DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 303–334, http://dx.doi.org/10.1515/JOS-2017-0016

Estimating Components of Mean Squared Error to Evaluate the Benefits of Mixing Data Collection Modes

Caroline Roberts[1] *and Caroline Vandenplas*[2]

Mixed mode data collection designs are increasingly being adopted with the hope that they may reduce selection errors in single mode survey designs. Yet possible reductions in selection errors achieved by mixing modes may be offset by a potential increase in total survey error due to extra measurement error being introduced by the additional mode(s). Few studies have investigated this empirically, however. In the present study, we compute the Mean Squared Error (MSE) for a range of estimates using data from a mode comparison experiment. We compare two mixed mode designs (a sequential web plus mail survey, and a combined concurrent and sequential CATI plus mail survey) with a single mode mail survey. The availability of auxiliary data on the sampling frame allows us to estimate several components of MSE (sampling variance, non-coverage, nonresponse and measurement bias) for a number of sociodemographic and target variables. Overall, MSEs are lowest for the single mode survey, and highest for the CATI plus mail design, though this pattern is not consistent across all estimates. Mixing modes generally reduces total bias, but the relative contribution to total survey error from different sources varies by design and by variable type.

Key words: Nonresponse error; measurement error; coverage error; sampling variance.

1. Introduction

Mixed mode data collection has been gaining popularity in survey research internationally. A number of developments working in parallel have contributed to this change in survey practice: (1) the need to find alternatives to traditional telephone surveys, due to the rapid increase in 'mobile only' households (Carley-Baxter et al. 2010; Blumberg and Luke 2013); (2) a widely reported decline in response rates (Brick and Williams 2013; De Leeuw and De Heer 2002); (3) an increase in costs associated with mitigating nonresponse (Massey and Tourangeau 2013) combined with cuts in research

[1] Institute of Social Sciences, University of Lausanne, Bâtiment Géopolis, Quartier Mouline, CH-1015 Lausanne, Switzerland. Email: caroline.roberts@unil.ch
[2] Centre for Sociological Research, KU Leuven, Parkstraat 45 – Box 3601, BE-3000 Leuven, Belgium. Email: caroline.vandenplas@kuleuven.be

Acknowledgments: This publication benefited from the support of the Swiss National Centre of Competence in Research LIVES – Overcoming Vulnerability: Life Course Perspectives, which is financed by the Swiss National Science Foundation (grant number: 51AU40-125770). The authors are grateful to the Swiss National Science Foundation for its financial assistance. We would particularly like to thank Dominique Joye (University of Lausanne) and Michèle Ernst Stähli (FORS) for their collaboration in the design and implementation of the mode experiment from which the data come. We would also like to express our gratitude to the anonymous reviewers and the editors of this special issue for their constructive feedback on earlier drafts of the manuscript. In particular, the meticulous comments and suggestions for revision of reviewer 1 helped extensively in refining the manuscript. Finally, we would like to thank Emilie Bornet and Mathias Humery at MIS Trend SA., for their careful management of the fieldwork, and their contribution to the analysis of costs.

budgets; and (4) advances in information and communication technologies increasing the opportunities for more cost- and time-efficient Internet-based data collection (Groves 2011). Mixed mode surveys that use different methods to administer questionnaires to different sample members (De Leeuw 2005; Dillman et al. 2009) have been adopted partly by necessity in response to these developments, prompting a need for research into their efficacy, and their impact on data quality.

Given these motivating factors, two common aims of mixed mode surveys are to reduce selection errors due to inadequate frame coverage and nonresponse in a single mode survey, and to reduce financial and/or time-related data collection costs. However, even if these aims are met, there is a risk that the potential benefits of mixing modes may be offset by a reduction in the accuracy of the estimates produced, due to compounding influences of the different modes used on the Total Survey Error (TSE). Differential measurement errors across modes in particular (and the need to adjust for them to improve the comparability of measurements), pose a significant risk to data quality that survey designers should take into consideration when weighing the decision about whether to mix modes and how to optimally design mixed mode surveys (Hox et al. 2017). To date, however, few studies have provided evidence as to the relative contribution to the TSE of error from different sources in different modes, and how this changes as a result of mixing modes.

In this article we compare the error properties of estimates produced by single and mixed mode surveys, and investigate the effect of mixing modes on survey errors given a fixed budget. To this end, we use data from a methodological experiment (Roberts et al. 2016) designed to compare the effectiveness of single and mixed mode data collection strategies for cross-sectional surveys with medium-length questionnaires (a completion time of around 30 minutes). We compare three survey designs: (1) a single mode mail survey; (2) a sequential mixed mode web plus mail survey; and (3) a combined concurrent and sequential mixed mode telephone (CATI) plus mail survey. To evaluate the impact of these design choices on the estimates produced, we calculate the mean squared error (MSE) of a range of variables from the questionnaire based on data available at the end of different phases of fieldwork (before and after the mode switch Designs 2 and 3). Specifically, we address the following research questions (RQs):

RQ1: Which survey design offers the lowest overall total error across a range of sociodemographic and target variables?

RQ2: What is the relative contribution to the MSE of error from different sources in each of the survey designs?

RQ3: How does the relative contribution of error from different sources vary as a function of combining modes? For example, do gains in response rates after mode switches translate into reductions in selection error, and to what extent is this offset or outweighed by increased measurement error?

The remainder of this section is structured in two parts. First, we consider the ways in which modes affect the accuracy of survey estimates, and how mixing modes can affect different sources of survey error. Then, we discuss the empirical challenges involved in detecting and measuring mode effects on different survey errors, and review evidence about the effect of mixing modes on error from different sources, and on the TSE of mixed mode estimates.

1.1. TSE and the Design of Mixed Mode Surveys

Survey design decisions are frequently taken from the perspective of the TSE paradigm, where the goal is to maximise the quality of the data collected, within the constraints imposed by the available budget (Biemer 2010). According to this approach, it has been argued that given equal budget and time constraints across different survey design options, researchers should opt for the design generating the lowest survey error across a range of variables (Biemer 2010; Biemer and Lyberg 2003). TSE is defined as the sum of errors from all possible sources that contribute to the difference between the value of an estimate based on the sample responding to a survey and the "true" value for the target population. Survey errors are sometimes categorised into *non-observational* errors (including sampling, coverage, nonresponse, and adjustment error), which affect the accuracy of inferences from the achieved sample to the population due to a failure to observe the entire population or an adequately representative sample of it; and *observational errors* (including specification, measurement, and data processing error), which affect the accuracy of inferences from responses given to the questionnaire to the true respondent characteristics of interest (Groves et al. 2009; Tourangeau 2017).

A major determinant of the TSE of estimates is the choice of data collection mode, which can influence the amount of non-observational and observational errors emanating from different sources. Notably, the choice of mode can affect (1) *coverage* error, by determining whether a population member has a chance of being selected to participate in a survey (e.g., if the sample design depends on a list of incomplete information needed to implement the survey in a particular mode (Carley-Baxter et al. 2010); (2) *nonresponse* error, because selected sample members may not have the possibility to participate in the chosen survey mode, or may be more or less willing to participate depending on the mode offered (Klausch et al. 2015a); (3) *measurement error*, because mode characteristics can influence how respondents come up with their answers to survey questions and the answers they give (Dillman et al. 2014; De Leeuw 2005); and (4) *processing error*, because, for example, noncomputerised methods of data entry and coding are more vulnerable to human error, or because interviewers may be less accurate in recording the responses given by respondents than the latter would be themselves (Groves et al. 2009). Furthermore, because data collection modes vary in terms of their associated fixed and variable costs (interviewer-administered modes being most expensive), mode choice partly determines the amount of (5) *sampling error* in statistics, because under a fixed budget constraint, a survey designer could afford to survey different sized samples using different modes (Vannieuwenhuyze 2014). This means that if the same survey were conducted using different modes of data collection the accuracy of the estimates produced would vary as a function of the amount of TSE produced by the chosen mode, and the composition of that error on each estimate would vary also (Tourangeau 2017).

The motivations for mixing modes hinge on the possibility to compensate for the error or cost disadvantages of one mode with the error or cost advantages of another (De Leeuw 2005). Mixed mode surveys typically involve combining modes in one of two different ways depending on the priorities of a given survey design. In "concurrent" mixed mode designs, sample members are either offered a choice between different ways of completing

the survey, or particular population subgroups are targeted in a different mode to the remainder of the sample, in the hope that a preferred or more accessible mode may encourage participation (Olson et al. 2012). In "sequential" mixed mode designs, the survey starts in one mode, and alternative modes are offered to nonrespondents at later stages of the fieldwork. In both types of design, the hope is that a more representative subset of sample members will participate as a result of combining modes, thereby reducing selection errors associated with noncoverage or nonresponse below what they would be if only one mode were used. Furthermore, by encouraging sampled units in sequential designs with a higher propensity to respond to participate via lower-cost modes (such as web or mail), overall costs may also be reduced and larger sample sizes (and hence, lower sampling errors) may be achieved (Hochstim 1967; Siemiatycki 1979; Lynn 2013; Vannieuwenhuyze 2014; Wagner et al. 2014).

As well as producing differentially selective samples, the fact that modes have unique measurement properties that can affect respondents' answers is well established. These are due to various method-related characteristics (e.g., the presence/absence of an interviewer, the use of visual vs. aural stimuli) interacting with question and respondent characteristics (De Leeuw 2005; Couper 2011) to produce differences in data quality, such as in the prevalence of response effects associated with satisficing (e.g., Chang and Krosnick 2009; Holbrook et al. 2003), or in the level of underreporting of socially undesirable behaviours and attitudes (Holbrook et al. 2003). Some face-to-face surveys explicitly incorporate mode switches (for all respondents) for modules of potentially "sensitive" questions which respondents answer more honestly in self-administered modes (De Leeuw 2005). However, where modes are mixed *between* respondents, two concerns arise with respect to measurement (and other observational) errors. First, differential measurement errors associated with each mode will be *compounded* in estimates, and the total contribution to the TSE from this error source may increase as a result. To the extent that any increase in measurement error offsets or outweighs any reduction in selection error achieved by mixing modes (leading to a net increase in TSE), advantages that could have been gained with a mixed mode design will be negated. Second, differential measurement errors will be *confounded* with selection errors in estimates, such that even if they do not cause an increase in the TSE (or even its measurement error component, if errors from different modes work in opposite directions – Tourangeau 2017), they will limit the possibility of making valid comparisons between subgroups surveyed in different modes (Vannieuwenhuyze et al. 2010).

Current recommendations for survey designers considering a mixed mode survey design are to address the twin risks of compounded and confounded errors at both the planning and analysis stage (Hox et al. 2017; De Leeuw and Berzelak 2017). To minimise the risk of differential measurement errors and enhance comparability across modes, for example, researchers can either opt for a *unified mode construction* approach to questionnaire design (Dillman et al. 2014), which maximises measurement equivalence by minimising differences in the way questions are asked in different modes (Hox et al. 2017; Tourangeau 2017) or an *optimal design* (or "best practices") approach (*ibid.*), which allows variation in how questions are asked in different modes to ensure estimates are obtained with the lowest possible measurement error in each mode. While the latter may be most effective at keeping the TSE of estimates from mixed mode surveys to a minimum

(*ibid.),* to enhance comparability across groups interviewed in different modes, it is recommended to use the unified mode strategy combined with a mixed mode survey design that simultaneously enables the isolation of mode-related measurement errors from selection errors, so that persistent differences in measurement may be corrected for statistically at the analysis stage (Hox et al. 2017).

