^2} = 1. \tag{28}$$

We note that the bias of GREG does not depend on λ^{m_j} and is determined by the size of random error. It is thus only negligible when the focal mode does not measure with random error. This conjecture points to a somewhat surprising shortcoming of the GREG estimator, which is biased in most practical scenarios of the re-interview design, because the focal mode usually measures with error.

To deal with this problem, we introduced the inverse regression estimator (IREG) in the previous section. It can be seen that its bias is

$$B(\hat{\bar{Y}}^{ireg}_{r_{mm}}) \approx P_j\Big((1 - \lambda^{m_j} v_{re}^{-1})(\bar{Y}_{r_{re}} - \bar{Y}_{r_j})\Big), \tag{29}$$

where $v_{re}^{-1} = (\lambda^{m_j})^{-1}$ and, consequently, $\hat{\bar{Y}}^{ireg}_{r_{mm}}$ is approximately unbiased.

Whereas we can take the bias properties of four of the considered candidate estimators from these equations, we will focus on the simulated trade-off of bias and variance (mean squared error) in the study presented in the next section.

4 Simulation study

In the present section, we explain, first, the parametrization of the model and the simulation conditions and, second, the results of the simulation.

4.1 Simulation set-up

In practice, the distributions of Y and the selection and measurement model parameters are unknown. It was therefore the goal of this study to assess the potential effects that different choices for the parameters have on the RMSE of the unadjusted and adjusted estimators by Monte Carlo Simulation.

Tables 4.1 and 4.2 give an overview on the parametrization of the selection and measurement model, respectively. We distinguish between three types of parameters. Fixed parameters were not varied over simulation conditions. Free parameters represented the simulation conditions, where the exact values applied in the simulation can be taken from the tables and are detailed below. Dependent parameters depended on the parametrization of the fixed and free parameters.

Table 4.1 ***Parameterization of the super-population selection model in the simulation***

Parameter	Eq.	Value(s) in simulation	Description
Fixed:			
$\bar{Y}_{r_{mm}}$	(2)	1	Mixed-mode response mean
$\sigma^2_{Y_r}$	(5)	1	Pooled within-stratum variance
P_1	(1)	0.5	Proportion of m_1 resp.
P_{re}	(4)	0.6	Proportion of re-int. resp.
Free:			
$SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$	(3)	$\{-0.5, -0.25, 0, 0.25, 0.5\}$	Selection effect of mode 1
$SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$	(4)	$\{0, 0.5\}$	Re-interview selection effect
Dependent:			
$\bar{Y}_{r_1}$	(3)	Dep. on $\bar{Y}_{r_{mm}}, SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$	Response stratum mean of m_1
$\bar{Y}_{r_2}$	(3)	Dep. on $\bar{Y}_{r_{mm}}, \bar{Y}_{r_1}, P_1$	Response stratum mean of m_2
$\bar{Y}_{r_{re}}$	(4)	Dep. on $\bar{Y}_{r_1}, SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$	Response stratum mean of re-int.
$\bar{Y}_{r_{nre}}$	(4)	Dep. on $\bar{Y}_{r_1}, \bar{Y}_{r_{re}}, P_{re}$	Non-resp. stratum mean of re-int.
P_2	(1)	0.5, dep. on P_1	Proportion of m_2 respondents
P_{nre}	(4)	0.4, dep. on P_{re}	Proportion of re-int. non-resp.
σ^2_Y	(5)	Dep. on $\sigma^2_{Y_r}, SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$	Population variance of Y

It can be seen that a fully parametrized selection model has six degrees of freedom. We fixed the values for $\bar{Y}_{r_{mm}} = 1$,$P_1 = 0.5$, $P_{re} = 0.6$, and $\sigma^2_{Y_r} = 1$, respectively. In earlier experimental research the chosen response proportion for P_1 and P_{re} were found for a web - face-to-face mixed-mode re-interview design (Klausch, Hox, & Schouten, 2015; Schouten et al., 2013). For other designs different response proportions may be expected, but we do not expect our results to vary strongly across choices of these parameters.

The strength of selectivity between modes, $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$, and the strength of re-interview selectivity, $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$, represented the free parameters in the selection model (equations 3 and 4). These parameters were varied from absent (0%) to strong selectivity ($\pm 50\%$ relative effect to $\bar{Y}_{r_{mm}} = 1$). Within each response stratum true scores y_i were generated from a Gaussian super-population, with means $\bar{Y}_{r_{re}}$, $\bar{Y}_{r_{nre}}$ and $\bar{Y}_{r_2}$ and variance $\sigma^2_{Y_{r_j}} = 1$, respectively. As explained in section 3.1, the resulting population response distribution of Y is a mixture of Gaussian distributions, with $\bar{Y}_{r_{mm}} = 1$ and population variance σ^2_Y, where σ^2_Y is given in equation (5).

Table 4.2 ***Parametrization of the super-population measurement model in the simulation***

Parameter	Eq.	Value(s) in simulation	Description
Fixed:			
μ^{m_b}	(9)	0	Benchmark mode systematic error
λ^{m_b}	(9)	1	Benchmark mode scale parameter
$(\sigma_u^{m_b})^2$	(9)	0	Benchmark mode error variance
Free:			
b	(9)	$\{1, 2\}$	Benchmark mode, $b \neq j$
μ^{m_j}	(6)	$\{-0.3, 0, 0.3\}$	Focal mode systematic error
λ^{m_j}	(6)	$\{0.75, 1, 1.25\}$	Focal mode scale parameter
c_j	(8)	$\{0.1, 0.2, ..., 1\}$	True-observed score correlation
Dependent:			
j	(6)	{1,2}, dep. on b	Focal mode, $b \neq j$
$(\sigma_u^{m_j})^2$	(8)	Dep. on c_j and σ^2_Y	Focal mode error variance

