Journal of Official Statistics
vol. 33, i. 2 (2017)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, p. 301, http://dx.doi.org/10.1515/JOS-2017-0015

Editorial – Special Issue on Total Survey Error (TSE)

Stephanie Eckman[1] *and Edith de Leeuw*[2]

In the past two decades, the focus in survey methodology and statistics has shifted from studying one source of error (e.g., nonresponse) to studying two or more sources of error simultaneously (e.g., nonresponse and coverage error, nonresponse and measurement error). The Total Survey Error (TSE) framework delineates and describes all the ways in which error can arise in survey data: in design, collection, processing, and analysis. Paul Biemer initiated the international total survey error workshop (ITSEW), and since then workshops have taken place annually in varying locations around the world. Information on these past workshops can be obtained at http://www.niss.org/search/node/itsew Presentations and papers from these workshops have been very influential, and as a result the TSE framework has had a growing impact on survey research (e.g., Biemer 2010).

In 2010, a special issue of *Public Opinion Quarterly* devoted to TSE contained a contribution by Bob Groves and Lars Lyberg titled "Total Survey Error: Past, Present and Future." This article provided a historical overview of survey errors. New technologies have stimulated changes in data collection, and new statistical and modeling tools have changed our analysis procedures. Reacting to these changes, Paul Biemer again took the initiative and in 2013 gathered a small group of experts to organize the first international conference on TSE. The TSE15 conference "Improving Data Quality in the ERA of Big Data" took place in Baltimore, MD USA in September 2015. An edited volume of invited papers from the conference has just been published in the Wiley series in survey methodology (Biemer et al. 2017).

The TSE15 conference also involved 129 contributed presentations. The articles published in this Special Issue are based on those talks; presenters were invited to develop full research articles and submit them to the Journal of Official Statistics.

We would like to thank the authors for their hard work on the articles and the reviewers for their insightful comments. We also thank the staff of JOS, especially Ingegerd Jansson and Susanna Emanuelsson, for their knowledge and help in producing this special issue.

References

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

Biemer, P.P., E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West. 2017. *Total Survey Error in Practice*. Wiley series in survey methodology. New York: Wiley.

Groves, R.M. and L.E. Lyberg 2010. "Total Survey Error: Past, Present and Future." *Public Opinion Quarterly* 74: 849–879. Doi: https://doi.org/10.1093/poq/nfq065.

[1] Fellow, Survey Research Division, RTI International Washington DC, U.S.A. Email: seckman@rti.org
[2] MOA-Professor Survey Methodology, Department of Methodology & Statistics, Utrecht University, Utrecht, The Netherlands. Email: e.d.deleeuw@uu.nl

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.

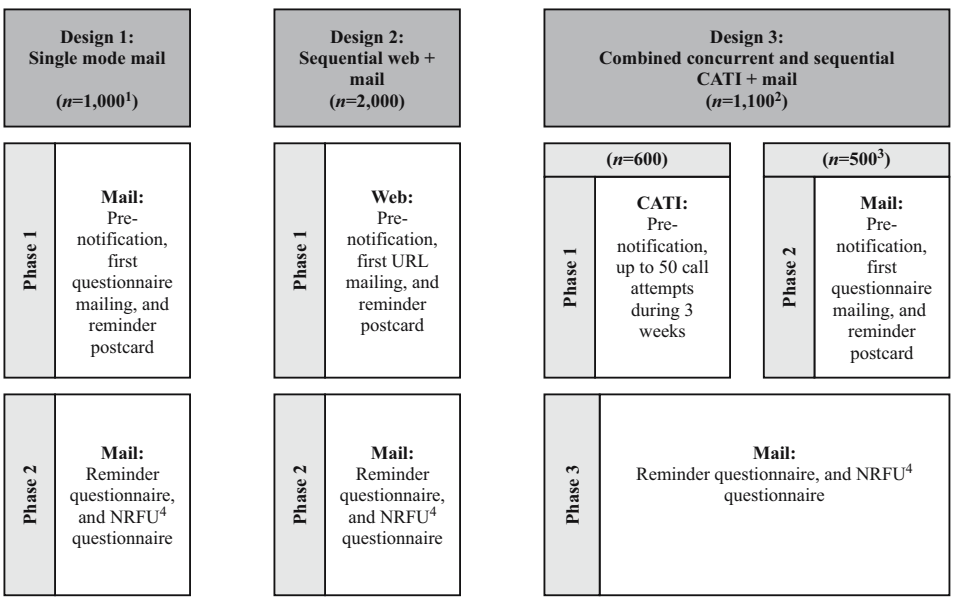

Fig. 1. Survey designs considered.