1.2. Estimating the Effect of Mixing Modes on Survey Errors and Available Evidence

To evaluate the effects of different survey design features on data quality, and to make an informed choice between competing (single or mixed mode) survey designs, researchers ideally need to be able to quantify and compare the different components of the TSE likely to affect the accuracy of the estimates produced. For this purpose, Biemer (2010) advocates the estimation of the MSE. MSE is an estimate-specific measure summarising how the statistic is affected by all possible sources of observational and non-observational errors, which may manifest as variance or bias in the estimate. To calculate the MSE, it is necessary to decompose the TSE into its separate components and estimate the relative contribution to the total made by each. The problem is that in practice this is rarely feasible for researchers, as it requires "an estimate of the parameter that is essentially error free" (Biemer 2010, 826). For example, to assess measurement bias, external records can be used to assess the accuracy of respondents' self-reports (e.g., Olson 2006; Kreuter et al. 2008; Sakshaug et al. 2010; Tourangeau et al. 2010). To assess nonresponse bias, these auxiliary data are needed for both respondents and nonrespondents (e.g., Klausch et al. 2015a; Kreuter et al. 2010; Kappelhof 2013). As such data are rarely available to researchers, the potential utility of the MSE as a metric for evaluating the effects of different survey design features or for comparing whole survey systems (Biemer 1988) has not been fully exploited.

Because in a mixed mode survey, measurement and nonresponse biases are confounded, to calculate the MSE of estimates the errors associated with each mode must be decomposed separately (Vannieuwenhuyze et al. 2010). Disentangling the error components makes it possible to identify and quantify both the compounded and confounded effects of mixed mode surveys on estimates and compare them with those produced by alternative survey designs. Furthermore, as mentioned, it is a necessary step for developing suitable adjustment methods to correct for persistent measurement differences between modes so that the benefits of mixed mode surveys aimed at reducing selection error may be maximised (Hox et al. 2017). Different approaches to the problem of how to disentangle confounded mode effects on selection and measurement errors have been proposed (Hox et al. 2017; Tourangeau 2017; Vannieuwenhuyze and Loosveldt 2012). Each approach depends on the availability of auxiliary data, which may already be available to researchers – such as register data (Klausch et al. 2015a), or data from the recruitment wave of a longitudinal survey (Hox et al. 2015) – or (more likely) need to be collected separately (Hox et al. 2017). The latter could include a randomised mode experiment embedded in a mixed mode design (De Leeuw et al. 2008), a single mode follow-up of a random sample of respondents (Klausch et al. 2014; Schouten et al. 2013), or a new or existing single mode survey (ideally) conducted alongside the mixed mode survey that can serve as a benchmark (Vannieuwenhuyze et al. 2010; Vannieuwenhuyze and Loosveldt 2012). With such data available, it is possible to estimate the effect of mode on

selection (e.g., by predicting response propensity in one mode compared to another), and then estimate the effect of mode on measurement while controlling for the selection effect (see Hox et al. (2017) for a description of alternative techniques).

The specific data requirements for disentangling mode effects in mixed mode surveys (and estimating MSEs) has meant that despite a large and often unwieldy literature on measurement differences, relatively few studies have been able to adequately deal with the confounding problem (*ibid.*). As a result, the available evidence as to the effects of mixing modes on different error sources, as well as on the TSE, remains somewhat inconclusive. In relation to selection errors, for example, several studies have analysed response rates and sample composition as proxies, and have confirmed that both can improve in mixed mode designs compared to single mode designs, depending on which modes are combined and how (e.g., Dillman et al. 2014; Eva et al. 2010; Fowler et al. 2002; Greene et al. 2008; Link and Mokdad 2006; Millar and Dillman 2011; Lynn 2013; Klausch et al. 2015a). However, few studies have attempted to estimate the magnitude of selection errors before and after switching modes, while controlling for measurement differences. One recent study using a combination of data from population registers, a single mode benchmark, and a re-interview design (Klausch et al. 2015a) observed an increase in response rates as a result of mixing modes sequentially (a face-to-face follow-up of web, mail and CATI surveys), but found no consistent reduction in selection errors present in estimates from the starting modes. Selection error was, however, reduced for some estimates from the mixed mode surveys as a result of bringing it closer in line with the selection error of the single mode (face-to-face) benchmark (*ibid.*).

In relation to measurement errors, research using appropriate methods to control for selection effects (e.g., Kreuter et al. 2008; Chang and Krosnick 2009; Heerwegh and Loosveldt 2011, Gordoni et al. 2012; Klausch et al. 2013) has found that differences between modes do persist once efforts to control for selection errors have been applied, particularly between self- and interviewer-administered modes (Hox et al. 2017). However, these studies have focused mainly on between-mode comparisons rather than on the cumulative effects on measurement error of combining data from different modes. Similarly, efforts to compute the MSE of survey estimates (e.g., Groves and Magilavy 1984; Peytchev et al. 2009) have tended to focus on single mode scenarios, even where mixed mode data were available (Kreuter et al. 2010; Olson 2006). These studies confirm that the MSE varies considerably by estimate, and changes as a result of efforts to reduce nonresponse. Meanwhile, the one study (to our knowledge) that has considered the relative difference in the MSE for a mixed mode design compared to alternative single mode benchmarks (Vannieuwenhuyze 2014), did not consider the effects of mode mixing on the separate components of MSE across multiple variables. The present study, therefore, addresses this gap in the literature.

2. Data and Methods

2.1. Data

Our analysis uses data from a mode experiment conducted in the French-speaking region of Switzerland during Autumn/Winter 2012–2013 (see Roberts et al. 2016 for details), designed to investigate errors and costs associated with conducting surveys with different

combinations of data collection modes (including CATI, web, and mail). The experiment was embedded within a survey on personal and social wellbeing. The population for the study was adults aged 15 years and over, registered as resident in French-speaking municipalities. The research was able to benefit from a simple random sample of eligible individuals supplied by the Swiss Federal Statistical Office (SFSO), drawn from their sampling frame based on population registers maintained by municipalities, which offers very high coverage of the (legally) resident population in Switzerland (Lipps et al. 2015), and also provides auxiliary sociodemographic data about the sample (described below). From the gross sample supplied, smaller random samples were drawn and randomly assigned to the experimental treatment groups, which varied according to the starting mode they were assigned to and subsequent procedures used to reduce nonresponse. The different treatments provide opportunities to compare different types of single and mixed mode survey design.

In the present study, we compare three survey designs: Design 1, a single mode mail survey, Design 2, a sequential mixed mode web plus mail survey, and Design 3, a combined concurrent and sequential mixed mode survey, consisting of a CATI plus mail follow-up of sample units for which publicly listed (in the Swisscom directory) fixed line telephone numbers were available and supplied by the SFSO with the sample (no additional procedures were used to obtain unlisted numbers to reduce the noncoverage rate), and a mail survey of sample members for whom telephone numbers were not supplied. The three designs are shown in Figure 1.

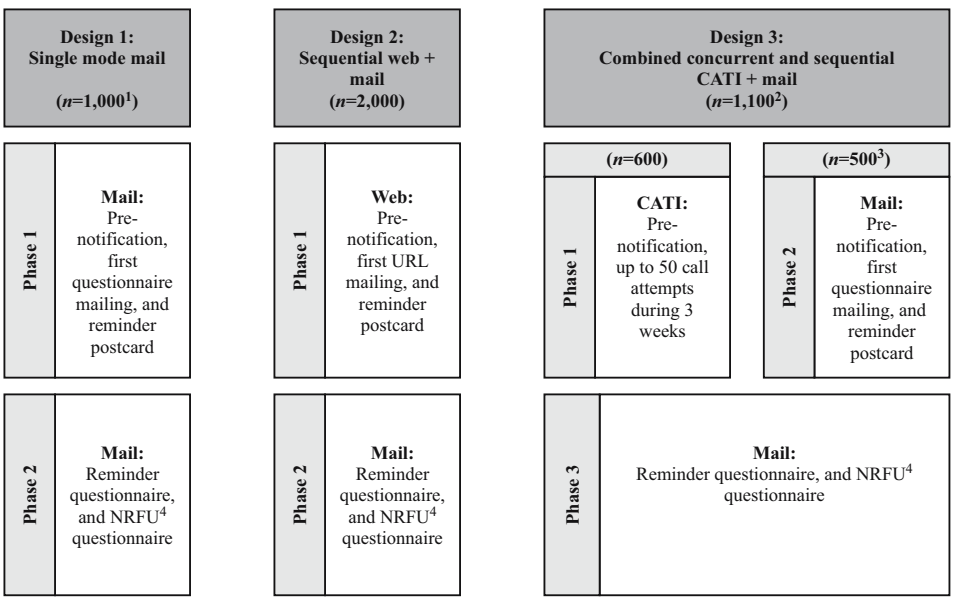

Fig. 1. Survey designs considered.
Notes: [1]*Consists of 500 sample units with known telephone numbers and 500 units without known telephone numbers.* [2]*Consists of 600 sample units with known telephone numbers and 500 sample units without known telephone numbers.* [3]*The 500 units in the mail condition of Design 3 are the same 500 sample units without a (known) telephone number used in Design 1.* [4]*NRFU respondents are only considered for the purpose of estimating bias in the target variables.*

Mainly for practical and budgetary reasons, the sample sizes in the original experiment differed between the treatment groups. In addition, the samples included an overrepresentation of units without a known telephone number. The purpose of this was to facilitate an analysis of coverage error in CATI surveys and the characteristics of units without publicly listed fixed line telephone numbers. The proportion of the gross sample supplied by the SFSO for which listed fixed line telephone numbers were unavailable was 41.2% (the noncoverage rate if the frame were used for a CATI survey and no additional efforts were made to find unlisted numbers). Design 1 included 1,000 cases (500 with telephone numbers and 500 without); Design 2 included 2,000 cases (1,000 with telephone numbers and 1,000 without); and Design 3 included 1,100 cases (600 with telephone numbers and 500 without). Note that the latter 500 cases without telephone numbers analysed in Design 3 are the same 500 cases without telephone numbers analysed in Design 1. We use design weights in all our analyses to adjust for differential inclusion probabilities for units with and without known telephone numbers in each of the survey designs (the weights for the cases without telephone numbers in Designs 1 and 2 were, therefore, equal before poststratification weighting – see below).