The measurement model has seven degrees of freedom (Table 4.2), where by benchmark mode assumption the parameters of the benchmark measurement model were fixed ($\mu^{m_b} = 0$, $\lambda^{m_b} = 1$, and $(\sigma_u^{m_b})^2 = 0$). The parameters of the focal mode were varied (equation 6): we introduced either no systematic error, $\mu^{m_j} = 0$, or moderate measurement error bias ($\pm 30\%$ relative to the population mean). Furthermore, scaling parameter λ^{m_j} was varied for moderate scale differences, scaling y^{m_j} up ($\lambda^{m_j} = 1.25$) and down ($\lambda^{m_j} = 0.75$). An important element of the focal mode is random error $(\sigma_u^{m_j})^2$, introduced by true-observed score correlation c_j, see equation (8), where c_j was varied between 0.1 (very high error variance) and 1 (no error variance).

A full factorial design was applied across the free parameters, giving rise to $5 * 2 * 2 * 3 * 3 * 10 = 1800$ separate super-population conditions. For each of the conditions, a population of size $N = 100,000$ was generated as a single realization from the super-population. From the generated populations, $K = 1000$ repeated simple random samples with expected size $n_{sample} = 2500$ were drawn without replacement. Every second m_1 respondent was randomly selected for a re-interview ($P_s = 0.5$). This proportion can be expected to yield, on average, a moderate re-interview sample size (expected reinterview $n_{re} = (P_1)(P_{re})(P_s)n_{sample} = 375$).

Subsequently, all estimators discussed under section 3 were administered for each sample. Besides the six adjusted estimators, we computed the unadjusted mixed-mode estimator (equation 11) and the unadjusted ”mode 1 only” estimator (equation 13). We

then estimated the root MSE (RMSE) across the $K = 1000$ repeated samples for any estimator $\hat{\bar{Y}}_{r_{mm}}$ as

$$\hat{RMSE}(\hat{\bar{Y}}_{r_{mm}}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\bar{Y}}_{r_{mm},k} - \bar{Y}_{r_{mm}}^{pop})^2}, \tag{30}$$

where $\bar{Y}_{r_{mm}}^{pop}$ the true population mean for the given condition. Since, as chosen above, $\bar{Y}_{r_{mm}}^{pop} \approx 1 = \bar{Y}_{r_{mm}}$ for any condition, the estimated RMSE also has the interpretation of approximated *relative* $RMSE$ $(= \hat{RMSE}(\hat{\bar{Y}}_{r_{mm}})/\bar{Y}_{r_{mm}})$.

4.2 Results

Figures 4.1 to 4.4 illustrate the key results of the simulation. Each figure displays the estimated RMSE of the two unadjusted and the six adjusted estimators for three levels of λ^{m_j} against the correlation between Y^{m_1} and Y^{m_2} observed in the re-interview ('re-interview correlation'). This correlation can be compared to the analogously estimable quantity in practice, as point of orientation. In the simulation the re-interview correlation is primarily impacted by the size of random error in the focal mode, which is a function of the population correlation c_j (equation 8) varied systematically from 0.1 to 1 (Table 4.2).

All figures display the condition where systematic measurement error was set to $+30\%$ ($\mu^{m_j} = 0.30$) of the mixed-mode response mean ($\bar{Y}_{r_{mm}} = 1$). We provide the figures for the 0% and -30% conditions in the *supplemental material*. The results presented here for $+30\%$ held in large parts for these conditions, so that this choice avoids redundancy. We highlight the few exceptions in the discussion below.

Furthermore, we focus the following discussion on RMSE, but we provide separate plots of the bias and variance components of RMSE in the *supplemental material*. Considering the variance plots, in particular, it can be seen that in most scenarios the variance component of RMSE only plays a dominant role for small to moderate re-interview correlations. For high correlations the dominant component of RMSE turns out to be bias. This is an important result, which however may be impacted by the size of the re-interview sample ($n_{re} = 375$ in the present study). We return to this aspect in the discussion. For brevity, we focus on RMSE in the following and point to the bias and variance components where necessary.

4.2.1 RMSE when mode 1 is the benchmark

Consider first figures 4.1 and 4.2 which display the situation when m_1 is the benchmark ($b = 1$). Figure 4.1 shows the condition when a re-interview selection effect (SE) was introduced (+50% of $\bar{Y}_{m_1}$), whereas it was absent (0%) in the results shown in figure 4.2. Each separate line in the figures represents a different relative SE between m_1 and m_2 ($SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$, cf. equation 3) varying from -50% to $+50\%$ of the mixed-mode

