Journal of Survey Statistics and Methodology (2017) ***5***, *409–432*

ADJUSTING MEASUREMENT BIAS IN SEQUENTIAL MIXED-MODE SURVEYS USING RE-INTERVIEW DATA

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 in the context of inference from mixed-mode data. Doing so requires auxiliary information, which we suggest collecting 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 by Monte Carlo simulation the performance of six candidate estimators that exploit re-interview information. In the simulation parameters are systematically varied that define the size and

THOMAS KLAUSCH is postdoctoral researcher at VU University Medical Center Amsterdam, Department of Epidemiology and Biostatistics, De Boelelaan 1089a, 1081HV Amsterdam, The Netherlands. BARRY SCHOUTEN is Senior Methodologist at Statistics Netherlands, Division of IT and Methodology, Henri Faasdreef 312, 2492JP Den Haag, The Netherlands, and Professor at Utrecht University, Social and Behavioral Sciences, Department of Methodology and Statistics, Padualaan 14, 3584CH Utrecht, The Netherlands. BART BUELENS is Senior Methodologist at Statistics Netherlands, Division of IT and Methodology, CBS-weg 11, 6412EX Heerlen, The Netherlands. JAN VAN DEN BRAKEL is Senior Methodologist at Statistics Netherlands, Division of IT and Methodology, CBS-weg 11, 6412EX Heerlen, The Netherlands, and Professor at Maastricht University School of Business and Economics, Tongersestraat 53, 6211LM Maastricht, The Netherlands.

The authors would like to thank the editor and two anonymous reviewers for helpful comments on earlier versions of this manuscript. The research described in this paper was funded by Statistics Netherlands.

*Address correspondence to Thomas Klausch, VU University Medical Center Amsterdam, Department of Epidemiology and Biostatistics, De Boelelaan 1089a, 1081HV Amsterdam, The Netherlands; E-mail: t.klausch@vumc.nl.

doi: 10.1093/jssam/smx022 Advance access publication 23 September 2017

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.

KEYWORDS: Error adjustment; Measurement bias; Measurement error; Missing data; Mixed-mode survey.

1. INTRODUCTION

Sequential mixed-mode surveys combine multiple modes of data collection sequentially to optimize the trade-off between survey non-response and data collection costs. Usually, a sequential design starts with a cost-efficient mode (e.g., web data collection), and non-respondents to the first stage are approached by another mode (e.g., face-to-face). This second stage often strongly improves survey response, perhaps resulting in a reduction in survey non-response bias (Klausch, Hox, and Schouten, 2015). However, any mode has particular measurement error properties, and this makes certain modes more or less suitable for the measurement of specific target variables. For example, socially desirable answering can introduce systematic measurement error and thus bias in estimates of statistics describing sensitive characteristics. This behavior is stronger in interviewer-administered than in self-administered modes. Generally, when one or more modes in a design have higher systematic error than others, the measurement bias of linear estimates (e.g., means or totals) is increased when compared to a design using only the best measurement mode. This problem is one of the major challenges of mixed-mode designs (De Leeuw, 2005).

The present paper contributes to the growing body of literature that discusses statistical adjustment of differences in measurement bias between modes, also called measurement effects (Suzer Gurtekin, 2013; Kolenikov and Kennedy, 2014; Vannieuwenhuyze, 2015). We focus on the scenario when effects cannot be prevented by designing questionnaires that measure equally accurately under all modes (Dillman, Smyth, and Christian, 2009). The primary difficulty in estimating and adjusting measurement effects is confounding with selection effects in mixed-mode data. Selection effects denote a difference in the true score distributions between mode-specific response samples (Vannieuwenhuyze and Loosveldt, 2013). This article suggests an innovative approach to the confounding problem in the important case when selection effects depend on the target variable and thus cannot be explained by auxiliary information. This selection mechanism is missing not at random, MNAR (Little and Rubin, 2002). We suggest using a research design called the mixed-mode re-interview (Schouten, van den Brakel, Buelens, van der Laan, and Klausch, 2013; Klausch, Schouten, and Hox, 2017), which is similar to an internal calibration design used in epidemiology (Guo and Little, 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. In the following, we provide a conceptual introduction to the measurement bias adjustment problem and the re-interview design (section 2), describe adjustment by six candidate estimators (sections 3), and evaluate their performance in a simulation study (section 4).

2. THE SEQUENTIAL MIXED-MODE RE-INTERVIEW DESIGN

The data from a mixed-mode survey with two modes are sketched in figure 1, i (extensions to more modes are provided in the Supplementary material). We distinguish three types of variables: the “true” scores of a target variable, Y; variable Y as measured by mode m_1, $Y^{(1)}$; and variable Y as measured by mode m_2, $Y^{(2)}$ (Klausch, Schouten, and Hox, 2017). Response is characterized by white areas, and unavailable data are characterized by grey areas. The true scores Y are unobserved, whereas $Y^{(1)}$ and $Y^{(2)}$ are partly observed; 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 (field B and C) are called “potential”. Non-response to both modes is omitted from the figure.

2.1 The Problem

Our objective is to estimate the mean of Y over units that respond to at least one of the modes; that is, over all the rows in figure 1, denoted $\bar{Y}_{r_{mm}}$ (i.e., the mixed-mode mean; cf. definition in section 3.1). Measurement error in $Y^{(1)}$ and $Y^{(2)}$ may bias the estimator obtained by simply averaging the observed values of $Y^{(1)}$ and $Y^{(2)}$ across all respondents. We seek to reduce the mean squared error (MSE) of this naïve unadjusted estimator. However, it is impossible to correct the measurement bias of both modes because true scores are unknown. We, therefore, assume one mode is a measurement benchmark (gold standard) setting its observed scores equal to Y, whereas the alternative mode is denoted focal mode. This choice depends on the combination of mode and question that evokes the least or no measurement error. For example, in a survey of alcohol use, answers on the phone or in person often suffer from social desirability bias and recall error. In a mixed-mode survey, self-administration (e.g., web) may, therefore, be set as the measurement benchmark.

To adjust measurement bias, we need to estimate the potential benchmark outcomes of either $Y^{(1)}$ or $Y^{(2)}$, which requires solving a missing data problem. Furthermore, due to confounding of measurement and selection effects,

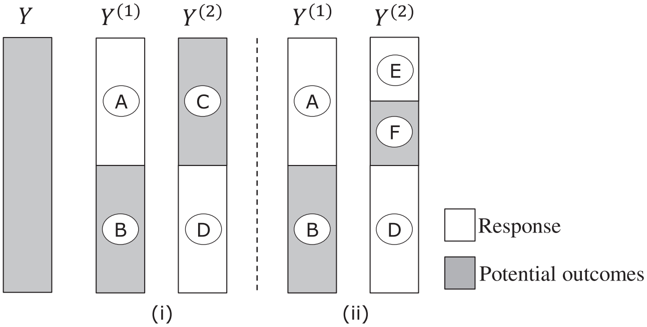

Figure 1. Missing data pattern of two sequential mixed-mode surveys: left (i) a simple sequential design, right (ii) a sequential design with re-interview. The true scores Y are not observed. Respondents in m_1 and m_2 provide $Y^{(1)}$ (field A) and $Y^{(2)}$ (field D). Field B denotes potential outcomes on $Y^{(1)}$ due to non-response in m_1. Field C denotes potential outcomes $Y^{(2)}$ due to response in m_1. Re-interview data in (ii) create overlap between the m_1 (field A) and m_2 (field E) response distributions. If sub-sampling is used for the re-interview, a random part of fields E and F is not observed.

a difference in respondent means on $Y^{(1)}$ and $Y^{(2)}$ can denote a measurement effect, a selection effect, or a combination. Using only the data collected by a simple sequential design, adjusting measurement bias is, therefore, not feasible. However, sometimes exogenous covariates are available, such as sampling frame information. One possible assumption is that, conditional on this data, the observed and missing distributions of the benchmark outcomes are equal; the outcomes are then missing at random, or MAR (Rubin, 1976; Little and Rubin, 2002). Under this assumption, estimation is straightforward. We study the case when alternative assumptions to MAR are considered more plausible.