Sample members in each survey design were sent a pre-notification letter to inform that they had been selected for the study and would shortly be contacted either by a telephone interviewer (Design 3), or by mail (Designs 1 and 2) with further instructions on how to participate. All sample members received an unconditional incentive of ten Swiss Francs (USD 10) in cash, which for the CATI group in Design 3 was included with the pre-notification, and for the other groups was sent in the second letter, together with the paper questionnaire (for the mail groups in Designs 1 and 3) or the URL and login details (for Design 2). A reminder/thank you postcard was sent to all sample members (including respondents) assigned to web and mail mode one week later. In Design 3 (CATI group), interviewers made up to 50 contact attempts over the course of a three-week period, with instructions to vary the days of the week and timing of calls to limit noncontacts. At the end of the CATI fieldwork, and two weeks after the postcard reminder in the mail and web groups, all nonrespondents in all three surveys were sent a reminder letter together with the paper questionnaire. One month following the end of the fieldwork period, nonrespondents from all surveys (except for office refusals and cases where addresses were found to be invalid) were additionally sent a reduced length "nonresponse follow-up" (NRFU) questionnaire by mail, the data from which we make use of here in our analysis of errors in target variables (described later).

For the purpose of our analyses, we distinguish two main phases of fieldwork in Designs 1 and 2, Phase 1 consisting of all mailings up to and including the postcard reminder, and Phase 2 consisting of the mailing of a reminder questionnaire and the NRFU questionnaire. For Design 3, we distinguish three phases, to assess the effect of adding the concurrent mail survey of sample units with no known telephone number independently of the CATI survey of sample units with known telephone numbers. Thus, in Design 3, Phase 1 refers to the CATI fieldwork, Phase 2 refers to the concurrent mail survey (equivalent to Phase 1 in Design 1), and Phase 3 refers to the mailing of the reminder questionnaire and the NRFU questionnaire to nonrespondents in both the phone and no-phone groups (see Figure 1).

The questionnaire for the survey included around 125 items, with mean CATI and web administration times of 25 minutes. About one third of the questions were measures of wellbeing. Another third were sociodemographic measures, and the remainder were questions on society in general. Data collection was carried out by the survey agency, M.I.S. Trend SA. Fieldwork started on the 22 November 2012, and was completed by 8 March 2013.

2.2. *Analytic Approach*

Our analysis is in two parts. First, we compare estimates from each of the survey designs for a range of variables to benchmark estimates to assess the total absolute error (RQ1), using two different approaches depending on the benchmark data available. We start by looking at estimates of sociodemographic characteristics of the sample, comparing self-reported characteristics with auxiliary data from the sampling frame to assess the total error in each. Then, we extend our analysis by looking at a set of target substantive variables from the questionnaire, using Design 1 as a benchmark against which to compare estimates from Designs 2 and 3, while applying poststratification weights based on auxiliary data from the sampling frame (specifically, the weighting model includes the variables age, marital status, country of birth, household size, and urbanisation).

To identify which survey design offers the lowest overall total error across a range of sociodemographic and target variables, we calculate two summary statistics of the absolute error: a) Cramer's V (following the approach used by Klausch et al. 2015a), and b) the MSE (as described below). Cramer's V provides a measure of the degree of correspondence between the survey data and the benchmark data. Based on Pearson's χ^2 statistic, it summarises the strength of the absolute deviations from independence (no selection error) for all categories of a nominal variable and determines whether there is a significant difference between the expected frequencies (provided by the benchmark) and the observed frequencies (provided by the survey) across one or more categories. Cramer's V renders the χ^2 statistic comparable across a variety of variables by scaling it to the interval of 0 and 1, which additionally provides a way of interpreting the effect size, which further facilitates comparisons (*ibid.*, 951). The values 0.10–0.30 indicate small absolute error, 0.30–0.50 indicate moderate absolute error, and values of 0.50–1.00 indicate large absolute error (*ibid.*). We calculate the V statistic for a) the sociodemographic variables from the register; and b) target questionnaire variables. We also compute the average of the error estimates for the two sets of variables. This allows us to examine the overall systematic effect of the different modes and mode combinations in each survey design on TSE for the two types of variable, and avoid some of the difficulties of interpreting inconsistent findings across variables, which are typically attributable to variable content, rather than the design of the survey (or some interaction between the two) (*ibid.,* 952). We used Rao-Scott chi-square tests, which is a design-adjusted version of Pearson's chi-square, and is suitable for selection probability weighted data (Rao and Scott 1987).

In the second part of our analysis, we estimate the following principal components of the total error of both the register and target variables: sampling variance, and noncoverage, nonresponse, and measurement bias, and use these components to calculate the MSE for each variable for each of the three survey designs, and to assess the relative

contribution to the MSE of error from different sources (RQ2). Finally, to assess the effect of mixing modes on the relative contribution to the total bias of each bias component (RQ3), we consider the relative contribution to the total bias made by selection and measurement bias following each fieldwork phase (Designs 2 and 3 only). We describe the procedures we use for estimating the bias in detail in the next section.

2.3. *Components of MSE Analysed*

Biemer and Lyberg (2003, 59) identify six major components of MSE, each of which poses to varying degrees a risk of variable and systematic error in survey estimates, including 1) specification error, 2) frame (coverage) error, 3) nonresponse error, 4) measurement error, 5) data processing error, and 6) sampling error. We do not consider all of these error types here, but, as mentioned, restrict ourselves to an analysis of sampling variance, and noncoverage (for a CATI survey based on listed, fixed line numbers), nonresponse, and measurement bias. We compute MSE as the sum of the sampling variance under each survey design and the square of the bias, using different procedures to estimate the bias for the sociodemographic and substantive variables. Specification, frame and nonresponse error are generally considered to pose low risk of variable error (*ibid.*), and for this reason we do not consider their contribution to the variance. Furthermore, as sampling error is considered to pose a low risk of systematic error (*ibid.*), we focus on the variable error component. We do not consider specification errors, as these were the same across all the survey designs (and are assumed to be small as most of the survey questions had been extensively pretested and fielded in two rounds of the European Social Survey). Neither do we separately consider data processing errors, which may have affected the quality of the data from the mail survey (which were entered manually), and are subsumed here within the estimates for measurement bias. Note that Groves and his colleagues (2009) additionally identify adjustment error as a separate contributor to non-observational errors in the TSE, while Biemer and Lyberg (2003) include adjustment error as part of post-survey data processing errors more generally (see also Biemer 2010). We do not estimate the error from the weighting adjustments we use here (design and poststratification weights), but it is important to note that part of our measurement and nonresponse error estimates for all three designs may be attributable to adjustment error (along with the other processing errors mentioned).

As we do not have repeated measurements to allow us to decompose the measurement error into bias and variance, we focus on measurement bias. Some of the substantive variables we analyse might be considered sensitive (e.g., measures of subjective wellbeing, measures of attitudes towards immigration), so the extent to which they are affected by social desirability bias might be expected to vary between interviewer- and self-administered modes (Holbrook et al. 2003). For sociodemographic variables, measurement error is more likely to take the form of classification errors (Biemer 2010) and is expected to be minimal. However, discrepancies between the register data and self-reports may also appear if somebody other than the named individual in the sample responded to the survey, which is more likely to occur in the web and mail groups, or due to data input errors. Both these error types are subsumed in the estimates of measurement bias.

To recap, we focus on the following components of MSE: noncoverage bias (B_{NC}) (in Design 3 only), nonresponse bias (B_{NR}), measurement bias (B_{MEAS}), and sampling variance (Var_{SAMP}). These components of bias are summed and squared to produce the total bias component of the MSE, then added to the sampling variance to obtain an estimate of the MSE, as follows:

$$MSE = (B_{NC} + B_{NR} + B_{MEAS})^2 + Var_{SAMP}$$

Given that bias from different sources varies by mode of data collection, we calculate the bias separately for each of the modes in the survey design and combine them additively, to see whether they compound or offset one another. Thus, MSE is decomposed further for the mixed mode survey designs, as follows:

Design 2:

$$MSE = ((B_{NC} + B_{NR} + B_{MEAS})_{WEB} + (B_{NC} + B_{NR} + B_{MEAS})_{MAIL})^2 + Var_{SAMP}$$

Design 3:

$$MSE = ((B_{NC} + B_{NR} + B_{MEAS})_{TEL} + (B_{NC} + B_{NR} + B_{MEAS})_{MAIL(phase\ 2)} + (B_{NC} + B_{NR} + B_{MEAS})_{MAIL(phase\ 3)})^2 + Var_{SAMPTEL} + Var_{SAMP\ MAIL}$$

2.4. *Calculating Bias*

For the sociodemographic variables, the calculation of bias is made possible by comparing different estimates derived from the sampling frame and survey data. These include: (1) the *sample register* estimate, which is the estimate based on the register data for each of the random samples randomly assigned to the three survey designs; (2) the *respondents' register* estimate, which is the estimate for the responding sample in each survey based on the register data; and (3) the *self-report* estimate, which is the estimate for the responding sample based on answers to the survey questions. For each one, we use design weights to correct for differential inclusion probabilities based on the availability of telephone numbers on the frame for sample members. For Design 3, we additionally compute (4) the *register coverage* estimate (for the CATI group only), which is the estimate based on the register data for the sample with known telephone numbers. We produce estimates based on the (cumulative) sample responding following each phase of fieldwork.

On the basis of these estimates, we compute bias for the sociodemographic variables as follows:

1. *Total bias = self-report estimate – sample register estimate*
2. *Noncoverage bias = register coverage estimate – sample register estimate*
3. *Nonresponse bias = respondents' register estimate – sample register estimate (or register coverage estimate for Design 3, Phase 1)*
4. *Measurement bias = self-report estimate – respondents' register estimate*

For the target variables, we calculate total bias by comparing estimates based on self-reports with estimates from the single mode mail benchmark survey (Design 1). To

decompose the measurement error from the selection error, we use a "MM calibration approach" (Vannieuwenhuyze and Loosveldt 2012, 87), in which we attempt to control for selection effects in the different survey designs to render the samples as comparable as possible to the benchmark survey, so that any remaining differences may be assumed to be caused by measurement effects (*ibid.*). Specifically, we apply poststratification weights based on auxiliary (sociodemographic) data from the sampling frame. The poststratification weights are used to adjust the response samples to the distribution of the auxiliary variables on the sample frame, after which we derive adjusted and unadjusted estimates based on the sample responding after each phase of fieldwork (as for the sociodemographic variables). Peytchev and his colleagues (2011) describe this approach as suitable for the given purpose. However, it should be noted (as previously mentioned) that the adjustment procedures used are themselves not free from error, and the effectiveness of such weighting procedures is limited by the availability of suitable auxiliary data (Hox et al. 2017). Thus, the bias we observe from both measurement and nonresponse in the target variables may partly be due to this limitation of the methods we use (a limitation we discuss further in Section 4).

The poststratification weights were computed by multiplying the design weight by the inverse response propensity score. Response propensity scores were estimated by a logistic regression equation including the following covariates: age, marital status, country of birth, household size, and urbanisation. In addition, the interaction terms age*marital status, marital status*country of birth, and marital status*urbanisation were added to improve model fit. Propensity scores were calculated separately for each of the survey designs and for sample members with and without telephone numbers.