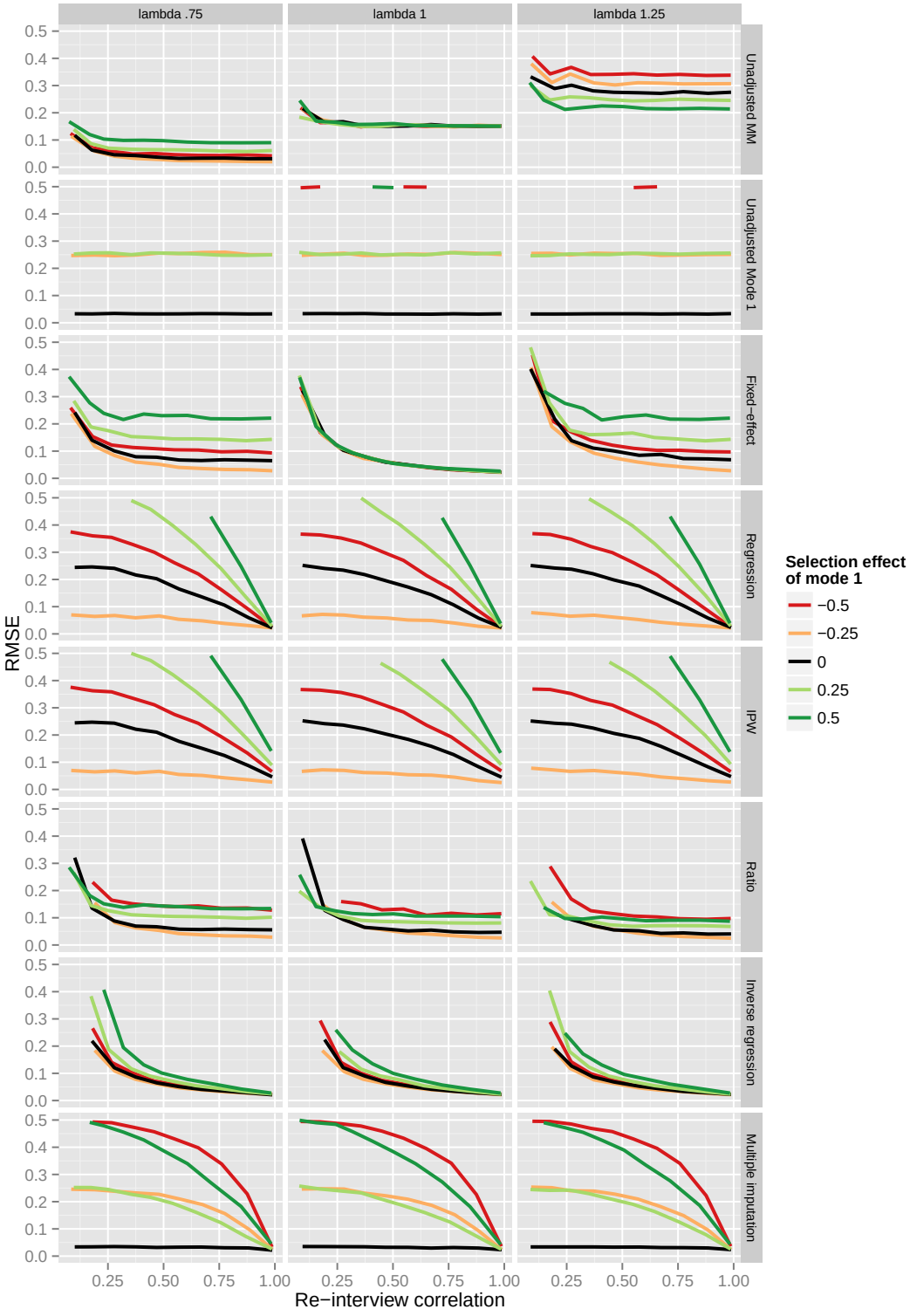

Figure 4.1 ***RMSE of adjusted and unadjusted estimators for benchmark mode*** $b = 1$***, meas. bias*** $\mu^{m_2} = 0.30$***, and re-interview*** $SE(\bar{Y}_{r_{re}}, \bar{Y}_{m_1}) = 0.50$***. IREG performs best under all conditions, followed closely by ratio, which is however biased more strongly. Fixed-effect performs well under*** $\lambda^{m_j} = 1$ ***only.***