2.2 Design and Use of a Re-Interview Extension

The re-interview method is an alternative way to collect auxiliary data. The design consists of a sequential mixed-mode survey, where additionally a subset of respondents in m_1 is followed up in m_2 (figure 1, ii). The re-interview data create overlap, so that we observe the outcomes in both modes in this subset of respondents (field E). This information is exploited during estimation. For simplicity, figure 1 (ii) shows the situation when all m_1 respondents are re-interviewed, and hence field F denotes re-interview non-response. However, sub-sampling of m_1 is possible and desirable in practice, as discussed below. If sub-sampling is used, response and non-response of units in fields E and F that are not sub-sampled is not observed.

Four aspects of the re-interview design are highlighted. First, re-interview studies are practically feasible. For example, Schouten et al. (2013) and Klausch, Hox, and Schouten (2015) executed a large-scale re-interview experiment in the Dutch Crime Victimization Survey (CVS), approaching respondents and non-respondents to the initial mode (e.g., web or mail) again in face-to-face. Another example is the American Community Survey by the US Census Bureau, which has used re-interview studies for estimating survey error on regular basis (Shin, 2012).

Second, the introduction of a re-interview to an ongoing mixed-mode design does not impact the standard fieldwork of the sequential mixed-mode survey. Since m_1 respondents are re-interviewed after their $Y^{(1)}$ answers have been recorded, the additional measurement occasion cannot "bias" the regular measurement process.

Third, the re-interview fieldwork incurs additional costs, but these can be reduced by restricting the re-interview to a sub-sample. An investment is justified only when moderate to large measurement bias is anticipated. Efficient subsampling schemes of m_1 respondents and non-respondents can be developed. An optimal scheme depends on benchmark mode, overall sample size, the response rate in m_1, and the response rates in the re-interview and follow-up in m_2, besides the particular error properties of the survey outcome variable. We also expect costs per re-interview to be slightly larger than for a follow-up unit. However, the trade-off between gain in MSE and required investment is complex and left for future study. In the simulation, we focus on moderate re-interview sample sizes.

Fourth, 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. We assume that the measurement error model in the re-interview and the regular m_2 model are identical and call this assumption "measurement equivalence". The assumption has to be scrutinized in practice; see the discussion section (Forsman and Schreiner, 1991; Biemer and Forsman, 1992).

2.3 Relationships to Earlier Literature

Measurement bias adjustment of regression coefficients has been discussed in epidemiological applications using so-called calibration samples (e.g., Freedman, Midthune, Carroll, and Kipnis, 2008; Guo and Little, 2011, 2013). A sample is available where benchmark measurements are observed together with outcomes under measurement error. Three adjusted estimators are applied. First, regression estimation regresses benchmark outcomes on outcomes under measurement error and uses the relationship to predict benchmark outcomes in the study set. Second, classical calibration proceeds by regressing outcomes with measurement error on benchmark outcomes and then inverting the regression equation to predict benchmark outcomes. Below, this estimator is referred to as inverse regression. Third, multiple imputation draws repeatedly from the conditional distributions of the benchmark outcomes in the study set to predict potential outcomes. All estimators assume that benchmark outcomes in the study set are MAR conditional on auxiliary data.

Unlike this research, the present study evaluates the performance of mean estimators when benchmark outcomes are missing not at random (MNAR) as opposed to MAR. That is, we assume response to a mode depends on the value of the partly unobserved target variable. MNAR holds, because auxiliary data are absent in the re-interview design (figure 1), and the observed information insufficiently explains the missing data mechanism, as it is partly observed under random measurement error.

Estimating means of variables that are MNAR is possible in pattern mixture models (PMM) as discussed in Little (1994) who demonstrates that, under normality, maximum likelihood estimators for means of MNAR variables are equivalent to the classical calibration (inverse regression) estimator. Using a calibration sample, West and Little (2013) apply the PMM approach and demonstrate good performance of the estimator. They also demonstrate substantial bias of multiple imputation and propensity score weighting estimators under MNAR.

The missing data setting in the re-interview design differs from West and Little (2013) and other classical calibration designs, in that the re-interview sub-sample can suffer from MNAR non-response causing a non-monotone missing data pattern. Patterns assumed in previous approaches are monotone as the calibration samples are complete. Our estimation strategy is detailed in section 3.3.

3. MEASUREMENT BIAS ADJUSTMENT

This section discusses the candidate estimators used in adjusting for measurement bias in re-interview designs. We first present a statistical model for the data generating process and then discuss a set of six adjusted estimators.

3.1 Fixed Response Model

For the simple sequential mixed-mode survey, we assume a fixed response model (Cochran, 1977) that separates all units $i = 1, \ldots, N$ in a population of size N into two response strata (units participating in either m_1 or m_2) and a non-response stratum. Since we focus on the respondent mean in this paper (cf. section 2), the non-response stratum is ignored in the following. Let fixed indicator variables r_{1i} and $r_{2i} = 1 - r_{1i}$ identify membership of unit i in the response strata of modes $j = \{1, 2\}$, and $N_{r_j} = \sum_{i=1}^{N} r_{ji}$ be the population sizes. Let $P_j = N_{r_j}/N$ denote the relative size of the strata, and let $\bar{Y}_{r_j} = N_{r_j}^{-1} \sum_{i=1}^{N} r_{ji} y_i$ be the stratum mean, where y_i is the true score of unit i on continuous target variable y. The mixed-mode mean is then given by $\bar{Y}_{r_{mm}} = N^{-1} \sum_{i=1}^{N} y_i = P_1 \bar{Y}_{r_1} + P_2 \bar{Y}_{r_2}$. The contrast $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = \bar{Y}_{r_1} - \bar{Y}_{r_{mm}} = P_2(\bar{Y}_{r_1} - \bar{Y}_{r_2})$ denotes the selection effect (SE) of the mode 1 respondent mean, $\bar{Y}_{r_1}$, relative to overall mean $\bar{Y}_{r_{mm}}$. It can be seen that the relative SE 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. The relative SE between modes is a major motivation for conducting mixed-mode surveys.

When a re-interview is added to the sequential mixed-mode design, the population response stratum in m_1 is split into a re-interview response and re-interview non-response stratum. Whether a unit is sub-sampled for the re-interview is a property of the design, addressed in the estimation section 3.3. Let $r_{re,i}$ ("re" for re-interview) denote the indicator whether unit i, which is a respondent in m_1 ($r_{1i} = 1$, field A), also responds in the re-interview. Note $r_{re,i}$ is not defined if $r_{1i} = 0$. Then, $P_{re} = N_{r_{re}}/N_{r_1}$ denotes the relative size of the re-interview response stratum to the number of m_1 respondents in the population, where $N_{r_{re}} = \sum_{i=1}^{N} r_{re,i}$. Let the population mean in the re-interview response stratum be $\bar{Y}_{r_{re}}$; then the contrast $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) = \bar{Y}_{r_{re}} - \bar{Y}_{r_1}$ denotes the re-interview SE, which occurs when systematically different respondents participate in the re-interview than in m_1. Such effects are practically relevant, because a different mode is offered (m_2), and there is respondent burden.

3.2 Measurement Model

Each mode is associated with a question-specific measurement error model describing the relation of true scores y_i to the observed outcomes in m_j, denoted $y_i^{(j)}$, as (Biemer and Stokes, 1991)

$$y_i^{(j)} = \mu^{(j)} + \lambda^{(j)}(y_i + u_i^{(j)}) \quad \forall \quad i, \tag{1}$$

where $\lambda^{(j)}$ is a scale parameter equal to 1 if m_j measures on the scale of the true score, and $u_i^{(j)}$ is an independently and identically distributed measurement error term with $u_i^{(j)} \sim iid(0, (\sigma_u^{(j)})^2)$. The parameter $\mu^{(j)}$ is systematic measurement error common to all units, whereas $(\sigma_u^{(j)})^2$ denotes the variance of measurement errors in the population and is called the random measurement error. We assume independence of true scores y_i and measurement errors $u_i^{(j)}$ for all i and j. Let $c_j = cor(Y^{(j)}, Y)$ be the population correlation between $Y^{(j)}$ and Y. The variance of random errors $(\sigma_u^{(j)})^2$ is related to the population variance of true scores Y, denoted σ_Y^2, and c_j by the identity

$$(\sigma_u^{(j)})^2 = \frac{1 - c_j^2}{c_j^2}\sigma_Y^2. \tag{2}$$

Biemer and Stokes (1991) call c_j the reliability coefficient.