For this part of our analysis, we restrict ourselves to target variables that were additionally included in the reduced-length NRFU questionnaire used in the original mode experiment. This allows us to add data from the NRFU respondents to compute the unadjusted self-report estimates before applying poststratification weighting, and thereby, reduce variation in the nonresponse adjustment weights and adjustment error. The motivation is to try to obtain the 'best possible' estimate from each survey design, with the least possible nonresponse bias, to compare against the benchmark. Note that it was not possible to use the same procedure to analyse the sociodemographic variables, as not all were included as questions in the NRFU questionnaire, nor was this necessary given the availability of register data. This means that the number of observations available in Designs 2 and 3 for the analysis of the target variables was slightly larger than for the sociodemographic variables and that the bias and variance estimates differ accordingly. Similarly, the poststratification weights used to analyse the target variables were slightly different to those used to analyse the sociodemographic variables. In any case, the number of NRFU respondents in all three designs was small – in Design 1 it was 50, in Design 2 it was 61, and the number in Design 3 was 64.

Total bias for the target variables is calculated by subtracting the unadjusted estimates from Designs 2 and 3 from the adjusted estimate from Design 1. Nonresponse bias is calculated by subtracting the unadjusted estimate from the adjusted estimate from each design. Measurement bias is calculated by subtracting the nonresponse bias from the total bias. Additionally, noncoverage bias is estimated for Design 3 by subtracting the Design 1 adjusted estimate for the sample with telephone numbers from the Design 1 adjusted

estimate for the full Design 1 sample. Note an additional limitation of our procedures is that we do not adjust for measurement (and/or processing) errors in our estimates of selection errors in the target variables.

2.5. Comparing Sampling Variance Across Survey Designs

The surveys under consideration have different sample sizes and different costs associated with them. The sample size being one of the major factors that influences the sampling variance, we needed a criterion to standardize the (responding) sample sizes across the survey designs. As a criterion, we chose the total cost to obtain the responding sample. Therefore, we computed the net sample size given a fixed budget constraint – in this case, USD 100,000. To do this, we make use of the cost data provided in Table 1 (which are based on calculations made by the fieldwork agency based on the budget agreed with the client for the fieldwork contract – i.e., they do not represent the actual costs to the survey agency). First we subtracted the fixed costs of each survey design from the budget (for the mixed mode surveys we added the fixed costs of the mail survey to the fixed costs for the starting mode for each design), and divided the remaining budget by the variable costs per sample member, which for respondents varied depending on which phase of the survey they responded in. This makes it possible to adjust the variance component of the MSE estimates to render them comparable across the three surveys. Based on the assumption that the bias component of the MSE would be unaffected by the size of the starting sample, and that the response rates achieved under a given design in the present study would not

Table 1. Unit costs of the survey designs (in USD).

	Design 1: Single mode	Design 2: Sequential	Design 3: Combined concurrent and sequential
	Mail	Web + mail	CATI + mail
Fixed costs[1]:	16 460.76	14 954.77	25 269.03
Variable costs per:			
Phase 1 respondent	22.29	15.75	77.75
Phase 2 respondent	25.71	23.40	22.29
Phase 3 respondent (CATI group)	–	–	20.83
Phase 3 respondent (Mail group)	–	–	25.71
Nonrespondent[2]	20.32	17.89	17.86/20.32
Sample member[3]	22.28	18.04	39.64
Total	22 278.52	36 076.03	43 599.48
Net sample size for USD 100k	2,493	2,449	979

Notes. [1]Does not include the fixed costs of the mail survey, which were added to the fixed costs for Designs 2 and 3 to compute the net sample size for the adjusted sampling variance. [2]Assumes nonrespondents receive maximum contacts under given survey design (not the case for office refusals). Variable costs for NRFU respondents were USD 1.80 higher, but NRFU respondents were not included in the calculation of the costs per randomly drawn sample member and net sample size. [3]Cost per randomly drawn sample member (includes USD 10 unconditional incentive).

change in a larger scale study, this allows us to compare the MSE of the different survey designs.

All survey estimates and their standard errors were calculated using design-weighted data with the "proc surveyfreq" and "proc surveymeans" procedures in SAS 9.3. These procedures rely on the Tailor Series Method to estimate the size of the sampling error in case of complex sampling designs (in this case, the oversampling of people without publicly listed telephone numbers).

2.6. *Variables Analysed*

The sociodemographic variables for which both self-report and register data were available were: respondent sex, age in years, marital status (single, married, widowed, divorced), country of birth (Switzerland, bordering countries, non-bordering countries), household size (number of persons), and availability of a fixed line telephone number for the sample member (listed in the Swisscom directory, unlisted, no fixed-line number available). Note that no self-report for this latter variable was available for the CATI group in Design 3 as all respondents were interviewed on their listed, fixed line telephone number.

The target variables analysed were: social trust, life satisfaction, happiness, frequency of feeling stressed in the past month, frequency of feeling depressed in the past week, self-rated health, interest in politics, support for immigration, and self-evaluation of income adequacy. Full details of question wording are available as supplemental material online (available at: http://dx.doi.org/10.1515/JOS-2017-0016)

3. Results

The results are presented as follows. First we address the question of which survey design offers the lowest overall total error (RQ1) by presenting estimates of the total error in the sociodemographic and target variables (Cramer's *V*) and the MSE. Then, we look at the relative contribution of different sources of error to the MSE in each of the survey designs (RQ2). Finally, we address the question of how the relative contribution to the Total Bias (TB) of bias from different sources changes as a function of mixing modes (RQ3). This allows us to assess the extent to which any reductions in Selection Bias (SEB) are offset by increased Measurement Bias (MEB). Before proceeding to the research questions, we first present the response rates for each of the survey designs.

3.1. *Response Rates*

Response rates, calculated as the number of completed interviews divided by the sample size (all sample members were considered eligible), were very similar across the three survey designs (see Table 2 for both unweighted and inclusion-probability weighted response rates). Design 1 obtained an overall (weighted) response rate of 66.2% (57.4% following Phase 1). Design 2 obtained a (weighted) response rate of 44.7% following Phase 1 (web), and 64.9% following Phase 2 (mail). In Design 3, the (weighted) response rate following Phase 1 (CATI) was 35.7% of the total Design 3 sample. Following Phase 2

Table 2. Unweighted and weighted response rates by survey design.

	Unweighted response rates (%)	Weighted response rates (%)	Sample size (n)
Design 1 ($n = 1{,}000$)			
Phase 1 respondent (mail)	56.5	57.4	565
Phase 2 respondent (mail)[1]	8.9	8.8	89
Total	**65.4**	**66.2**	**654**
Design 2 ($n = 2{,}000$)			
Phase 1 respondent (web)	44.5	44.7	889
Phase 2 respondent (mail)	19.9	20.2	399
Total	**64.4**	**64.9**	**1288**
Design 3 ($n = 1{,}100$)			
Phase 1 respondent (CATI)	33.1	35.7	364
Phase 2 respondent (mail)	23.4	21.1	257
Phase 3 respondent (phone)	5.2	5.6	57
Phase 3 respondent (no phone)	4.2	3.8	46
Total	**65.8**	**66.2**	**724**

Notes. [1]Phase 2/3 and total response rates do not include respondents to the NRFU questionnaire, data from whom are used in the estimation of bias in the target variables. The number responding to the NRFU in each design were: Design 1 ($n = 50$), Design 2 ($n = 61$), and Design 3 ($n = 64$).

(concurrent mail phase), the response rate was 56.8%, and following Phase 3 (sequential mail phase), it was 66.2%.

3.2. Absolute Error

Overall, the absolute error in the sociodemographic variables was small, as indicated by values of Cramer's V (shown in Table 3) only exceeding 0.10 for one variable in the web phase of Design 2 (age), and for two variables in Design 3 (country of birth in the CATI phase, and having a registered fixed line telephone number in all three phases). Due to space limitations, we do not describe the nature of the errors here. Interested readers can refer to Tables A1 and A2 (sociodemographic variables) and A3 and A4 (target variables) in the supplemental material online to interpret the effects (available at: http://dx.doi.org/10.1515/JOS-2017-0016). This latter variable had the largest absolute error in Phase 1 of the CATI survey (0.402, which can be interpreted as a moderate effect size), reflecting the noncoverage error present in Phase 1 of Design 3. At the end of all three phases, the absolute error on this variable was reduced to 0.153. Mixing modes in Design 2 was similarly effective for reducing the absolute error on age. Although the absolute errors were generally small, the chi-square tests revealed significant differences between estimates based on self-reports and estimates based on register data for three of the sociodemographic variables in all three survey designs, which persisted after all phases of fieldwork: country of birth, household size (though only at the ten per cent level in survey 1), and having a listed fixed line telephone number.

In the target variables (shown in the lower half of Table 2), the absolute error was generally even smaller than that for the sociodemographic variables, never exceeding 0.10 for Designs 1 and 2, and only just doing so for three variables in the CATI phase of Design

Table 3. Absolute error on sociodemographic variables from the sampling register and key target variables (Cramer's V).

	Design 1: Single mode mail		Design 2: Sequential web + mail		Design 3: Combined concurrent and sequential CATI + mail		
Variables	Phase 1	Phases 1 + 2	Phase 1	Phases 1 + 2	Phase 1	Phases 1 + 2	Phases 1 + 2 + 3
Gender	0.030	0.029	0.016	0.028	0.034	0.054*	0.037
Age	0.037	0.018	0.130***	0.047	0.097*	0.024	0.020
Marital Status	0.066§	0.056	0.054*	0.040	0.040	0.025	0.025
Country of birth	0.072*	0.073*	0.093***	0.085***	0.123***	0.064*	0.070**
Household size	0.059§	0.060§	0.066**	0.057**	0.061§	0.071**	0.069**
Telephone number	0.083**	0.066*	0.069***	0.072***	0.402***	0.189***	0.153***
Health	0.001	0.011	0.065**	0.022	0.051§	0.043	0.062*
Interest in politics	0.027	0.006	0.026	0.015	0.038	0.011	0.002
Immigration attitude	0.016	0.012	0.023	0.002	0.013	0.006	0.010
Income evaluation	0.030	0.019	0.079***	0.053*	0.115***	0.048§	0.039
Social trust	0.013	0.009	0.017	0.027	0.065*	0.041	0.038
Life satisfaction	0.025	0.013	0.048*	0.032	0.104***	0.053**	0.051*
Happiness	0.012	0.015	0.031	0.016	0.125***	0.070*	0.067§
Stress	0.023	0.014	0.070**	0.043§	0.096***	0.085**	0.074**
Depression	0.003	0.005	0.008	0.006	0.010	0.006	0.014

Notes. ***$p < 0.001$. **$p < 0.01$. *$p < 0.05$. § $p < 0.10$. *P*-values derived from Rao-Scott Chi-square tests of association.

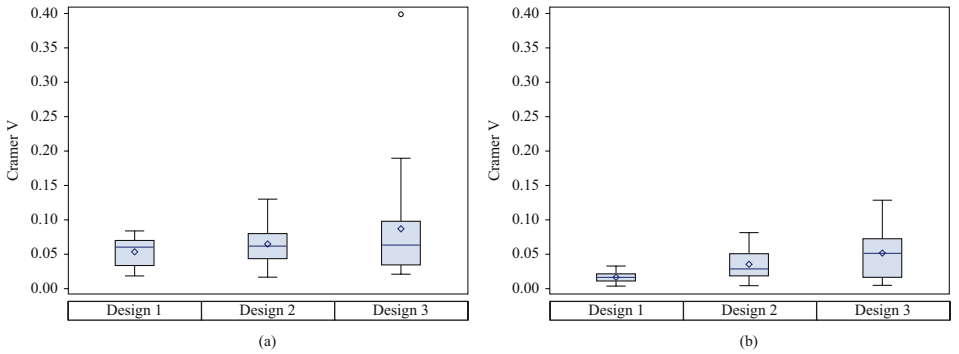

Fig. 2. Distribution of Cramer's V statistics (see Table 3) measuring absolute deviations from the benchmarks for the three survey designs: (a) sociodemographics; (b) target variables.