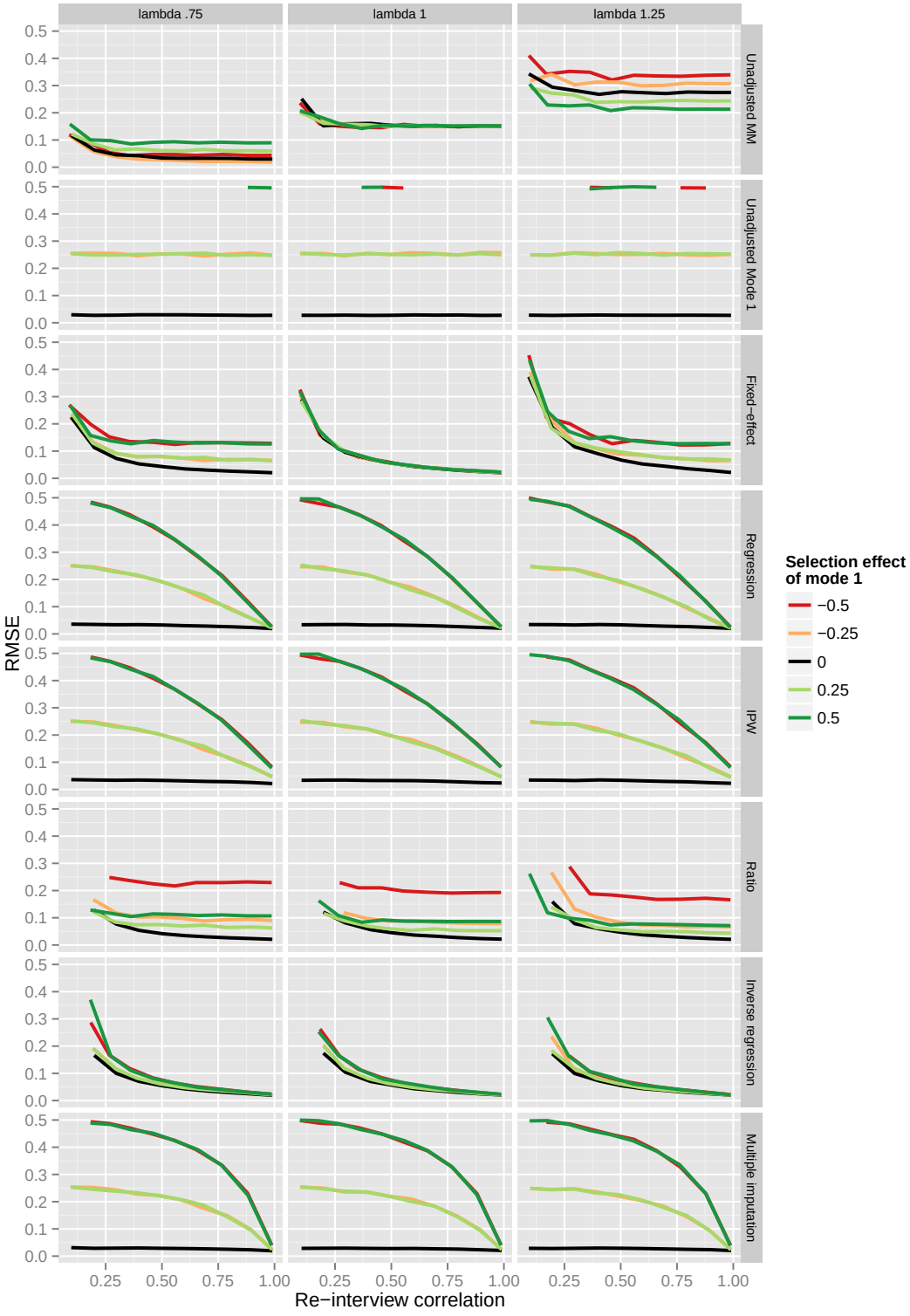

Figure 4.2 ***RMSE of adjusted and unadjusted estimators for benchmark mode*** $b = 1$***, meas. bias*** $\mu^{m_2} = 0.30$***, and re-interview*** $SE(\bar{Y}_{r_{re}}, \bar{Y}_{m_1}) = 0$***. IREG performs best but ratio shows high RMSE. Fixed-effect performs insufficiently and can only reduce RMSE fully when*** $\lambda^{m_j} = 1$***.***

response mean. For clarity, we limit the vertical axis to .50 (equivalent to 50% relative RMSE), so that higher RMSE is not displayed.

In interpreting the further results, we have to compare the performance of the adjusted to the unadjusted mixed-mode estimator (equation 11). Considering the unadjusted mixed-mode estimator, we see that the baseline RMSE, which we sought to reduce by adjustment, varied considerably across conditions of λ^{m_j} and $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$. Whereas it was at a constant level of its bias at $B(\bar{Y}_{r_{mm}}^{unadj}) = P_j \mu^{m_j} = 0.5 * 0.3 = 15\%$ if $\lambda^{m_j} = 1$, it was considerably higher, up to 30 to 40%, for $\lambda^{m_j} > 1$ and lower, 1 to 15%, for $\lambda^{m_j} < 1$. In addition, we can verify that using m_1 only to estimate $\bar{Y}^{r_{mm}}$ ('Unadjusted Mode 1' estimator, equation 13) was insufficient and lead to large RMSE (dominated again by its bias term).

Across the six adjusted estimators, we identified the IREG estimator to outperform all other estimators when mode 1 is the benchmark ($b = 1$). Whereas the estimator was unbiased (equation 29), its variance component could, however, be considerable when focal mode random error was high, as indicated by low re-interview correlation. However, RMSE fell below 10% for a re-interview $cor > .50$ and below 5% for $cor > .70$ (Figure 4.1). In the absence of re-interview sample selectivity, these values even improved slightly (Figure 4.2). Whereas the unadjusted mixed-mode estimator could outperform IREG when $\lambda^{m_j} = 0.75$, the strength of IREG lay in its reliable performance under all conditions.

Furthermore, we notice that all other adjusted estimators performed worse under at least some conditions. This held for the regression, IPW and multiple imputation (MI) estimator in particular. These estimators showed high ($> 10\%$) RMSE unless re-interview correlation was very high ($> .90$), a situation seldom expectable in practice.

For the fixed-effect estimator we notice its good performance if $\lambda^{m_j} = 1$. This finding was expectable as bias vanishes under this condition (equation 24), so that the remainder in both figures is a variance component. However, considering $\lambda^{m_j} \neq 1$ we found that the estimator can suffer from serious error (see in particular Figure 4.1).

A similar observation can be made for the ratio estimator, which suffered from moderate RMSE when re-interview selectivity is 50% (Figure 4.1), but RMSE increased drastically for 0% (Figure 4.2). A finding not shown here, but available in the supplemental material, was that the ratio estimator performed slightly better for all conditions of λ^{m_j} in the absence of systematic term μ^{m_j}, since its bias then vanishes (equation 25). However, we then found the ratio estimator to have higher variance (i.e. if $\mu^{m_j} = 0$), which makes it an inefficient choice even in the absence of bias. Variance of the ratio estimator even increased further in presence of negative systematic measurement error $\mu^{m_j} = -0.30$, also shown in the supplemental material.

4.2.2 RMSE when mode 2 is the benchmark

Now consider the situation when we regarded m_2 as the benchmark ($b = 2$) shown in figures 4.3 and 4.4. The findings for the unadjusted estimators were identical, but for the adjusted estimators we found some differences to the situation when $b = 1$. We address these in the following.