The measurement equivalence assumption, mentioned in section 2.2, implies that the random error $u_i^{(j)}$ is independent of response to the main interview or the re-interview and that model parameters $\mu^{(j)}$ and $\lambda^{(j)}$ are the same in all response strata.

3.3 Candidate Estimators Using Re-Interview Data

The unadjusted estimator of the mixed-mode mean is

$$\widehat{\bar{Y}}^{unadj}_{r_{mm}} = \frac{1}{\widehat{N}_{r_1} + \widehat{N}_{r_2}} \sum_{i=1}^{N} I_i d_i \left(r_{1i} y_i^{(1)} + r_{2i} y_i^{(2)}\right), \tag{3}$$

where d_i denote 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 $\hat{N}_{r_j} = \sum_{i=1}^{N} I_i d_i r_{ji}$. Its bias is $B(\widehat{\bar{Y}}^{unadj}_{r_{mm}}) \approx P_1((\lambda^{(1)} - 1)\bar{Y}_{r_1} + \mu^{(1)}) + P_2((\lambda^{(2)} - 1)\bar{Y}_{r_2} + \mu^{(2)})$ where the two terms denote weighted measurement biases contributed by modes 1 and 2. If m_1 or m_2 represent a measurement benchmark, one of the measurement bias terms is zero. We now suggest six candidate estimators that can correct for this bias using re-interview data. As discussed in section 2, this requires estimating the potential benchmark outcomes (figure 1, ii) by using missing data methods. If mode 1 is the benchmark, figure 1 (ii) shows that benchmark outcomes in field B are missing and need to be estimated, but the re-interview (in the focal mode) lacks observations in field F. If mode 2 is the benchmark, outcomes in field F need to be estimated, but focal mode outcomes in field B are missing. Estimators for both situations follow the same form. For brevity, we give estimators for the case when m_1 is the benchmark. The situation when m_2 is the benchmark follows analogously.

To introduce random sub-sampling of m_1 respondents, let indicator $s_{re,i}$ determine whether unit i is selected for a re-interview. Furthermore, let $P_s = \sum_{i=1}^{N} s_{re,i} / \sum_{i=1}^{N} r_{1i}$ denote the proportion of sub-sampled respondents for the re-interview. If $P_s = 1$, all m_1 respondents are approached for a re-interview, whereas choices $P_s < 1$ make the design cost efficient. The suggested estimators all assume simple random sub-sampling.

We consider two classes of estimators referred to as π-estimators and y-estimators, respectively (Särndal and Lundström, 2005; Kang and Schafer, 2007). The 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). Denote the propensity for unit i to be observed in the benchmark mode m_1 by π_i and estimate it as:

$$\begin{aligned}\pi_i &= P(r_{1i} = 1 | y_i^{(2)}, r_{2i} = 1 \cup (r_{re,i} = 1 \cap s_{re,i} = 1)) \\ &= \frac{1}{1 + \exp(-(\theta_0 + \theta_1 y_i^{(2)}))} \quad \forall \quad i.\end{aligned} \tag{4}$$

In the re-interview setting there are missing observations due to re-interview non-response and sub-sampling, which requires conditioning model (4) on observed auxiliary vector $y^{(2)}$ and the set of all i for which $r_{2i} = 1 \cup (r_{re,i} = 1 \cap s_{re,i} = 1)$. The inverse propensity weighting (IPW) estimator is (Rosenbaum, 1987):

$$\begin{aligned}\widehat{\overline{Y}}_{r_{mm}}^{ipw} &= \frac{1}{\widehat{N}_1 + \widehat{N}_2} \left(\sum_{i=1}^{N} I_i d_i y_i^{(1)} r_{re,i} s_{re,i} \widehat{\pi}_i^{-1} + \sum_{i=1}^{N} I_i d_i y_i^{(1)} (1 - r_{re,i} s_{re,i}) \right) \\ &= \frac{1}{\widehat{N}_1 + \widehat{N}_2} \sum_{i=1}^{N} I_i d_i y_i^{(1)} (r_{re,i} s_{re,i} \frac{(1 - \widehat{\pi}_i)}{\widehat{\pi}_i} + 1) \quad \forall \quad i.\end{aligned} \tag{5}$$

As can be seen from the first equation, the estimator consists of two sums. The first sum is a standard weighting estimator of the total of benchmark $Y^{(1)}$ across the group of re-interview respondents and focal mode respondents $(r_{2i} = 1 \cup (r_{re,i} = 1 \cap s_{re,i} = 1))$. However, it omits the set of re-interview non-respondents and units not sub-sampled (Field F), for which the total is given by the second sum of benchmark outcomes. If m_2 is benchmark, $\pi_i = P(r_{re,i} = 1 \cap s_{re,i} = 1 | y_i^{(1)}, r_{1i} = 1)$ follows. An alternative IPW estimator omits units not sub-sampled and applies sub-sampling weights P_s^{-1}. This estimator uses $\pi_i = P(r_{re,i} | y_i^{(1)}, r_{1i} = 1, s_{re,i} = 1)$ and has very similar performance if sub-sampling weights are not extreme.

Unlike π-estimators, y-estimators seek good predictions of the potential benchmark outcomes $y^{(1)}$ using a suitable model for $y^{(1)}$ and finally sum over the joint vector of observed and predicted scores (Kang and Schafer, 2007). A general form of the y-estimator is

$$\widehat{\overline{Y}}_{r_{mm}}^{yest} = \frac{1}{\widehat{N}_1 + \widehat{N}_2} \sum_{i=1}^{N} I_i d_i (r_{1i} y_i^{(1)} + r_{2i} \widehat{y}_i^{(1)}) \quad \forall \quad i, \tag{6}$$

where $\hat{y}_i^{(1)}$ represent the estimated potential (unobserved) benchmark outcomes for respondents in the focal mode (m_2). The y-estimator is based on a y-model that describes the relationship of benchmark to alternative mode outcomes. It is then assumed that the model also holds in the response stratum to mode m_2 and can be used to transform observed focal mode $y^{(2)}$ to benchmark $y^{(1)}$ (Little and Rubin, 2002). An intuitive y-model corrects each focal mode outcome $y_i^{(2)}$ by the fixed observed mean difference in the re-interview sample as a simple estimate of measurement effect, i.e.,

$$\hat{y}_i^{(1)} = y_i^{(2)} - (\hat{\bar{Y}}_{re}^{(2)} - \hat{\bar{Y}}_{re}^{(1)}), \tag{7}$$

where $\hat{\bar{Y}}_{re}^{(2)}$ and $\hat{\bar{Y}}_{re}^{(1)}$ the respondent means of focal and benchmark mode outcome in the re-interview. We call this the "fixed-effect" estimator, denoted as $\hat{\bar{Y}}_{r_{mm}}^{fe}$. While it may be realistic to omit scale differences between modes for some types of survey variables, it may be too simplistic for many others. Two estimators for survey data with non-response that account for these scale differences are the ratio estimator ($\hat{\bar{Y}}_{r_{mm}}^{ratio}$), which uses prediction

$$\hat{y}_i^{(1)} = y_i^{(2)} \frac{\hat{\bar{Y}}_{re}^{(1)}}{\hat{\bar{Y}}_{re}^{(2)}}, \tag{8}$$

and a standard or generalized regression estimator ($\hat{\bar{Y}}_{r_{mm}}^{greg}$), which uses prediction

$$\hat{y}_i^{(1)} = \hat{\bar{Y}}_{re}^{(1)} - \hat{\beta}_{re}(\hat{\bar{Y}}_{re}^{(2)} - y_i^{(2)}), \tag{9}$$

(Särndal and Lundström, 2005), where β_{re} denotes the (population) "slope" of the linear regression of $Y^{(1)}$ on $Y^{(2)}$ in the re-interview stratum. Unlike the regression estimator which uses focal outcomes $y^{(j)}$ as covariates, the inverse regression estimator (IREG) models $y^{(2)}$ as outcome in $y_i^{(2)} = \nu_0 + \nu_{re} y_i^{(1)} + \epsilon_i$, and then inverts the regression equation to impute $y^{(1)}$ (Brown, 1990; Little, 1994)

$$\hat{y}_i^{(1)} = \hat{\bar{Y}}_{re}^{(1)} - \frac{1}{\hat{\nu}_{re}}\left(\hat{\bar{Y}}_{re}^{(2)} - y_i^{(2)}\right). \tag{10}$$

In practice, the parameters β_{re}, ν_{re}, as well as $\bar{Y}_{r_{re}}^{(2)}$ and $\bar{Y}_{r_{re}}^{(1)}$, are estimated by their sample analogues in the re-interview response stratum. Bootstrapped standard errors provide a robust method for variance estimation, because no closed-form variance expressions of the above point estimators currently exist.