3 (income evaluation, life satisfaction, and happiness). Following all three phases of Design 3, the values for Cramer's V for these variables were reduced to below 0.10. The chi-square tests revealed that following all phases in Design 2, a statistically significant difference (at the five per cent level) compared to the benchmark remained on only one variable: income evaluation. Following all phases in Design 3, estimates for three variables were statistically significantly different from those in the mail survey: the proportion in good health, the proportion satisfied with their life, and the proportion feeling stressed.

Summarising across both sets of variables, Design 1 had the lowest absolute error on both the sociodemographics and the target variables, followed by Design 2 and then Design 3. This pattern of results is depicted in Figure 2, which shows the empirical distribution of the Cramer's V values for each design across the two sets of variables shown in Table 3 in the form of boxplots, where the mean values are represented by diamonds, and the median values by the bars. For the sociodemographic variables (left-hand side of Figure 2), the median values for Cramer's V are quite similar for all three survey designs, but the mean value is lowest in Design 1 and highest in Design 3. The interquartile ranges of the V statistics for Designs 1 and 2 are more similar, while it is wider for Design 3, and the full range of values for Designs 2 and 3 is wider than for Design 1, indicating stronger variance across the variables. Turning to the target variables (right-hand side of Figure 2), we see a similar pattern, though the absolute error, as noted, is lower overall than for the sociodemographics. Estimates based on Design 3 vary most from the benchmark, as reflected in a slightly higher median value for the V statistics, the higher mean value, and the larger range of values overall.

3.3. Mean Squared Error

The MSE estimates for the three survey designs, which are displayed in Table 4, along with the TB and the Sampling Variance (SV), mirror the above findings. On average, the adjusted MSE, which is based on the net sample size obtainable under a fixed budget of USD 100,000, is highest for Design 3 on both types of variable though largest for the sociodemographic measures than for the target variables (32.71 for the demographics and 15.51 for the target variables). Design 1 has the lowest average MSE on both types of

Table 4. Total bias, sampling variance and MSE for register variables and target variables by survey design after all fieldwork phases.

	Design 1: Single mode mail			Design 2: Sequential web + mail			Design 3: Combined concurrent and sequential CATI + mail		
	Total bias	Sampling variance	MSE	Total bias	Sampling variance	MSE	Total bias	Sampling variance	MSE
Register variables:									
Male (%)	−2.91	1.00	9.49	0.28	1.06	1.14	−1.95	2.54	6.34
Aged 15–24 (%)	0.39	0.41	0.56	2.21	0.56	5.43	1.40	1.28	3.25
Aged 65+ (%)	−1.27	0.67	2.28	−2.64	0.68	7.63	−0.58	1.48	1.82
Married (%)	2.65	1.00	8.01	0.08	1.06	1.07	1.48	2.56	4.76
Born in Switzerland (%)	6.28	0.88	40.32	8.20	0.89	68.16	6.51	2.19	44.58
1 person household (%)	−3.23	0.57	10.99	−2.39	0.61	6.31	−2.99	1.47	10.39
Listed phone number (%)	1.74	0.95	3.98	4.22	0.94	18.77	12.48	1.99	157.86
Average [1]	**2.64**	**0.78**	**10.80**	**2.86**	**0.83**	**15.50**	**2.34**	**1.93**	**32.71**
Target variables:									
Good health (%)	0.88	0.55	1.33	1.65	0.55	3.26	4.52	1.10	21.51
Interested in politics (%)	0.63	1.01	1.41	1.53	1.05	3.41	−0.24	2.54	2.60
Anti-immigration (%)	1.14	0.85	1.33	0.19	0.87	0.91	0.96	2.16	3.09
Low income (%)	−1.77	0.83	1.41	−4.89	0.80	24.67	−3.56	2.00	14.70
Trusts others (%)	0.87	0.94	2.16	−2.62	0.95	7.79	3.75	2.45	16.51
Satisfied with life (%)	1.16	0.75	3.96	2.84	0.74	8.82	4.49	1.71	21.84
Happy (%)	1.12	0.58	1.70	1.21	0.60	2.06	4.94	1.16	25.52
Stressed (%)	−1.06	0.59	2.09	−3.29	0.55	11.39	−5.54	1.15	31.89
Depressed (%)	0.33	0.41	1.83	0.41	0.43	0.59	0.92	1.08	1.93
Average	**1.00**	**0.72**	**1.91**	**2.07**	**0.73**	**6.99**	**3.21**	**1.71**	**15.51**

Notes. MSE Mean Squared Error and Sampling Variance adjusted according to the net sample size affordable given budget constraint. [1]For Total Bias, average is based on absolute values.

variable (10.80 for the sociodemographics, and 1.91 for the target variables), while Design 2's average MSE values are in-between (15.50 for the sociodemographics, and 6.99 for the target variables).

Although overall the total error was lowest in Design 1 and highest in Design 3, the magnitude of the MSE values was estimate specific, and varied by survey design (the full MSE and component errors for all categories of the variables in Table 4, are available online in Tables A5, A6, and A7 in the supplemental material available at: http://dx.doi.org/10.1515/JOS-2017-0016). Notably, despite having the lowest overall total error, Design 1 had the highest MSEs of all three surveys for three of the sociodemographic variables: the proportion of men, people who are married, and the people living in single-person households. Design 2 had the highest MSEs for estimates of the proportion in the youngest and oldest age groups, and of people born in Switzerland, while Design 3 had the highest MSE only for the estimate of the proportion with a listed fixed line telephone number. For the target variables, Design 2 also had the highest MSE for the estimates of the proportion interested in politics and the proportion finding it difficult to live on their present income. However, on the remainder, the MSEs were highest in Design 3.

3.4. Components of MSE

Next, to address RQ2, we consider the relative contribution of different sources of error – Sampling variance (SV) and Total bias (TB) – to the MSE in each of the survey designs (also shown in Table 4). Sampling variance (SV) is highest in Design 3, due to the higher fixed and variable costs of telephone interviewing, and hence the lower net sample size affordable with a fixed budget of USD 100,000 (see Table 1). SV was very similar for the other two survey designs: despite the higher fixed costs of the mixed mode design (almost twice those of the mail survey), the lower variable costs associated with web mean that similar sample sizes are achievable when the budget constraint is imposed. The sampling variances for Design 1 and Design 2 ranged from 0.41 to 1.01 and 0.43 to 1.06, respectively. This means that the sampling errors of estimated percentages would range from $\pm 0.64\%$ to $\pm 1.00\%$ for Design 1 and from $\pm 0.66\%$ to $\pm 1.03\%$ for survey 2. For Design 3, the sampling variance ranged between 1.08 and 2.54, rendering the margin of sampling error higher as well – between $\pm 1.04\%$ and $\pm 1.60\%$. Thus, precision is considerably lower in Design 3 compared to Designs 1 and 2.

For the target variables, as with the MSE and Cramer's *V* estimates, the total (absolute) bias was largest on average in Design 3 (3.21 percentage points), and lowest in Design 1 (1.0 percentage point compared to 2.07 in Design 2 – see Table 4). By comparison, the differences between the surveys on the sociodemographic variables were only minimal and across the estimates presented in Table 4, the average bias was actually lowest in Design 3 (2.31 percentage points compared to 2.64 in Design 1 and 2.86 in Design 2). Note, however, that across estimates for all categories of the variables we analysed, Design 3 had the largest average biases on both the sociodemographic variables and the target variables (tables available in the supplemental material online at http://dx.doi.org/10.1515/JOS-2017-0016). As for the MSE, the size of the total bias varied by estimate and by survey. The absolute biases for the sociodemographics estimates shown in Table 4 ranged from 0.39 (% aged 15–24 years) to 6.28 (% born in Switzerland) percentage points

in Design 1; from 0.08 (% married) to 8.20 (% born in Switzerland) in Design 2; and from 0.58 (% aged 65 plus) to 12.48 (% with a listed phone number). On the target variables, the absolute biases in Design 1 ranged from 0.33 (% depressed) to 1.77 (% on low income); from 0.19 (% anti-immigration) to 4.89 (% on low income) in Design 2; and in Design 3, from 0.24 (% interested in politics) to 5.54 (% happy). The largest biases were distributed between the three survey designs following exactly the same pattern as the MSEs. Thus, the overall differences observed in the MSEs are not explained by the differences in the SVs, but rather, by the contribution made by total bias.

3.5. Decomposition of Bias

To assess the relative contribution to the total bias made by non-coverage (NCB), nonresponse (NRB) and measurement biases (MEB), we can consider both the relative (absolute) size of the errors, as well as their direction – that is whether the biases have an additive or compensatory effect on the total. Before considering the effect of mixing modes on the contribution to the total of bias from different sources (RQ3), we first compare the composition of biases in estimates at the end of Phase 1 in Designs 2 (web) and 3 (CATI) to the total Design 1 (mail) bias estimates (RQ2).

3.5.1. Sociodemographic Variables

As with the total bias, the relative contribution to the total from the different sources of bias varied by estimate and survey design, and a different pattern of findings was evident for the sociodemographic variables compared to the target variables. Across the three modes used in Phase 1 of each design, NRB (together with NCB in Design 3) made a larger contribution than MEB to the total bias in the sociodemographic estimates, with only three exceptions (the percentage aged 15–24 in Design 1; and in Design 2, the percentage married and the percentage with a listed telephone number). For some variables the different sources of bias combined additively to increase the overall positive or negative bias. This was the case for three out of the seven of the sociodemographic estimates in Design 1 (% male, % married, and % in a single-person household), six of the estimates in Design 2 (all except the % males); while three of the estimates in Design 3 (% aged 15–24, % aged 65, and % with a listed phone number) had positive biases composed of positive, additive contributions from NCB, NRB, and MEB (second half of Table 5). In the remainder, the different sources of bias worked in opposite directions. In Design 1, this pattern occurred for three of the estimates (% married, % born in Switzerland, and % with a listed phone number). In each case, positive NRB was offset by a smaller, negative MEB. In Design 2, only one estimate (% male) had opposing biases – large positive NRB was offset by an almost negligible negative MEB. In Design 3, for two estimates (% married and % born in Switzerland), the positive total bias was composed of a positive SEB offset slightly by a negative MEB; for another estimate (% living in a single person household), the negative total bias was composed of a negative SEB barely offset by a negligible positive MEB; while for another (% male), the negative total bias was composed of a small positive NCB, a larger negative NRB and a smaller negative MEB. These findings provide a clear indication that the choice of data collection mode affects the composition of errors

Table 5. Total Bias and magnitude of bias from different sources, including noncoverage (NCB), nonresponse (NRB), and measurement (MEB) by phase of fieldwork.