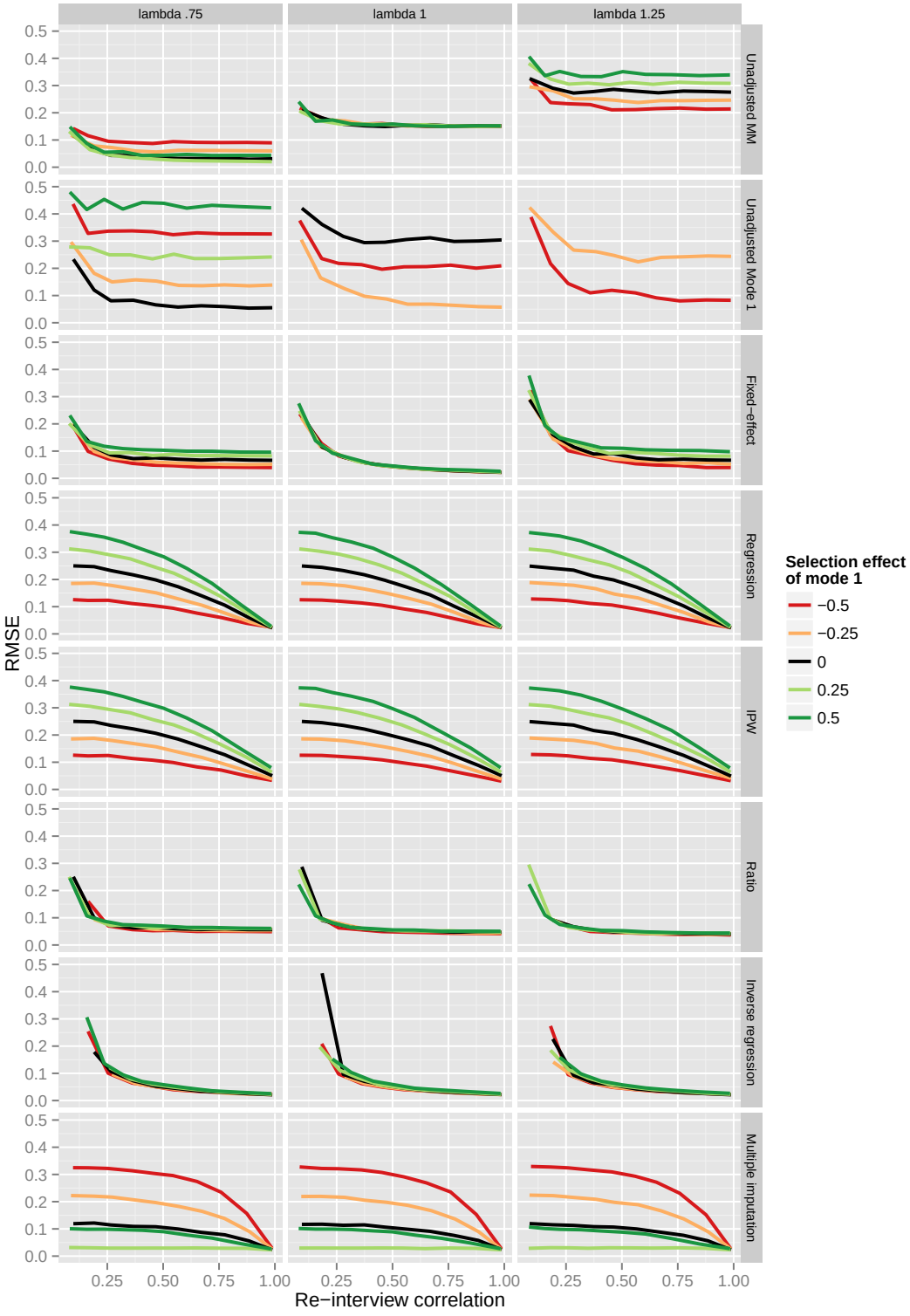

Figure 4.3 ***RMSE of adjusted and unadjusted estimators for benchmark mode*** $b = 2$***, meas. bias*** $\mu^{m_1} = 0.30$***, and re-interview*** $SE(\bar{Y}_{r_{re}}, \bar{Y}_{m_1}) = 0.50$***. IREG, ratio and fixed-effect perform well. However, fixed-effect can only fully reduce RMSE when*** $\lambda^{m_j} = 1$ ***and ratio maintains residual RMSE on low level.***