Finally, we consider simultaneous multiple imputation for measurement error adjustment (Guo and Little, 2013) using the MICE algorithm (multiple imputation by chained equations; van Buuren, 2012). The procedure specifies the conditional distributions of benchmark and focal mode outcomes as normal regression models and initially completes the missing data by draws from the observed distributions (fields A for $Y^{(1)}$ and E + D for $Y^{(2)}$, respectively). The procedure then alternates between the two variables predicting the potential outcomes by drawing from their predictive distributions, converging to the bivariate distribution like a Gibbs sampler. We evaluate this procedure with five imputed data sets pooled by Rubin's rules.

3.4 Residual Bias of the Adjusted Estimators

We approximate the bias of the fixed-effect, ratio, regression, and inverse regression estimators (cf. proof in Supplementary material) and give conditions for unbiasedness of MICE. For the first three we have:

$$B(\hat{\bar{Y}}^{fe}_{r_{mm}}) \approx P_j((1-\lambda^{(j)})(\bar{Y}_{r_{re}} - \bar{Y}_{r_j})), \tag{11}$$

$$B(\hat{\bar{Y}}^{ratio}_{r_{mm}}) \approx P_j(\mu^{(j)} \frac{\bar{Y}_{r_{re}} - \bar{Y}_{r_j}}{\lambda^{(j)}\bar{Y}_{r_{re}} + \mu^{(j)}}), \tag{12}$$

$$B(\hat{\bar{Y}}^{greg}_{r_{mm}}) \approx P_j((1-\lambda^{(j)}\beta_{re})(\bar{Y}_{r_{re}} - \bar{Y}_{r_j})), \tag{13}$$

where $\beta_{re} = \sigma_Y^2/(\lambda^{(j)}(\sigma_Y^2 + (\sigma_u^{(j)})^2))$ and index $j = 1,\ 2$ indicates the focal mode. Bias of the ratio and GREG estimator are approximated using Taylor linearization (Särndal and Lundström, 2005), where the remainder terms vanish in large samples. Bias of IPW is hard to approximate analytically and is simulated instead. However, we expect IPW to perform similarly to GREG.

From (11)–(13), $\hat{\bar{Y}}^{fe}_{r_{mm}}$, $\hat{\bar{Y}}^{ratio}_{r_{mm}}$, and $\hat{\bar{Y}}^{greg}_{r_{mm}}$ are unbiased if the contrast $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_j}) = \bar{Y}_{r_{re}} - \bar{Y}_{r_j} = 0$. If m_2 is benchmark, this is simply the re-interview SE, $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$. If m_1 is benchmark, $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_2}) = SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$ $+P_2^{-1}SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$ depends on re-interview SE and SE of mode 1 relative to $\bar{Y}_{r_{mm}}$. Only when both SEs balance, $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = -P_2 SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$, or are zero, are the estimators unbiased. Note again that $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) \neq 0$ is a major reason to conduct mixed-mode surveys and $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) = 0$ is unlikely (cf. section 3.1). In practice, we, therefore, expect bias in these estimators regardless of benchmark.

However, further conditions may create unbiasedness. From (11), $\hat{\bar{Y}}^{fe}_{r_{mm}}$ is unbiased if $\lambda^{(j)} = 1$ and, from (12), $\hat{\bar{Y}}^{ratio}_{r_{mm}}$ is unbiased if $\mu^{(j)} = 0$. Thus, $\hat{\bar{Y}}^{fe}_{r_{mm}}$ corrects a systematic error difference between modes if there is no scale difference and conversely $\hat{\bar{Y}}^{ratio}_{r_{mm}}$ if there is no systematic error. From (13), $\hat{\bar{Y}}^{greg}_{r_{mm}}$ is unbiased if $\lambda^{(j)}\beta_{re} = 1$ and thus $\sigma_Y^2/(\sigma_Y^2 + (\sigma_u^{(j)})^2) = 1$. The bias of GREG thus does not depend on $\lambda^{(j)}$ and is determined by the size of random error variance. As the focal mode usually measures with error, GREG is biased in most scenarios.

The bias of MICE is hard to derive, but it is well known that imputation estimators are consistent under MAR (Little and Rubin, 2002). However, no auxiliary data is available outside the focal and benchmark outcomes that would allow MAR in the re-interview design. Due to random measurement error in the focal mode, response is MNAR for benchmark outcomes even if focal outcomes were fully observed (if the SE between modes is not zero). Consistency of the MICE estimator therefore requires mean equality of the observed and unobserved parts of benchmark Y (cf. figure 1 ii). If m_1 is benchmark, it is sufficient if $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = 0$, so that there is no SE between m_1 and m_2. For the m_2 benchmark,

$$SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = \frac{P_1 P_{re} P_s}{P_1(1 - P_{re} P_s)} SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) \tag{14}$$

is sufficient, so that both SEs are zero or balanced. As in the discussion above, these situations are unlikely in practice. Note the conditions of MICE and GREG differ (13), and this difference emerges due to the missing data pattern of the re-interview design. Whereas MICE completes all data, GREG omits focal mode non-respondents for which benchmark data is already available. As for GREG, however, MICE is unbiased if $\sigma_u^{(j)} = 0$ as then $Y^{(1)}$ and $Y^{(2)}$ are linear combinations of each other. In other situations, we expect the estimators' performances to differ.

By analogy to (13), IREG is approximately unbiased:

$$B(\hat{\bar{Y}}^{ireg}_{r_{mm}}) \approx P_j((1 - \lambda^{(j)} v_{re}^{-1})(\bar{Y}_{r_{re}} - \bar{Y}_{r_j})) = 0, \tag{15}$$

because $v_{re}^{-1} = (\lambda^{(j)})^{-1}$. This property of the IREG estimator is a result of the measurement equivalence assumption. We return to this point in the discussion.

4. SIMULATION STUDY

In practice, the distributions of Y and the response and measurement model parameters are unknown. In this section we assess the potential effects that different choices of the parameters have on the root mean square error (RMSE) of the unadjusted and adjusted estimators, by Monte Carlo simulation.

4.1 Simulation Set-up

Tables 1 and 2 give an overview on the parametrization of the response and measurement models. There are five parameters to be specified in the fixed response model. Three of these parameters were "fixed" and two were varied ("free"), as listed in table 1. We fixed the values for $\bar{Y}_{r_{mm}} = 1$, $P_1 = 0.5$, and $P_{re} = 0.6$. Values of P_1 and P_{re} were based on a web - face-to-face mixed-mode re-interview design (Schouten et al., 2013; Klausch, Hox, and Schouten, 2015), where about 50 percent ($= P_1$) of respondents replied in web and about 60 percent ($= P_{re}$) of web respondents participated in the face-to-face re-interview. We sub-sample every second m_1 respondent for the re-interview ($P_s = 0.5$).

Table 1. Parametrization of the Super-Population Response Model in the Simulation

Parameter	Value(s) in simulation	Description
Fixed:		
$\bar{Y}_{r_{mm}}$	1	Mixed-mode mean
P_1	0.5	Rel. response rate to mode 1
P_{re}	0.6	Response rate in re-int.
$(P_s)^*$	$(0.5)^*$	(Prop. of sub-sampled m_1 resp.)*
Free:		
$SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$	$\{-0.5, -0.25, 0, 0.25, 0.5\}$	Selection effect of mode 1
$SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$	$\{0, 0.5\bar{Y}_{r_1}\}$	Re-interview selection effect
Dependent:		
$\bar{Y}_{r_1}$	Dep. on $\bar{Y}_{r_{mm}}$, $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$	Response stratum mean of m_1
$\bar{Y}_{r_2}$	Dep. on $\bar{Y}_{r_{mm}}, \bar{Y}_{r_1}, P_1$	Response stratum mean of m_2
$\bar{Y}_{r_{re}}$	Dep. on $\bar{Y}_{r_1}$, $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$	Response stratum mean in re-int.
P_2	0.5, dep. on P_1	Rel. response rate to mode 2
σ_Y^2	Dep. on $SE(\bar{Y}_{r_1}, \bar{Y}_{r_2})$, $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$	Population variance of Y

prop.: proportion; resp.: response/respondents; re-int.: re-interview; rel.: relative
* P_s is a feature of the sampling design, not the super-population response model.