	Design 1: Single mode mail			Design 2: Sequential web + mail					
	Phases 1 + 2			Phase 1			Phases 1 + 2		
	Total bias	NRB	MEB	Total bias	NRB	MEB	Total bias	NRB	MEB
Male (%)	−2.91	−1.70	−1.22	1.74	1.91	−0.17	0.28	0.49	−0.21
Aged 15–24 (%)	0.39	−0.09	0.48	3.64	3.06	0.58	2.21	1.02	1.19
Aged 65+ (%)	−1.27	−0.84	−0.43	−10.74	−9.77	−0.97	−2.64	−1.29	−1.35
Married (%)	2.65	3.61	−0.96	−0.93	−0.24	−0.69	0.08	0.98	−0.90
Born in Switzerland (%)	6.28	6.46	−0.18	9.52	7.62	1.90	8.20	6.40	1.80
1 person household (%)	−3.23	−2.70	−0.52	−5.45	−4.41	−1.04	−2.39	−2.46	0.07
Listed phone number (%)	1.74	3.51	−1.77	2.99	1.35	1.64	4.22	2.11	2.11
Good health (%)	0.88	0.88	0.00	5.02	2.36	2.66	1.65	1.83	−0.18
Interested in politics (%)	0.63	0.63	0.00	2.82	−1.61	4.43	1.53	1.82	−0.29
Anti-immigration (%)	1.14	1.14	0.00	−2.20	0.79	−2.99	0.19	−0.07	0.26
Low income (%)	−1.77	−1.77	0.00	−7.56	−1.59	−5.97	−4.89	−1.47	−3.42
Trusts others (%)	0.87	0.87	0.00	−1.72	0.26	−1.98	−2.62	0.00	−2.62
Satisfied with life (%)	1.16	1.16	0.00	4.46	0.76	3.70	2.84	1.36	1.48
Happy (%)	1.12	1.12	0.00	2.48	0.74	1.74	1.21	1.22	−0.01
Stressed (%)	−1.06	−1.06	0.00	−5.64	0.28	−5.92	−3.29	−0.36	−2.93
Depressed (%)	0.33	0.33	0.00	−0.52	0.22	−0.74	0.41	−0.39	0.80

Table 5. Continued.

	Design 3: Combined concurrent and Sequential CATI + mail									
	Phase 1				Phases 1 + 2			Phases 1 + 2 + 3		
	Total bias	NCB	NRB	MEB	Total bias	NRB	MEB	Total bias	NRB	MEB
Male (%)	−1.95	0.18	−1.85	−0.27	−3.84	−2.23	−1.61	−1.95	−0.90	−1.05
Aged 15–24 (%)	2.19	0.96	0.94	0.28	0.23	−0.09	0.32	1.40	0.84	0.56
Aged 65+ (%)	4.71	4.33	0.38	0.00	1.25	1.54	−0.29	−0.58	−0.19	−0.39
Married (%)	5.66	3.74	3.85	−1.92	2.88	3.92	−1.04	1.48	2.22	−0.74
Born in Switzerland (%)	13.47	10.51	6.33	−3.23	6.26	5.63	0.63	6.51	5.73	0.78
1 person household (%)	−4.52	−2.98	−1.58	0.05	−3.76	−2.93	−0.83	−2,99	−2.36	−0.63
Listed phone number (%)	41.16	41.16	0.00	0.00	15.78	3.95	11.84	12.48	3.49	8.99
Good health (%)	4.93	−1.82	−0.05	6.80	3.66	−0.37	4.03	4.52	0.54	3.98
Interested in politics (%)	4.55	4.28	1.50	−1.27	1.16	1.66	−0.50	−0.24	0.68	−0.92
Anti-immigration (%)	−0.83	−1.56	1.16	−0.44	0.92	1.00	−0.08	0.96	0.82	0.14
Low income (%)	−11.56	−7.17	−0.46	−3.81	−4.33	−1.78	−2.55	−3.56	−2.10	−1.46
Trusts others (%)	8.55	2.74	2.39	3.42	4.94	1.58	3.36	3.75	0.56	3.19
Satisfied with life (%)	10.69	3.96	0.43	6.31	5.04	0.86	4.18	4.49	0.99	3.50
Happy (%)	10.50	2.31	0.21	7.98	5.46	0.66	4.80	4.94	0.44	4.50
Stressed (%)	−8.17	2.53	0.49	−11.19	−6.59	0.20	−6.79	−5.54	0.00	−5.54
Depressed (%)	0.50	0.20	−0.08	0.38	0.26	−0.22	0.48	0.92	0.18	0.74

in estimates, and in particular – as would be expected – the relative contribution to the total made by selection errors.

3.5.2. Target Variables

While the total bias in the sociodemographic variables mainly stemmed from SEB, in the target variables, MEB made a more important contribution, illustrating the potential for different modes to also produce different measurements, especially on subjective measures (though it should be borne in mind, as previously mentioned, that some part of the estimated contribution of MEB may in fact be SEB that is not adequately controlled for by the poststratification weighting). In Design 2 (web only), the contribution from the MEB exceeded the contribution from NRB on all nine of the estimates. In Design 3 (CATI only), the pattern was more mixed due to the additional contribution to bias made by the NCB. Here, the MEB contribution was larger than that of the combined SEB on five of the nine variables. In both Designs 2 and 3, the biases had an additive effect on the total on four of the nine variables (in both surveys, these were: % low income (underestimated compared to the mail benchmark survey), % satisfied with life, and % happy (both overestimated compared to the benchmark); plus % in good health in Design 2, and % trusting others in Design 3 (again, both overestimated compared to the benchmark).

In the remaining target variables, the biases worked in opposite directions. For example, in Design 2, NRB resulted in an underrepresentation of people interested in politics (by 1.61 percentage points), however, a positive MEB (of 4.43%) on this variable (respondents by web overreporting their interest in politics relative to the mail survey) overrode the effects of the other bias. In Design 3, the opposite pattern was observed. The NCB and NRB resulted in an overrepresentation of people interested in politics (of 5.78%), and this was offset by a negative MEB (of −1.27% respondents in CATI slightly underreporting their interest in politics relative to the mail survey). For the remaining four target variables in Design 2, the MEB made a much larger opposite contribution than the NRB to the total bias, such that the compensatory effect of the two was only minimal. In Design 3, however, two other target variables had substantial biases made up of different sources working in opposite directions. These were the percentage in good health, where a negative SEB (mainly from NCB) was overridden by a large positive MEB (overreporting of good health in CATI compared to mail); and the percentage reporting feeling stressed, where the positive SEB was counteracted by a large negative MEB (underreporting of stress in CATI compared with the mail survey).

3.6. Effect of Mixing Modes on Bias Components

Finally, we consider the effect of mixing modes on the composition of biases (RQ3). Our primary interest is in whether mixing modes helps to reduce the SEB associated with the starting modes, whether any reduction in SEB is offset by increases in MEB, and the relative contribution made by both sources to changes in the TB. In sum, we find that TB is almost uniformly reduced as a result of mixing modes in the combinations considered in this study. SEB is reduced for most of the sociodemographic estimates in both designs as a result of mixing modes, but for the target variable estimates, the positive effect of adding the mail mode differs by survey design, reducing NRB on more variables in Design 3 than

in Design 2 (where the NRB in some estimates actually increased). The effect of mixing modes on the MEB varies by estimate type. For the sociodemographic variables, the MEB generally increased, while for the target variables it decreased. However, increases in MEB rarely outweighed reductions in the SEB. In the following, we consider in detail the effect of mixing web and mail modes in a sequential design (Design 2), before considering the effects of mixing CATI and mail both concurrently and sequentially (Design 3).

3.6.1. Design 2: Web Plus Mail

Comparing estimates of bias across Phases 1 and 2 of Design 2 (shown in the top-right half of Table 5), we find that TB was reduced as a result of mixing modes on all but two estimates. These include the proportion with a fixed line telephone number (where TB increased from 2.99% to 4.22%); and the proportion reporting that they trust other people (where TB increased from −1.72% to −2.62%). The size of the NRB was reduced on five out of seven of the sociodemographic variables (the two exceptions are the % with a listed phone number, where the positive NRB increased, and the % married, where a negligible under-estimate became a slightly larger over-estimate), but on only four of the nine target variables (% in good health, % anti-immigration, % on low income, and % trusting others). In the remaining target variables, the NRB either increased in the same direction (as was the case for the measures of life satisfaction and happiness); or in the opposite direction (as was the case for the measures of stress and depression, where small positive NRBs became slightly larger negative NRBs; and interest in politics, where an underrepresentation of people interested in politics in the web phase was converted to a greater overrepresentation of this group following the mail phase).

MEB in the sociodemographic estimates produced by Design 2 increased on five of the seven variables. On the remaining two, there was a negligible reduction in MEB on the estimate of the proportion born in Switzerland (from 1.90 to 1.80 percentage points), and a slightly larger reduction on the estimate of the proportion living in a single-person household (from −1.04 to 0.07 percentage points). On the three sociodemographic estimates where NRB went down and MEB went up after Phase 2 (% male, % aged 15–24 and % aged 65+), the size of the increase in MEB did not outweigh that of the decrease in NRB. By contrast, MEB was reduced as a result of mixing modes on seven out of nine of the target variable estimates produced by Design 2. The two exceptions were the proportion (under-) reporting that they trust others (which increased from −1.98 to −2.62), and the proportion reporting feeling depressed (where the total bias was negligible anyway). For all target variables, the change in the MEB was greater in magnitude than the change in the NRB, but the reduction in MEB for most target variables which resulted from switching to the benchmark mode meant that, ultimately, only one estimate (% trusting others) saw an increase in MEB, which offset the reduction in NRB and contributed to an increase in TB (note however, that even in this instance, the size of the NRB was only 0.26 at Phase 1, and 0.00 at Phase 2).

3.6.2. Design 3: Combined Concurrent and Sequential CATI Plus Mail

Comparing estimates of bias across the three phases of Design 3 (shown in the bottom-right half of Table 5), we find that between Phases 1 and 2, TB was reduced on six out of seven sociodemographic estimates (the exception being the proportion of males where TB

increases from −1.95 to −3.84); and on eight out of nine target variable estimates (the exception being the proportion with anti-immigration attitudes, where there was an increase in TB from −0.83 to 0.92). We assume that the addition of Phase 2 (the concurrent mail phase) eliminates the NCB, so we compare the combined magnitude of the SEB (NCB plus NRB) in Phase 1 with the NRB in Phase 2 to draw conclusions about the effects of concurrent mode mixing on SEB. Correspondingly, we find that SEB is reduced on all but one sociodemographic variables (% male, where SEB increases from −1.68 to −2.23 percentage points), and on all but two of the target variables (% with anti-immigration attitudes, where SEB increases from −0.40 to 1.00 per cent). Nevertheless, the elimination of the NCB following Phase 2 is met with a net increase in the estimated NRB for five out of the seven sociodemographic estimates, and six of the nine target variables. The addition of Phase 3 (mail follow-up of nonrespondents) sees further increases in NRB for six estimates (for the sociodemographic estimates, there are small increases in the NRB for the proportion aged 15–24 (from −0.09 to 0.84), and the proportion born in Switzerland (from 5.63 to 5.73), but the remainder benefit from the mail follow-up and reduce in size.