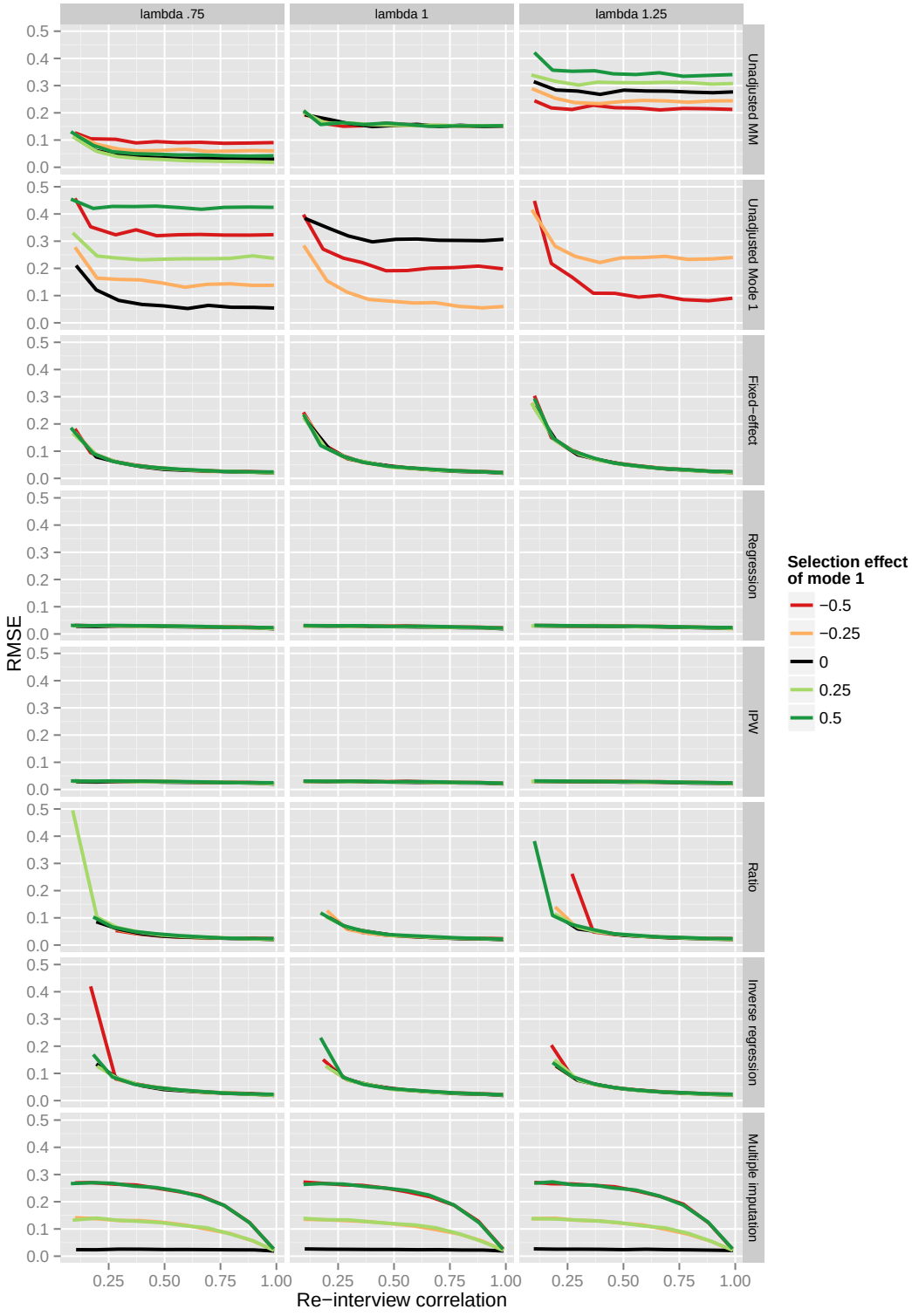

Figure 4.4 ***RMSE of adjusted and unadjusted estimators for benchmark mode*** $b = 2$***, meas. bias*** $\mu^{m_1} = 0.30$***, and re-interview*** $SE(\bar{Y}_{r_{re}}, \bar{Y}_{m_1}) = 0$***. All adjusted estimators except MI perform well.***

First, we found the IREG estimator again performed well, if re-interview correlations exceed a moderate level ($> .40$) regardless of the size of the re-interview SE. It can be seen that IREG even lead to somewhat smaller RMSE at each level of re-interview correlation compared to $b = 1$.

Second, also the finding that GREG, IPW, and MI have high RMSE was repeated, even though RMSE levels were lower than for $b = 1$ (Figure 4.3). However, RMSE of GREG and IPW nearly vanished in the absence of a re-interview SE (Figure 4.4). This is an immediate consequence of absent bias and minimal variance under this condition (cf. equation 27 for bias of GREG; simulated bias and variances can be taken from supplemental material). Since it is hard to diagnose the size of the re-interview SE in practice, however, we recommend against the use of the estimators as we did for m_1 as benchmark.

Third, the fixed-effect estimator had low RMSE under both re-interview SE conditions and regardless of levels of scaling parameter λ^{m_j}. At first glance, this finding is surprising, given that RMSE varies considerably across λ^{m_j} when m_1 is the benchmark (cf. Figure 4.1). This conjecture warrants a closer look at the role of the variance and bias components. From the supplemental material we may take that from re-interview $cor = .35$ variance was approximately negligible and bias became the dominant part of RMSE. In figures 4.3 and 4.4 this point is reached when the RMSE graphs are, roughly, horizontal. From bias equation (24) we may take that in the absence of a re-interview SE, the fixed-effect estimator is unbiased, so that RMSE approaches zero (Figure 4.4). In the presence of a re-interview SE some residual bias remains, but it is small ($< 10\%$) for moderate choices of λ^{m_j} (Figure 4.3). By equation (24) we may quantify maximum absolute bias for this simulation as $|B(\hat{\bar{Y}}^{fe}_{r_{mm}})| = |0.5(1 - \lambda^{m_j})(0.5)| = 6.25\%$. We thus find that even for relatively extreme choices for λ^{m_j} and $SE(Y_{r_1}, Y_{r_{mm}})$, bias and RMSE of the fixed-effect estimator do not reach extreme levels, which was reflected by the simulation results.

Fourth, we made a similar conjecture for the ratio estimator, which showed low levels of RMSE under all levels of λ^{m_j}. It can be shown that the maximum absolute bias of the ratio estimator is 5.77% (reached for $SE(Y_{r_1}, Y_{r_{mm}}) = -0.5, \lambda^{m_j} = 1$; cf. equation 25). However, as for $b = 1$ we observed that the variance of the ratio estimator is sensitive to random error (low re-interview correlations) when $\mu^{m_j} = 0$ or $\mu^{m_j} = -.30$, as can be seen in the variance plots available in the supplemental material. We found that from re-interview $cor > .50$ the ratio estimator performed well in all μ^{m_j} conditions and it performed even slightly better than the fixed-effect estimator.

In summary, when mode 1 was the benchmark, only the IREG estimator performed reliably, whereas ratio and fixed-effect estimators performed only well under special circumstances (fixed-effect if $\lambda^{m_j} = 1$, ratio if $\mu^{m_j} = 0$). When mode 2 was the benchmark, IREG performed again well, but also the ratio and fixed-effect estimators showed low levels of RMSE when re-interview correlation was moderate, even in the most extreme scenarios considered here. GREG, IPW, and MI performed badly in most considered scenarios due to their high bias.