Table 2. Parametrization of the Super-Population Measurement Model in the Simulation

Parameter	Value(s) in simulation	Description
Fixed:		
$\mu^{(b)}$	0	Benchmark mode systematic error
$\lambda^{(b)}$	1	Benchmark mode scale parameter
$(\sigma_u^{(b)})^2$	0	Benchmark mode error variance
Free:		
b	$\{1, 2\}$	Benchmark mode, $b \neq j$
$\mu^{(j)}$	$\{-0.3, 0, 0.3\}$	Focal mode systematic error
$\lambda^{(j)}$	$\{0.75, 1, 1.25\}$	Focal mode scale parameter
c_j	$\{0.1, 0.2, \ldots, 1\}$	True-observed score correlation
Dependent:		
j	$\{1,2\}$, dep. on b	Focal mode, $b \neq j$
$(\sigma_u^{(j)})^2$	Dep. on c_j and σ_Y^2	Focal mode error variance

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 response model. $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$ was varied from absent (0 percent) to strong selectivity (± 50 percent relative effect to $\bar{Y}_{r_{mm}} = 1$). $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$ was either set to 0 or to 50 percent bias relative to $\bar{Y}_{r_1}$ (i.e., $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) = 0.5\,\bar{Y}_{r_1}$). Within each response stratum, true scores y_i were generated from three Gaussian super-populations: $y_i \sim N(\bar{Y}_{r_{re}}, 1)$ if $r_{1i} = 1 \cap r_{re,i} = 1$; $y_i \sim N(\bar{Y}_{r_{nre}}, 1)$ if $r_{1i} = 1 \cap r_{re,i} = 0$; and $y_i \sim N(\bar{Y}_{r_2}, 1)$ if $r_{1i} = 0$. The resulting population response distribution of Y is a mixture of Gaussians with $\bar{Y}_{r_{mm}} = 1$ and population variance σ^2_Y, which is a function of the within-stratum variance (set to 1) and the between-stratum variances determined by the selection effects.

In the measurement model we distinguish two situations with either m_1 or m_2 as the benchmark. We write $b = \{1, 2\}$ for benchmark mode and $j = \{1, 2\}$ for focal mode ($j \neq b$). The measurement model then has six further parameters that need to be specified (table 2). The parameters of the benchmark measurement model were fixed: $\mu^{(b)} = 0$, $\lambda^{(b)} = 1$, and $(\sigma^{(b)}_u)^2 = 0$. The parameters of the focal mode were varied (1). We introduced either no systematic error, $\mu^{(j)} = 0$, or moderate systematic measurement error (±30 percent relative to the population mean). The scaling parameter $\lambda^{(j)}$ was varied for moderate scale differences, scaling $y^{(j)}$ up ($\lambda^{(j)} = 1.25$) and down ($\lambda^{(j)} = 0.75$). The variance of random error $(\sigma^{(j)}_u)^2$ was controlled by varying true-observed score correlation c_j between 0.1 (very high error variance) and 1 (no error variance). We sampled u from a normal distribution, i.e., $u^{(j)}_i \sim N(0, (\sigma^{(j)}_u)^2)$. 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 condition, we generated a population of size $\mathrm{N} = 100{,}000$ from the super-population. We then drew $K = 1{,}000$ simple random samples with expected size $n_{sample} = 2{,}500$ without replacement from each population. Every second m_1 respondent was randomly selected for a re-interview ($P_s = 0.5$), yielding on average a moderate re-interview sample size (expected re-interview $n_{re} = (P_1)(P_{re})(P_s)n_{sample} = 375$). For each data set we computed the six adjusted estimators as well as two unadjusted estimators. Specifically, the first unadjusted estimator is given by (3) and the second one is given by

$$\hat{\hat{Y}}^{unadj2}_{r_{mm}} = \frac{1}{\hat{N}_{r_1}} \sum_{i=1}^{N} I_i d_i r_{1i} y^{(1)}_i. \tag{16}$$

This estimator mimics the estimator for a single-mode survey in the first mode rather than the mixed-mode design. We finally estimated the root MSE as $R\hat{M}SE(\hat{\hat{Y}}_{r_{mm}}) = (K^{-1} \sum_{k=1}^{K} (\hat{\hat{Y}}_{r_{mm},k} - \bar{Y}^{pop}_{r_{mm}})^2)^{1/2}$ where $\bar{Y}^{pop}_{r_{mm}}$ is the true population mean for the given condition. Since $\bar{Y}^{pop}_{r_{mm}} \approx 1 = \bar{Y}_{r_{mm}}$, the estimated RMSE also has the interpretation as an approximate *relative RMSE* $(= R\hat{M}SE(\hat{\hat{Y}}_{r_{mm}}) / \bar{Y}_{r_{mm}})$.

4.2 Results

Figures 2–5 illustrate the key results of the simulation. Each figure plots the estimated RMSE of the two unadjusted and the six adjusted estimators for three levels of focal mode $\lambda^{(j)}$ against the correlation between $Y^{(1)}$ and $Y^{(2)}$ observed in the re-interview ("re-interview correlation"). This correlation can be compared to re-interview correlations for several variables from the Crime Victimization and Labor Force Survey ranging from 0.39 to 0.75 (Schouten et al., 2013). 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 (2) which is varied systematically from 0.1 to 1 (table 2).

All figures display the condition where focal mode systematic measurement error was set to +30 percent ($\mu^{(j)} = 0.30$) of the mixed-mode mean ($\bar{Y}_{r_{mm}} = 1$). We provide the figures for the 0 percent and –30 percent conditions in the Supplementary material. The results presented here for +30 percent generally held for these conditions. We highlight the few exceptions below. Furthermore, we focus on RMSE, but we provide bias and variance plots in the Supplementary material. Considering the variance plots, 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 is bias. This result may, however, be impacted by the size of the re-interview sample ($n_{re} = 375$ in the present study).

4.2.1 RMSE when mode 1 is the benchmark. Figures 2 and 3 display m_1 as benchmark. Figure 2 shows the condition when a re-interview SE was introduced (+50 percent of $\bar{Y}_{r_1}$), whereas it was absent (0 percent) in the results shown in figure 3. Each separate line represents a different $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$, introduced by varying $\bar{Y}_{r_1}$ from –50 percent to +50 percent of $\bar{Y}_{r_{mm}}$. We limit the vertical axis to 0.50 (equivalent to 50 percent relative RMSE). Higher RMSE is not displayed.

The RMSE of the unadjusted estimator varied considerably across $\lambda^{(2)}$ and $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$. If $\lambda^{(2)} = 1$, RMSE was constant across $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$ at $B(\bar{Y}_{r_{mm}}^{unadj}) = P_j \mu^{(2)} = 0.5 * 0.3 = 15\%$, but RMSE was much higher, up to 40 percent, for $\lambda^{(2)} > 1$ and lower, 1 to 15 percent, for $\lambda^{(2)} < 1$. As expected, using only m_1 to estimate $\bar{Y}_{r_{mm}}$ ("Unadjusted Mode 1" estimator, equation (16)) led to bias, and a large RMSE (dominated again by its bias term).

The IREG estimator outperformed the other adjusted estimators in most cases. Whereas the estimator was unbiased (15), its variance could, however, be considerable when focal mode random error was high (i.e., at low re-interview correlations). However, IREG's RMSE fell below 10 percent for a re-interview $cor > 0.50$ and below 5 percent for $cor > 0.70$ (figure 2). Without re-interview SE, these values improved slightly (figure 3).