As in Design 2, MEB increased between Phases 1 and 2 on five out seven of the sociodemographic estimates (the two exceptions were the proportion married, where the MEB decreased from −1.92 to −1.04; and the proportion born in Switzerland, where MEB decreased from −3.23 to 0.63). By contrast, MEB decreased for eight out of nine of the target variables, as a result of introducing the benchmark mode (although note that this positive effect is over-estimated here as the same cases are considered in both Design 3 and the benchmark). The one exception was the estimate of the proportion feeling depressed, where total bias was negligible anyway (0.26). Following Phase 3, there was relatively little change in MEB. It only exceeded 0.60 percentage points for one estimate – the proportion with a listed phone number. Here, the TB following Phase 3 remained high (12.48), resulting from an overrepresentation of people with listed numbers in the responding sample (by 3.49), and a strong tendency among respondents to overreport (8.99) that their phone number was listed in the directory. Change in MEB between Phases 2 and 3 was similarly small for the target variables. Here only three variables saw an increase in MEB (% interested in politics, % with anti-immigration attitudes, and % depressed). In all cases, the increase was small (not exceeding 0.42 percentage points), and only exceeded the reduction in NRB observed for the same variables between Phases 2 and 3 for one variable (% depressed, where TB was still only 0.92 following Phase 3).

4. Discussion and Conclusion

A frequently cited motivation for mixing modes of data collection is to try to raise response rates, and thereby reduce selection errors associated with noncoverage and nonresponse. A concern often raised in relation to this is that reductions in selection error may be offset by an increase in measurement error, causing a net increase in the MSE of survey estimates. In this study, we were able to benefit from auxiliary data from population registers that formed the basis of the sampling frame in order to address these concerns in comparisons between a single mode mail survey (Design 1), a sequential mixed mode web plus mail survey (Design 2), and a combined concurrent and sequential CATI plus mail

survey (Design 3). We used these data to decompose the TSE into its component sources and calculate the MSE of estimates produced to draw conclusions about the effect of mixing modes on overall accuracy, and on the relative contribution to accuracy of the individual sources of error. Specifically, we sought to identify which of the three designs offered the lowest overall total error across a range of sociodemographic and target questionnaire variables (RQ1); what was the relative contribution to the MSE of error from different sources (RQ2), and how mixing modes affected the composition of errors across different estimates (RQ3).

As with other studies that have investigated TSE in survey estimates (e.g., Groves and Magilavy 1984; Olson 2006; Peytchev et al. 2009), we found considerable variation across estimates and across survey designs. On average, MSE was lowest in the single mode mail survey (in part due to the decision to effectively "discount" the measurement error by using this survey as the benchmark for the target variables), and highest in the CATI plus mail design (RQ1). Nevertheless, while the largest MSEs for most of the target variables were observed in the CATI plus mail design, for the different sociodemographic estimates the largest MSEs were divided between all three designs. We found differences in the relative contribution of each error source by type of variable and by survey design (as well as some estimate-specific patterns) (RQ2). Bias on sociodemographic variables was generally the result of selection error; in the target variables, measurement error was generally dominant, which is perhaps not surprising as subjective measures are often more susceptible to response biases (although of course, the true value of these variables is unknown).

Overall, total bias on the estimates analysed was reduced as a result of mixing modes, with few exceptions, providing clear evidence that the TSE does not necessarily increase as a result of mixing modes (RQ3), however, mixing modes did not always have the predicted effect on the separate sources of bias. Indeed, the effect of mixing modes on the bias components varied by survey design and type of variable. Mixing web and mail had the effect of reducing NRB in most of the sociodemographic variables as intended, but increased it in over half of the target variables. Meanwhile, mixing CATI and mail concurrently effectively decreased the overall combined selection error from NCB and NRB in most of the sociodemographic *and* target variables, and the addition of the sequential mail follow-up led to further reductions in NRB in over half the variables. However, the elimination in the NCB in estimates was actually accompanied by an increase in NRB on a majority of both types of variable following the concurrent mail phase, and further increases occurred for six of the variables following the sequential mail phase, meaning that three estimates ended up with larger NRBs following all three phases than they had after Phase 1. In contrast, mixing modes (in both the web plus mail and CATI plus mail designs) generally had the effect of increasing MEB in the sociodemographic variables, but decreasing it in the target variables (though in both designs there were, again, some exceptions to this pattern).

The higher MSEs in the CATI plus mail survey were in part attributable to the higher sampling variances due to smaller sample sizes, which in turn, were the result of the higher combined fixed costs of mixing CATI and mail surveys, and the higher variable costs per sample member. However, the interpretation of the relative contribution of bias and variance to MSE and of its magnitude is obscured somewhat by the fact that larger errors

are weighted more heavily than smaller ones as a result of the squaring of bias terms. A further difficulty is that it is not clear what the threshold for MSE should be for researchers to conclude that the TSE is severe. For these reasons, Cramer's *V* may be preferred over MSE as an overall estimate of the total error (in categorical variables), because of the possibility of interpreting the size of the effects. Based on the *V* statistics, we conclude that the error in this study was generally small (and consistent with the results of Klausch et al. (2015a), slightly smaller on the target variables than on the sociodemographic variables), but the overall conclusions drawn from these two indicators regarding the relative quality of the three surveys were ultimately not different.

Our findings largely mirror those of other studies. Response rates were remarkably similar in all three designs, but in the mixed mode surveys, this was only possible as a result of switching modes. Increases in response rates in the mixed mode surveys did correspond to overall reductions in bias, but as mixing modes affected the composition of errors from different sources this could have implications for the comparability of the data across population subgroups. As others have found (e.g., Millar and Dillman 2011), the single mode mail survey fared well compared to the mixed mode surveys. However, this conclusion is not independent of the decision to use it as the benchmark. In fact, in the sociodemographic variables it was evident that deviations from the register-based estimates could not only be attributed to selection errors, but also to measurement bias (which as acknowledged previously could have included error from other sources). For this reason, we should hesitate to conclude that the mail mode per se offers better accuracy than the other modes. Indeed, sampling variances in the mail survey were very similar to those of the web plus mail design, and with a larger budget, further gains in precision (e.g., for subgroup analyses) would likely be possible in such a design due to its lower variable costs (Vannieuwenhuyze 2014). This could potentially offset the disadvantage of greater measurement bias in the mail mode (especially if combined with efforts to minimise processing errors and ideally, to correct for measurement differences between the modes). With these considerations in mind, our findings contribute to the mounting evidence that survey designs that combine web and mail offer a number of cost and error advantages over designs combining interviewer- and self-administered methods (Dillman et al. 2014).

Our analysis of target variables from the questionnaire employed a calibration method that relied on auxiliary data from the sampling frame to 'correct' the selection errors observed on these variables. This method may be suboptimal as a way of disentangling mode-related selection and measurement effects (Vannieuwenhuyze and Loosveldt 2012; Schouten et al. 2013; Klausch et al. 2015b; Hox et al. 2017), and so our estimates of bias from different sources are dependent on the nonresponse weighting adjustment and, therefore, are themselves not error free. It is highly likely that despite the random assignment of sample members to survey designs, selection into a particular mode was non-random with respect to variables for which no exogenous auxiliary data are available. Furthermore, there is evidence that using the kinds of sociodemographic variables used here for poststratification weighting may not succeed in correcting for selection errors if they are uncorrelated with the target variables (Peytcheva and Groves 2009). This may limit the accuracy of our bias estimates for the target variables, but it is not uncommon for methodologists to construct weights based on sociodemographic variables, so our methods

at least reflect common survey practice. Furthermore, it is relatively rare to have access to auxiliary data of the kind we were able to make use of, and the possibility to make use of them in this way makes an important contribution to the relatively sparse literature comparing TSE in mixed mode survey designs to single mode designs.

Given that our choice of benchmark mode affects our conclusions with respect to the accuracy of the target variables, it would be of interest to consider an alternative mode as a benchmark. Given the interest among large-scale survey programmes in finding out how lower cost mixed mode surveys compare with single mode face-to-face surveys, a useful extension of our analysis would be to use the 2012 Swiss European Social Survey (ESS) as the comparison survey, as the fieldwork was carried out at the same time as the mode experiment reported here, and the questionnaire carried many of the same questions. However, comparisons would likely be compromised by the fact that the questionnaire for the mode experiment was considerably shorter than that of the ESS, and the order of questions was not identical. Furthermore, the response rate for the ESS was lower than that for the mail survey conducted as part of this study, which could mean the responding sample is less representative of the population. A mail survey comparison offered certain other advantages for the present study. For one, some of the questions were relatively sensitive, for which self-administered modes generally provide superior measures (Kreuter et al. 2008). For another, interviewer-administered surveys can suffer from interviewer-related effects other than social desirability bias (and in face-to-face surveys these are confounded with clustering in the sample design). Another promising alternative could be to use a hybrid mixed mode benchmark (Klausch et al. 2015b), for example, combining the measurement quality of the web survey with the selection error of the mail survey, but it is not clear this would offer any advantages.

Smith (2011, 465) has argued that the lack of available measures of true population values for most survey variables represents a major limitation of the TSE perspective. He argues for a refinement that emphasises 'total survey measurement variation' and takes into account the inherent challenges and potential for error involved in making comparisons across studies. Likewise, Biemer (2010) has argued that the emphasis on accuracy in the TSE paradigm may undermine other more pertinent criteria on which to select between competing survey designs. Following the TSE approach, useful extensions to the present study would be to try to deconstruct the complex interactions between different error sources, as these have generally received little attention, particularly in comparisons across studies (Smith 2011, 474), and to explore in more detail the conditions under which errors from different sources offset one another or serve to cancel each other out. At the same time, however, researchers should be conscious of the possible limits of the TSE paradigm in the current survey climate.

5. References

Biemer, P.P. 1988. "Measuring Data Quality." In *Telephone Survey Methodology*, edited by R.M. Groves, P.P. Biemer, L. Lyberg, J.T. Massey, W. II, Nicholls, and J. Waksberg, 341–375. New York: Wiley.

Biemer, P.P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74: 817–848. Doi: http://dx.doi.org/10.1093/poq/nfq058.

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, NJ: Wiley.

Blumberg, S.J. and J.V. Luke. 2013. "*Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July–December 2012*." National Center for Health Statistics. December. Available at: http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201306.pdf (accessed March 2017).

Brick, J.M. and D. Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." *The ANNALS of the American Academy of Political and Social Science* 645: 36–59. Doi: http://dx.doi.org/10.1177/0002716212456834.

Carley-Baxter, L.R., A. Peytchev, and M.C. Black. 2010. "Comparison of Cell Phone and Landline Surveys: A Design Perspective." *Field Methods* 22(1): 3–15. Doi: http://dx.doi.org/10.1177/1525822X09360310.