5 Discussion

The present paper introduced a new approach for estimating and adjusting measurement bias (also called measurement effects) in mixed-mode surveys towards a benchmark mode by using re-interview data. This data is obtained from a subset of respondents to the first mode in a sequential design and it is employed as auxiliary data in a set of six adjusted candidate estimators. We evaluated by simulation, whether any of the estimators can outperform the unadjusted mean estimator in terms of mean squared error.

Earlier literature that attempts to estimate or adjust measurement effects can be criticized for potentially high bias, because researchers often had to assume that selection is ignorable conditional on weak auxiliary information (Vannieuwenhuyze, 2015; Vannieuwenhuyze & Loosveldt, 2013). Importantly, this study is among the first to demonstrate how estimating measurement effects in the presence of non-ignorable selection effects is practically feasible.

Our results demonstrated the potential of the re-interview approach for adjusting measurement effects, but the final choice of estimator depends on the analyst's expectations about the measurement error model of the focal mode and choice of benchmark mode. Generally, the IREG estimator performed well in all considered scenarios, whereas the ratio and fixed-effect estimator may be viable alternatives to be used if $\mu^{m_j} = 0$ or $\lambda^{m_j} = 1$ can be assumed, respectively. Given that the analyst often has insufficient information on the type of measurement error model, the IREG estimator may be the safest option in practice, but its use requires at least a moderate re-interview correlation of .50, better .70, to control its variance.

These conclusions were based on a large number of possible scenarios varying many parameters of the measurement and selection error models while keeping only a small number of parameters fixed (cf. Tables 4.1 and 4.2). Furthermore, our method can be extended for use with more than two modes. The choice of re-interview mode then becomes a more central decision in the design, as discussed in more detail for the cases of the Dutch Labor Force Survey and the American Community Survey in the appendix to this article .

Nevertheless, some limitations to the present study show up relevant paths for further theoretical and empirical work. Firstly, we assumed measurement equivalence between the re-interview and mode 2 (equation 10). As we discussed in section 2, the fieldwork design is a relevant factor determining the plausibility of this assumption. Longer time lags let appear equivalence more plausible because answers given under mode 1 tend to be forgotten. A time frame of several weeks appears to us as sufficient in many practical scenarios. However, to further re-assure against potential violations future research could develop designs and estimators for testing and adjusting the degree of measurement in-equivalence in re-interview designs.

Secondly, we assumed that the realizations of the benchmark mode represent the true scores towards which focal mode measurements are adjusted. A relevant path for further work is evaluating the robustness of our results to the introduction of random

error in the benchmark mode. Furthermore, designs which allow estimating random error of survey items, such as test-retest designs in the same mode, could be evaluated as a means to adjust for random error in GREG and IREG estimators.

Thirdly, despite the large variety of scenarios we considered in our simulation, some parameters were held constant. These parameters included the size of the overall response sample, the size of mode-specific response proportions, the size of the re-interviewed sample, and the response rate in the re-interview. In particular, the size of the re-interview response sample, which is a function of all aforementioned parameters, may have an impact on the efficiency of adjustments. In the present simulation, the variance of the adjusted estimators was small for all estimators when a moderate re-interview correlation was present. In further research, it should be evaluated, however, how robust these results are to changes in the size of the re-interview response sample (in this study, the expected re-interview n was $n_{re} = 375$). Another relevant factor we left out of consideration is the cost of the re-interview, which is also determined by the size of the re-interview sample (besides the mode). For fixed survey budgets a re-interview decreases overall sample size leading to an increase in sampling error of the re-interview mixed-mode design compared to a design without re-interview. Future work should therefore address the trade-off between cost, re-interview sample sizes, and error.

Appendix: Extending the re-interview design to multiple modes

So far, we have limited the number of modes to two. However, many real surveys use three or four modes. These include the Dutch Labor Force Survey (LFS) and the American Community Survey (ACS). Here, we consider the extension to three modes. Since the number of possible designs with three modes is already quite large, we restrict ourselves to the designs of the ACS and LFS. Extensions to other designs or designs with more than three modes can be done analogously.

Figure 5.1 presents the extension of figure 1 to the ACS and LFS. The ACS and LFS both start with a self-administered mode, and then do a follow-up of nonrespondents with telephone and face-to-face. In order to be able do employ telephone, a listed telephone number needs to be available. In figure 2, areas with a subscript 1 and a subscript 0 represent subsamples with and without a listed number, respectively. The LFS collects areas A_0, A_1, F_1 and B_0, where F_1 is a telephone follow-up and B_0 a face-to-face follow-up. The ACS also does a follow-up of telephone nonrespondents by face-to-face, i.e. it also collects area G_1.

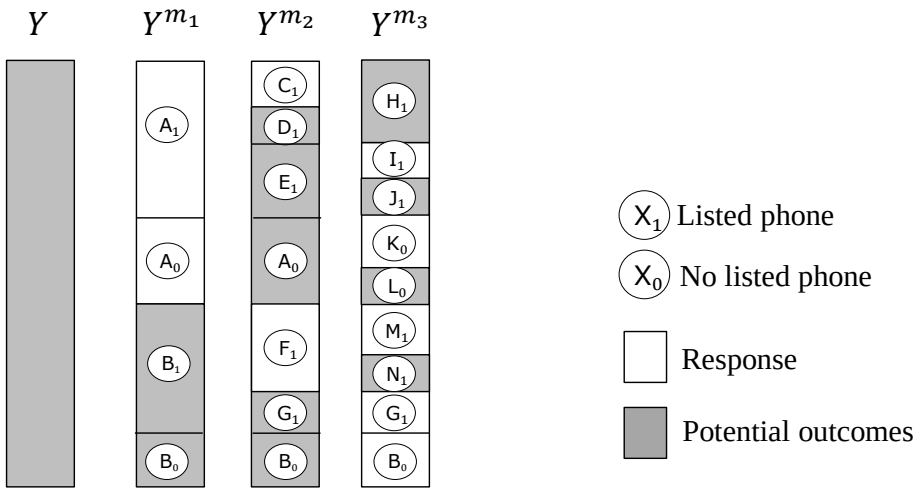

Figure 5.1 ***Schematic illustration of the missing data pattern a sequential mixed-mode re-interview design with three modes.***