In a few cases, IREG was inferior to the alternative adjusted estimators. Note that MICE had small error in the trivial case when there was no SE between m_1 and m_2, $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = 0$ (cf. section 3.4). As expected, MICE performed differently than GREG, which had small RMSE if $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}})$

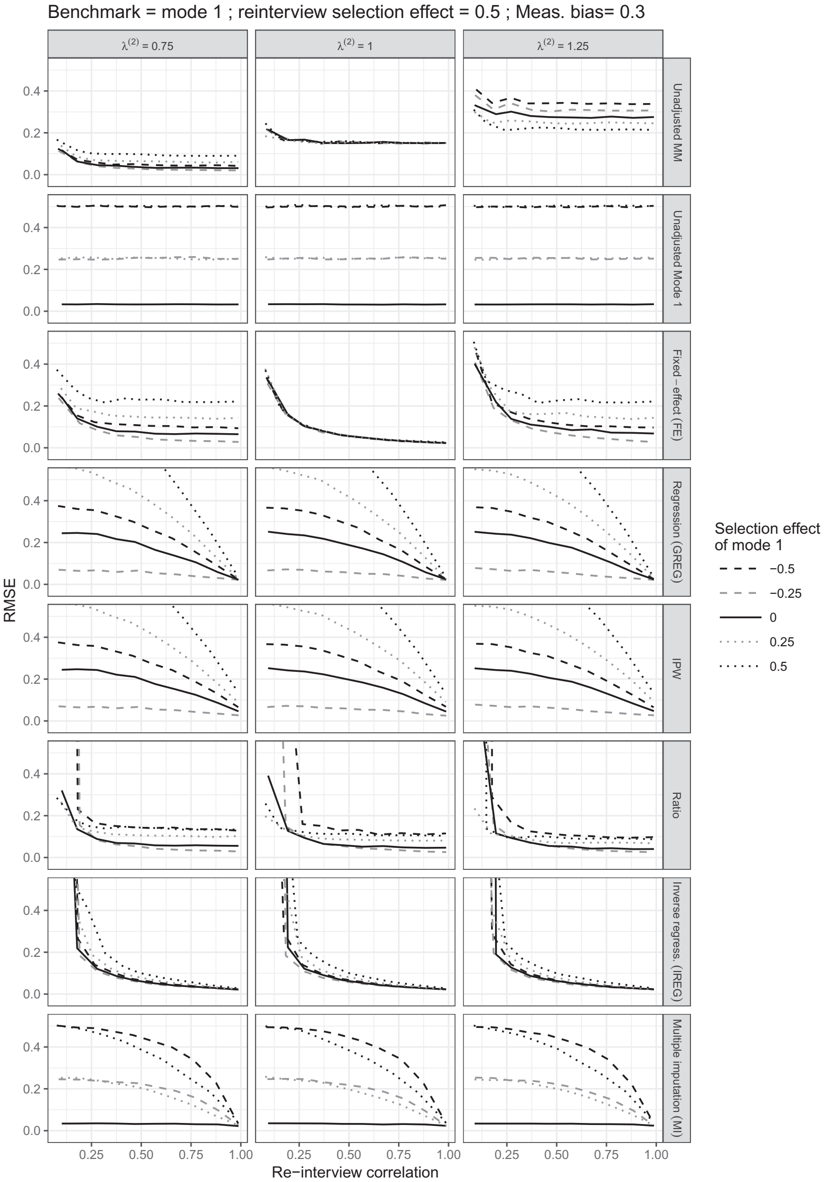

Figure 2. RMSE of adjusted and unadjusted estimators for benchmark mode $b = 1$, meas. bias $\mu^{(2)}=0.30$, and a re-interview SE of 50% relative to $\bar{Y}_{r_1}$ ($SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})=0.5\bar{Y}_{r_1}$). IREG performs best if re-interview cor >0.50, followed closely by ratio which is biased more strongly. Fixed-effect performs well under $\lambda^{(2)} = 1$ only. All others show high RMSE under some conditions.

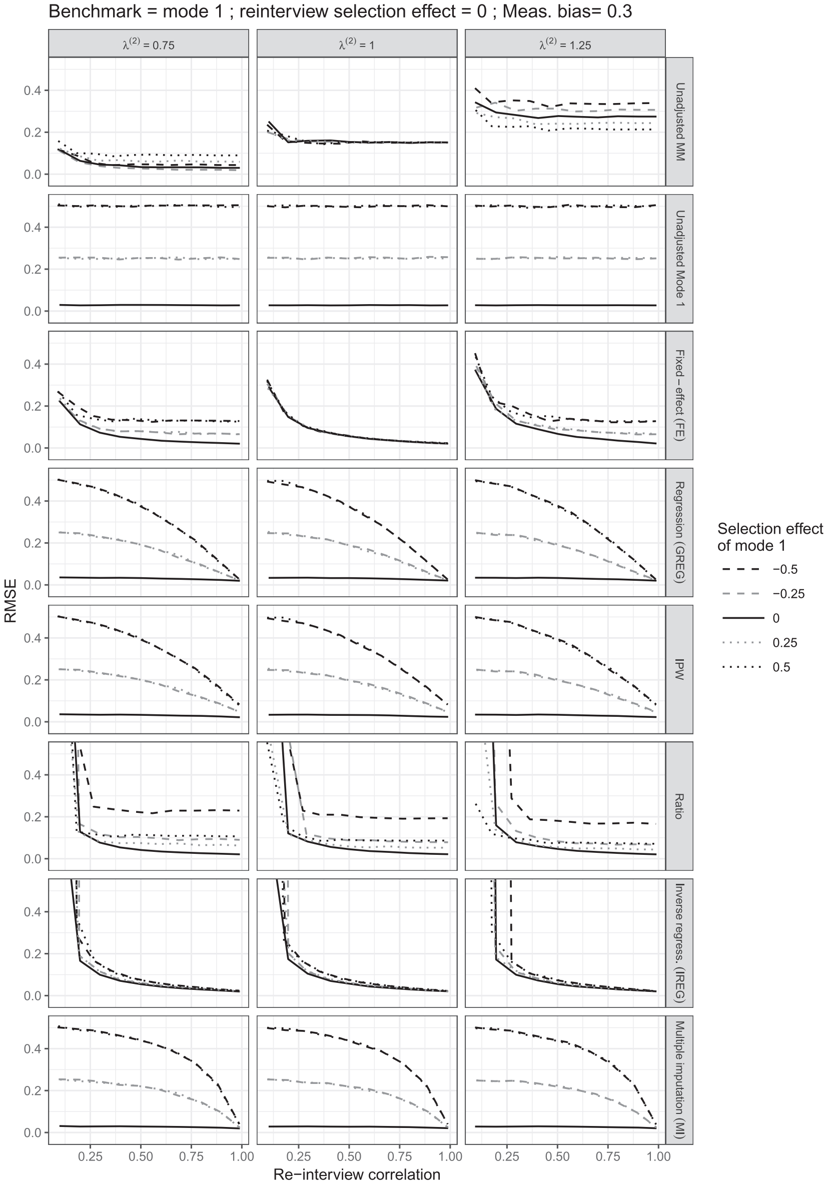

Figure 3. RMSE of adjusted and unadjusted estimators for benchmark mode *b* = *1*, meas. bias $\mu^{(2)}$=0.30, and no re-interview SE ($SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$=0). IREG performs well if re-interview cor >0.50. Ratio shows high RMSE under some conditions. Fixed-effect performs well under $\lambda^{(2)} = 1$ only. All other estimators show high RMSE under some conditions.