Chang, L. and J.A. Krosnick. 2009. "National surveys via RDD telephone interviewing versus the internet. Comparing sample representativeness and response quality." *Public Opinion Quarterly* 73: 641–678. Doi: http://dx.doi.org/10.1093/poq/nfp075.

Couper, M.P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly* 75(5): 889–908. Doi: http://dx.doi.org/10.1093/poq/nfr046.

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

De Leeuw, E.D. and N. Berzelak. 2017. "Survey mode or survey modes?" In *The Sage Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T.W. Smith, and Y.-C. Fu, 142–156. London: Sage Publications.

De Leeuw, E. and W. de Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In *Survey nonresponse*, edited by R. Groves, D. Dillman, J. Eltinge, and R.J.A. Little, 41–54. New York: Wiley.

De Leeuw, E.D., J.J. Hox, and D.A. Dillman. 2008. "Mixed mode surveys: When and why." In *International Handbook of Survey Methodology*, edited by E.D. de Leeuw, J.J. Hox, and D.A. Dillman, 299–316. New York/London: Erlbaum/Taylor & Francis.

Dillman, D.A., G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B.L. Messer. 2009. "Response Rate and Measurement Differences in Mixed-Mode Surveys using Mail, Telephone, Interactive Voice Response (IVR) and the Internet." *Social Science Research* 38: 1–18. Doi: http://dx.doi.org/10.1016/j.ssresearch.2008.03.007.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-mode Surveys: the Tailored Design Method* (4th Edition). Hoboken: Wiley.

Eva, G., G. Loosveldt, P. Lynn, P. Martin, M. Revilla, W. Saris, and J. Vannieuwenhuyze. 2010. *Assessing the Cost-Effectiveness of Different Modes for ESS Data Collection*. London: City University.

Fowler, F.J., P.M. Gallagher, V.L. Stringfellow, A.M. Zaslavsky, J.W. Thompson, and P.D. Cleary. 2002. "Using Telephone Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members." *Medical Care* 40: 190–200.

Gordoni, G., P. Schmidt, and Y. Gordoni. 2012. "Measurement invariance across face-to-face and telephone modes: the case of minority-status collectivistic oriented groups." *International Journal of Public Opinion Research* 24(2): 185–207. Doi: http://dx.doi.org/10.1093/ijpor/edq054.

Greene, J., H. Speizer, and W. Wiitala. 2008. "Telephone and Web: Mixed-Mode Challenge." *Health Services Research* 43: 230–248. Doi: http://dx.doi.org/10.1111/j.1475-6773.2007.00747.x.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675.

Groves, R.M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75(5): 861–971. Doi: http://dx.doi.org/10.1093/poq/nfl033.

Groves, R.M., F.J. Fowler Jr., M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology (Wiley Series in Survey Methods)*, 2nd Ed. Hoboken, NJ: John Wiley & Sons.

Groves, R.M. and L.J. Magilavy. 1984. "An Experimental Measurement of Total Survey Error." In *Proceedings of the Section on Survey Research Methods: American Statistical Association*, 698–703. Alexandria, VA: American Statistical Association. Available at: http://ww2.amstat.org/sections/srms/Proceedings/ (accessed March 2017).

Heerwegh, D. and G. Loosveldt. 2011. "Assessing mode effects in a national crime victimization survey using structural equation models: social desirability bias and acquiescence." *Journal of Official Statistics* 27: 49–63.

Hochstim, J.R. 1967. "A Critical Comparison of Three Strategies of Collecting Data from Households." *Journal of the American Statistical Association* 62: 976–989. Doi: http://dx.doi.org/10.2307/2283686.

Holbrook, A., M. Green, and J. Krosnick. 2003. "Telephone Versus Face-to-face Interviewing of National Probability Samples with Long Questionnaires." *Public Opinion Quarterly* 67: 79–125. Doi: http://dx.doi.org/10.1086/346010.

Hox, J., E.D. de Leeuw, and T. Klausch. 2017. "Mixed mode research: Issues in design and analysis." In *Total Survey Error in Practice: Improving Quality in the Era of Big Data*, edited by P.P. Biemer, E.D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, C. Tucker, and B.T. West, 511–530. Hoboken, NJ: John Wiley and Sons, Inc.

Hox, J.J., E.D. de Leeuw, and E.A.O. Zijlmans. 2015. "Measurement equivalence in mixed mode surveys." *Frontiers in Psychology: Quantitative Psychology and Measurement*. Doi: http://dx.doi.org/10.3389/fpsyg.2015.00087.

Kappelhof, J.W.S. 2013. "The Effect of Different Survey Designs on Nonresponse in Surveys of Non-Western Minorities in The Netherlands." *Survey Research Methods* 8(2): 81–98. Doi: http://dx.doi.org/10.18148/srm/2014.v8i2.5784.

Klausch, L.T., J.J. Hox, and B. Schouten. 2013. "Measurement effects of survey mode on the equivalence of attitudinal rating scale questions." *Sociological Methods and Research* 42: 227–263. Doi: http://dx.doi.org/10.1177/0049124113500480.

Klausch, T., J.J. Hox, and B. Schouten. 2015a. "Selection Error in Single- and Mixed Mode Surveys of the Dutch General Population." *Journal of the Royal Statistical Society Series A* 178: 945–961. Doi: http://dx.doi.org/10.1111/rssa.12102.

Klausch, T., B. Schouten, and J.J. Hox. 2014. "The Use of Within-subject Experiments for Estimating Measurement Effects in Mixed-mode Surveys." *Statistics Netherlands Discussion Paper,* 2015/06. Available at: https://www.cbs.nl/en-gb/background/2014/11/the-use-of-within-subject-experiments-for-estimating-measurement-effects-in-mixed-mode-surveys (accessed March 2017).

Klausch, T., B. Schouten, and J.J. Hox. 2015b. "Evaluating Bias of Sequential Mixed-mode Designs Against Benchmark Surveys." *Sociological Methods and Research* 1–34. Doi: http://dx.doi.org/10.1177/0049124115585362.

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

Kreuter, F., G. Müller, and M. Trappmann. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74: 880–906. Doi: http://dx.doi.org/10.1093/poq/nfq0.

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

Lipps, O., N. Pekari, and C. Roberts. 2015. "Undercoverage and Nonresponse in a List-Sampled Telephone Election Study." *Survey Research Methods* 9(2): 71–82. Doi: http://dx.doi.org/10.18148/srm/2015.v9i2.6139.

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

Massey, D.S. and R. Tourangeau. 2013. "Where Do We Go from Here? Nonresponse and Social Measurement." *The Annals of the American Academy of Political and Social Science* 645(1): 222–236. Doi: http://dx.doi.org/10.1177/0002716212464191.

Millar, M.M. and D.A. Dillman. 2011. "Improving Response to Web and Mixed-Mode Surveys." *Public Opinion Quarterly* 75(2): 249–269. Doi: http://dx.doi.org/10.1093/poq/nfr003.

Olson, K. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias." *Public Opinion Quarterly* 70(5): 737–758. Doi: http://dx.doi.org/10.1093/poq/nfl038.

Olson, K., J.D. Smyth, and H. Wood. 2012. "Does Providing Sample Members with Their Preferred Survey Mode Really Increase Participation Rates?" *Public Opinion Quarterly* 76(4): 611–635. Doi: http://dx.doi.org/10.1093/poq/nfs024.

Peytchev, A., R.K. Baxter, and L.R. Carley-Baxter. 2009. "Not All Survey Effort is Equal. Reduction of Nonresponse Bias and Nonresponse Error." *Public Opinion Quarterly* 73(4): 785–806. Doi: http://dx.doi.org/10.1093/poq/nfp037.

Peytcheva, E. and R.M. Groves. 2009. "Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates." *Journal of Official Statistics* 25(2): 193–201.

Rao, J.N.K. and A.J. Scott. 1987. "On simple adjustments to chi-square tests with sample survey data." *The Annals of Statistics* 15(1): 385–397. Doi: http://dx.doi.org/10.1214/aos/1176348654.

Roberts, C., D. Joye, M. Ernst Stähli, and R. Sanchez Tome. 2016. Mixing modes of data collection in Swiss social surveys: Methodological Report of the LIVES-FORS Mixed Mode Experiment. *LIVES Working Paper Series,* 2016/48. Doi: http://dx.doi.org/10.12682/lives.2296-1658.2016.48.

Sakshaug, J.W., T. Yan, and R. Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-mode Survey of Sensitive and

Non-sensitive Items." *Public Opinion Quarterly* 74(5): 907–933. Doi: http://dx.doi.org/10.1093/poq/nfq057.

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(6): 1555–1570. Doi: http://dx.doi.org/10.1016/j.ssresearch.2013.07.005.

Siemiatycki, J. 1979. "A Comparison of Mail, Telephone, and Home Interview Strategies for Household Health Surveys." *American Journal of Public Health* 69: 238–245.

Smith, T.W. 2011. "Refining the Total Survey Error Perspective." *International Journal of Public Opinion Research* 23(4): 464–484. Doi: http://dx.doi.org/10.1093/ijpor/edq052.

Suzer-Gurtekin, Z., S. Heeringa, and R. Vaillant. 2012. "Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-mode Surveys." *Proceedings of the JSM, Section on Survey Research Methods* 4711-2. Available at: https://www.niss.org/sites/default/files/VII%201%20Suzer-Gurtekin_itsew2013.pdf (accessed April 2016).

Tourangeau, R. 2017. "Mixing modes: Tradeoffs among coverage, nonresponse, and measurement error." In *Total Survey Error in Practice: Improving Quality in the Era of Big Data*, edited by P.P. Biemer, E.D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, C. Tucker, and B.T. West. 115–132. Hoboken, NJ: John Wiley and Sons, Inc.

Tourangeau, R., R.M. Groves, and C.D. Redline. 2010. "Sensitive topics and reluctant respondents: Demonstrating a link between nonresponse bias and measurement error." *Public Opinion Quarterly* 74(3): 413–432. Doi: http://dx.doi.org/10.1093/poq/nfq004.

Vannieuwenhuyze, J.T.A., G. Loosveldt, and G. Molenberghs. 2010. "A Method for Evaluating Mode Effects in Mixed-mode Surveys." *Public Opinion Quarterly* 74: 1027–1045. Doi: http://dx.doi.org/10.1093/poq/nfq059.

Vannieuwenhuyze, J.T.A. 2014. "On the Relative Advantage of Mixed-Mode versus Single-Mode Surveys." *Survey Research Methods* 8(1): 31–42. Doi: http://dx.doi.org/10.18148/srm/2014.v8i1.5500#sthash.xSmtK1fH.dpuf.

Vannieuwenhuyze, J.T.A. and G. Loosveldt. 2012. "Evaluating Relative Mode Effects in Mixed-mode Surveys: Three Methods to Disentangle Selection and Measurement Effects." *Sociological Methods and Research* 42: 82–104. Doi: http://dx.doi.org/10.1177/0049124112464868.

Wagner, J., J. Arrieta, H. Guyer, and M.B. Ofstedal. 2014. "Does Sequence Matter in Multimode Surveys: Results from an Experiment." *Field Methods* 26(2): 141–155. Doi: http://dx.doi.org/10.1177/1525822X13491863.

Received February 2016
Revised March 2017
Accepted April 2017