The areas for re-interview depend on the benchmark mode. They are also slightly different between the LFS and ACS. Since re-interviews follow the order of modes in the sequence (m_1 - m_2 - m_3), the m_1 respondents with a telephone need to be randomly divided into two groups, E_1 and H_1. When m_1 is the benchmark, a re-interview is done on A_1 and A_0, leading to observations in C_1 and K_0. For the ACS, also I_1 is observed. Areas D_1 and L_0 are nonrespondents to the re-interviews. Additionally, area J_1 are non-respondents to the re-interview in the ACS. When m_2 is the benchmark mode, then

C_1 and M_1 are observed for both ACS and LFS, and D_1 and N_1 is non-response to the re-interview. When m_3 is the benchmark mode, then K_0 and M_1 are observed, and, additionally, I_1 for the ACS.

The estimation strategy can be shaped by making pairs with the benchmark mode. For example, when m_1 is the benchmark, the pairs (m_1, m_2) and (m_1, m_3) are formed, and the corresponding potential outcomes are estimated or imputed analogously to the two mode setting described in section 3. The recommended estimator per pair must be chosen based on conjectures about the measurement models and bias terms. However, following the simulation results of section 4, the inverse regression estimator will often be the preferred estimator.

References

Alwin, D. F. (2007). *Margins of Error*. Hoboken: Wiley.

Bang, H. & Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, *61*(4), 962–972. doi:10.2307/3695907

Bethlehem, J. (1988). Reduction of Nonresponse Bias Through Regression Estimation. *Journal of Official Statistics*, *4*(3), 251–260. Retrieved from http://www.jos.nu/Articles/abstract.asp?article=43251

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

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys.* New Jersey: Wiley.

Biemer, P. P. & Stokes, L. (1991). Approaches to the Modeling of Measurement Errors. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 487–517). Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley.

Buelens, B. & van den Brakel, J. A. (2014). Measurement Error Calibration in Mixed-mode Sample Surveys. *Sociological Methods & Research*. doi:10.1177/0049124114532444

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

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

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

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, *78*(1), 3–20. doi:10.1111/j.1751-5823.2010.00102.x

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. doi:10.1214/07-STS227

Klausch, T. (2014). *Informed Design of Mixed-Mode Surveys: Evaluating mode effects on measurement and selection error* (Unpublished PhD-Thesis, Utrecht University, Utrecht, The Netherlands).

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions. *Sociological Methods & Research*, *42*(3), 227–263. doi:10.1177/0049124113500480

Klausch, T., Hox, J., & Schouten, B. (2015). Selection error in single- and mixed mode surveys of the Dutch general population. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. doi:10.1111/rssa.12102

Klausch, T., Schouten, B., & Hox, J. J. (2015). Evaluating Bias of Sequential Mixed-mode Designs Against Benchmark Surveys. *Sociological Methods & Research*. doi:10.1177/0049124115585362

Kolenikov, S. & Kennedy, C. (2014). Evaluating Three Approaches to Statistically Adjust for Mode Effects. *Journal of Survey Statistics and Methodology*, *2*(2), 126–158. doi:10.1093/jssam/smu004

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. doi:10.1093/poq/nfn063

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

Lord, F. M. & Norvick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Welsley.

Lynn, P. (2013). Alternative Sequential Mixed-Mode Designs: Effects on Attrition Rates, Attrition Bias, and Costs. *Journal of Survey Statistics and Methodology*, *1*(2), 183–205. doi:10.1093/jssam/smt015

Raghunathan, T. E., Lepkowski, J., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

Rosenbaum, P. R. & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55. doi:10.1093/biomet/70.1.41

Rosenbaum, P. R. & Rubin, D. B. (1985). The Bias Due to Incomplete Matching. *Biometrics*, *41*(1), 103–116. doi:10.2307/2530647

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

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. doi:10.1198/016214504000001880

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

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Schafer, J. L. & Kang, J. D. Y. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313. doi:10.1037/a0014268

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1), 101–113.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, *42*(6), 1555–1570. doi:10.1016/j.ssresearch.2013.07.005

Suzer Gurtekin, Z. T. (2013). *Investigating the Bias Properties of Alternative Statistical Inference Methods in Mixed-Mode Surveys* (PhD thesis, University of Michigan, Michigan).

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

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064. doi:10.1080/10629360600810434

Vannieuwenhuyze, J. (2014). On the Relative Advantage of Mixed-Mode versus Single-Mode Surveys. *Survey Research Methods*, *8*(1), 31–42. Retrieved April 8, 2014, from https://ojs.ub.uni-konstanz.de/srm/article/view/5500

Vannieuwenhuyze, J. (2015). Mode Effects on Variances, Covariances, Standard Deviations, and Correlations. *Journal of Survey Statistics and Methodology*, *3*(3), 296–316. doi:10.1093/jssam/smv009

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. doi:10.1177/0049124112464868

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045. doi:10.1093/poq/nfq059

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

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

Colofon

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

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information