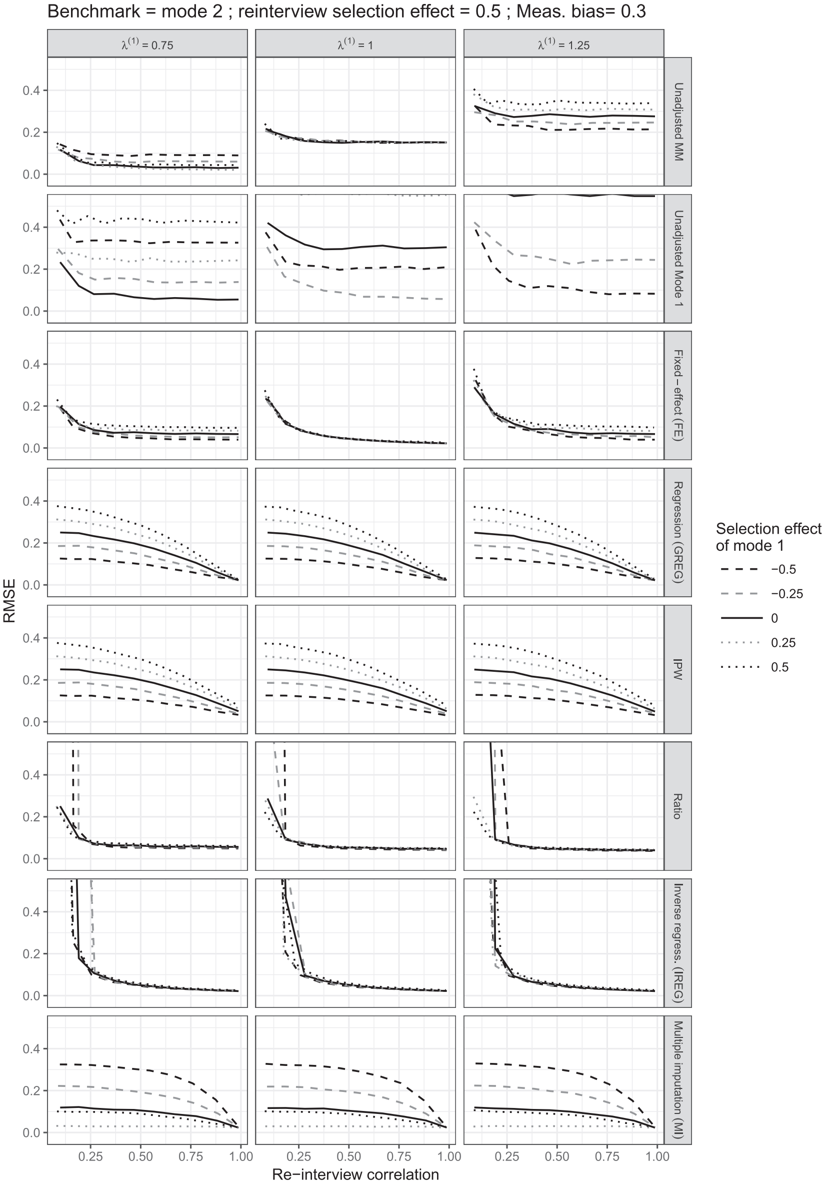

Figure 4. RMSE of adjusted and unadjusted estimators for benchmark mode $b = 2$, meas. bias $\mu^{(1)} = 0.30$, and a re-interview SE of 50% relative to $\bar{Y}_{r_1}$ ($SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) = 0.5\bar{Y}_{r_1}$). IREG, ratio and fixed-effect perform well. However, fixed-effect can only fully reduce RMSE when $\lambda^{(1)} = 1$ and ratio maintains residual RMSE on a low level.

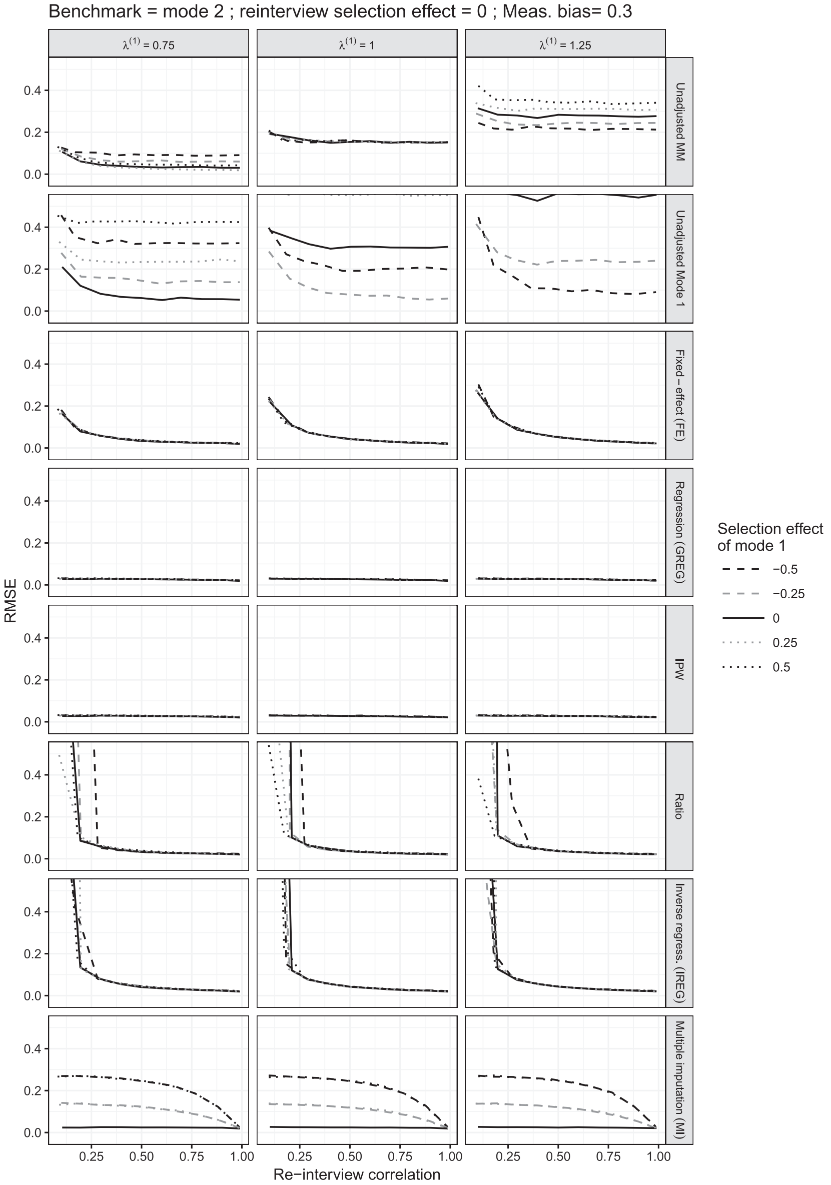

Figure 5. RMSE of adjusted and unadjusted estimators for benchmark mode $b = 2$, meas. bias $\mu^{(1)} = 0.30$, and no re-interview SE ($SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) = 0$). All adjusted estimators except MI perform well.

$= -0.25$ in figure 2 and if $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = 0$ in figure 3, as it was then approximately unbiased (13). Otherwise, GREG and MICE showed high (>10 percent) RMSE, unless re-interview correlation was very high (>0.90). RMSE of IPW followed GREG closely. The fixed-effect estimator performed well if $\lambda^{(2)} = 1$ as bias then vanishes (11). However, if $\lambda^{(2)} \neq 1$, the estimator had serious error (figure 2). Similarly, the ratio estimator had moderate RMSE when re-interview selectivity was 50 percent (figure 2), but RMSE increased drastically for 0 percent (figure 3). The ratio estimator performed slightly better if systematic term $\mu^{(2)} = 0$, since its bias then vanishes; see Supplementary material (12). However, it then had higher variance which increased further for $\mu^{(2)} = -0.30$.

4.2.2 RMSE when mode 2 is the benchmark. Figures 4 and 5 display the scenarios with and without re-interview SE for the m_2 benchmark. The IREG estimator again performed well, if re-interview correlations exceeded a moderate level (>0.40) and regardless of the size of the re-interview SE. IREG even had somewhat smaller RMSE at each level of re-interview correlation compared to the m_1 benchmark.

The bias of GREG now depends on the size of the re-interview SE (13), $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1})$, leading to biased GREG in figure 4 and no bias in figure 5 where $SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) = 0$ (the remainder being variance). IPW performed comparably. As can be seen in figure 5, both estimators outperformed IREG here, making them better alternatives if the re-interview can be considered non-selective relative to m_1. This may be achievable by design, but we judge this scenario less likely in practice (cf. section 3.1). Again MICE performed differently than GREG and IPW, as expected. MICE had low RMSE in figure 5, when, trivially, $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = SE(\bar{Y}_{r_{re}}, \bar{Y}_{r_1}) = 0$. Elsewhere MICE showed higher error, except in figure 4 when $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = 0.25$ where (14) holds approximately. MICE performed better than GREG and IREG here and also well when $SE(\bar{Y}_{r_1}, \bar{Y}_{r_{mm}}) = 0.5$.

The fixed-effect estimator performed better than under the m_1 benchmark. Variance was negligible for re-interview cor >0.35, i.e., where RMSE graphs are, roughly, horizontal. Without a re-interview SE, the fixed-effect estimator then was unbiased (11), so that RMSE approached zero (figure 5). With a re-interview SE, some residual bias remained, but it was small (<10 percent) for our choices of $\lambda^{(1)}$ (figure 4). We can quantify maximum absolute bias for this simulation at 9.38 percent; see (11). Similarly, the ratio estimator had low RMSE regardless of $\lambda^{(1)}$ in the Figures shown where $\mu^{(1)} = 0.3$. The maximum absolute bias here was 5.66 percent (for $SE(Y_{r_1}, Y_{r_{mm}}) = 0.5, \lambda^{(1)} = 0.75$; equation (12)). However, bias was larger when $\mu^{(1)} = -0.3$; see Supplementary material. In addition, variance was sensitive to random error then or when $\mu^{(1)} = 0$.

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 using re-interview data. This data is obtained from a subset of respondents to the first mode in a sequential design. The design creates partial overlap between the response distributions of both modes, which is subsequently exploited in a set of six adjusted candidate estimators. We evaluated by simulation whether any of the estimators outperform the unadjusted estimator in terms of RMSE. Earlier literature that attempts to estimate or adjust measurement effects can be criticized for potentially high bias, because researchers assumed that selection is MAR, conditional on weak auxiliary information (Vannieuwenhuyze and Loosveldt, 2013; Vannieuwenhuyze, 2015). This study is among the first to demonstrate how estimating and adjusting measurement effects in the presence of MNAR selection effects is practically feasible (see also West and Little, 2013).

The final choice of estimator depends on the analyst's expectations about the measurement error model of the focal mode, the choice of benchmark mode, and the anticipated selection effects (SEs). The inverse regression estimator (IREG) generally performed well in all considered scenarios, except when re-interview correlation was small. The IREG estimator did well even if mode-specific response depends on the target variable, which is a common situation in survey practice. The ratio and fixed-effect estimators are viable alternatives if systematic error $\mu^{(j)} = 0$ or scale parameter $\lambda^{(j)} = 1$ in the focal mode, respectively. This decision depends on the type of survey question.

Use of GREG, IPW, and MICE was shown to be problematic if there are selection effects between modes and m_1 is benchmark. Since absence of such effects is an unrealistic situation – mixed-mode surveys are designed to evoke selection effects – we recommend against these estimators unless there is additional exogenous information on which data are MAR. We considered the situation when such data are not available. If m_2 is benchmark, GREG and IPW outperform IREG if the re-interview is not selective relative to m_1 (figure 5). Careful design (e.g., long time lags) may help to justify this assumption, but we still consider it strong because the re-interview switches the mode.

Given that the analyst often has insufficient information on the type of measurement error model and selection mechanism, the IREG estimator may be the safest option in practice, but its use requires at least a moderate re-interview correlation of 0.50, or ideally 0.70, to control its variance. A mixed-mode re-interview study by Schouten et al. (2013) shows that such correlations are empirically observable.

Earlier research has studied measurement error adjustment using calibration samples similar to the re-interview sample (cf. section 2.3). This literature focusses on measurement bias of regression coefficients under MAR and finds that IREG is biased and GREG and multiple imputation are unbiased

(Freedman et al., 2008; Guo and Little, 2011, 2013). Our results differ because of a difference in estimands (regression coefficients versus means) and because these references assume MAR. Estimating means in a PMM, West and Little (2013) compared a Bayesian variant of IREG to multiple imputation and IPW in a MNAR setting. They found IREG performs best (there referred to as PMM), whereas multiple imputation and IPW performed badly unless data are MAR. These findings match ours. Contrary to us, the authors report similar performance of imputation and IPW, whereas in our results, MICE and GREG/IPW differed. This difference emerges because the estimators handle the non-monotone pattern of the re-interview design differently (cf. section 3.4).

Some limitations of the present study introduce paths for further research. First, our design focused on two modes. We note that the method can be extended for use with more modes in a relatively straightforward way. The choice of re-interview mode then becomes a central decision, however, as discussed for the cases of the Dutch Labor Force Survey and the American Community Survey in the Supplementary material.

Second, we assumed a specific measurement model conditional on which IREG is unbiased. When the model does not hold, our version of IREG may fail. Future research should, therefore, extend the theory and results about IREG to more complex measurement models (e.g., heteroscedasticity and non-linearity).

Third, we assumed measurement equivalence between the re-interview and mode two. The time lag between occasions, which may lie in the range of several weeks, is important for assuring the validity of the assumption, because longer lags increase chances of answers being forgotten. Nevertheless, there remains the possibility that re-interviews do not produce the same measurements as standard interviews (Biemer and Forsman, 1992). Future empirical research should address this question (Forsman and Schreiner, 1991).

Fourth, we assumed a measurement benchmark. Even a well-reasoned benchmark mode choice may not be optimal, especially if none of the modes in the design measures without error. It may then still be useful to adjust towards the most plausible mode, because this measure most likely still reduces total survey error.

Finally, we assumed a large sample and did not take into account the cost of the re-interview, which is determined by the size and mode of the re-interview sample. 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 research should address the trade-off between cost and efficiency and determine the most efficient subsampling designs. Further important extensions of the methodology include estimation for discrete outcomes and complex samples.

Supplementary Material

Supplementary materials are available online at https://academic.oup.com/jssam.

References

Biemer, P. P., and G. Forsman (1992), "On the Quality of Re-Interview Data with Application to the Current Population Survey," *Journal of the American Statistical Association*, 87, 915–923.

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

Brown, C. H. (1990), "Protecting Against Nonrandomly Missing Data in Longitudinal Studies," *Biometrics*, 46, 143–155.

Cochran, W. (1977), *Sampling Techniques* (3rd ed.). New York: Wiley.

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

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

Forsman, G., and I. Schreiner (1991), "The Design and Analysis of Reinterviews: An Overview," in *Measurement Error in Surveys*, eds. P. P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz, and S. Sudman, pp. 279–302, Hoboken, New Jersey: Wiley.

Freedman, L. S., D. Midthune, R. J. Carroll, and V. Kipnis (2008), "A Comparison of Regression Calibration, Moment Reconstruction and Imputation for Adjusting for Covariate Measurement Error in Regression," *Statistics in Medicine*, 27, 5195–5216.

Guo, Y., and R. J. A. Little (2011), "Regression Analysis with Covariates that have Heteroscedastic Measurement Error," *Statistics in Medicine*, 30, 2278–2294.

———. (2013), "Bayesian Multiple Imputation for Assay Data Subject to Measurement Error," *Journal of Statistical Theory and Practice*, 7, 219–232.

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

Klausch, T., J. Hox, and B. Schouten (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)*, 178(4), 945–961.

Klausch, T., B. Schouten, and J. J. Hox (2017), "Evaluating Bias of Sequential Mixed-mode Designs Against Benchmark Surveys," *Sociological Methods & Research*, 46(3), 456–489.

Kolenikov, S., and C. Kennedy (2014), "Evaluating Three Approaches to Statistically Adjust for Mode Effects," *Journal of Survey Statistics and Methodology*, 2, 126–158.

Little, R. J. A. (1994), "A Class of Pattern-Mixture Models for Normal Incomplete Data," *Biometrika*, 81, 471–483.

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

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

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

Schouten, B., J. van den Brakel, B. Buelens, J. van der Laan, and T. Klausch (2013), "Disentangling Mode-Specific Selection and Measurement Bias in Social Surveys," *Social Science Research*, 42, 1555–1570.

Shin, H. B. (2012), "2010 ACS Content Test Evaluation Report Covering Computer and Internet," *US Census Bureau* https://www.census.gov/content/dam/Census/library/working-papers/2012/acs/2012_Shin_01.pdf. Accessed August 3, 2017.

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

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

Vannieuwenhuyze, J. (2015), "Mode Effects on Variances, Covariances, Standard Deviations, and Correlations," *Journal of Survey Statistics and Methodology*, 3, 296–316.

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

West, B. T., and R. J. A. Little (2013), "Non-Response Adjustment of Survey Estimates Based on Auxiliary Variables Subject to Error," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62, 213–231.