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.

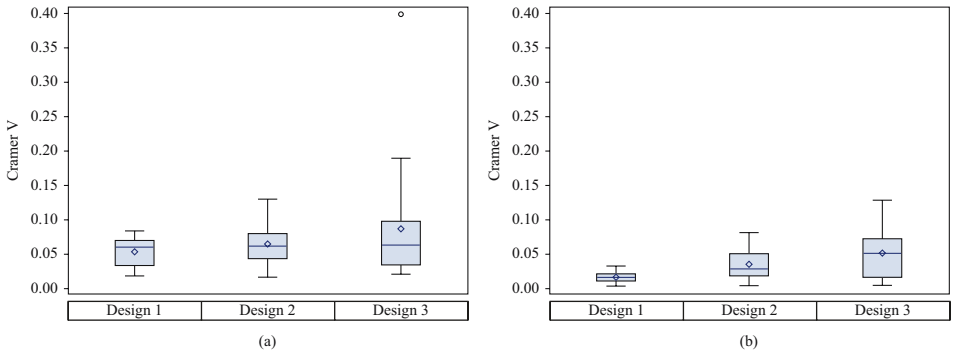

Fig. 2. Distribution of Cramer's V statistics (see Table 3) measuring absolute deviations from the benchmarks for the three survey designs: (a) sociodemographics; (b) target variables.

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

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 335–366, http://dx.doi.org/10.1515/JOS-2017-0017

Total Survey Error and Respondent Driven Sampling: Focus on Nonresponse and Measurement Errors in the Recruitment Process and the Network Size Reports and Implications for Inferences

Sunghee Lee[1], *Tuba Suzer-Gurtekin*[1], *James Wagner*[1], *and Richard Valliant*[1]

This study attempted to integrate key assumptions in Respondent-Driven Sampling (RDS) into the Total Survey Error (TSE) perspectives and examine TSE as a new framework for a systematic assessment of RDS errors. Using two publicly available data sets on HIV-at-risk persons, nonresponse error in the RDS recruitment process and measurement error in network size reports were examined. On nonresponse, the ascertained partial nonresponse rate was high, and a substantial proportion of recruitment chains died early. Moreover, nonresponse occurred systematically: recruiters with lower income and higher health risks generated more recruits; and peers of closer relationships were more likely to accept recruitment coupons. This suggests a lack of randomness in the recruitment process, also shown through sizable intra-chain correlation. Self-reported network sizes suggested measurement error, given their wide dispersion and unreasonable reports. This measurement error has further implications for the current RDS estimators, which use network sizes as an adjustment factor on the assumption of a positive relationship between network sizes and selection probabilities in recruitment. The adjustment resulted in nontrivial unequal weighting effects and changed estimates in directions that were difficult to explain and, at times, illogical. Moreover, recruiters' network size played no role in actual recruitment. TSE may serve as a tool for evaluating errors in RDS, which further informs study design decisions and inference approaches.

Key words: Sampling hard-to-reach populations; chain referral; network-based sampling; measurement error; nonresponse error.

1. Introduction

This article attempts to provide a framework for evaluating Respondent-Driven Sampling (RDS) by integrating its key assumptions into the Total Survey Error (TSE), (Groves 1989) as suggested by Lee (2009). RDS, introduced by Heckathorn (1997, 2002), has gained tremendous popularity due to rising demands for data on rare, hidden and/or elusive populations, for example, sexual minorities (for example, Ramirez-Valles et al. 2005), injection drug users (for example, Burt et al. 2010), racial and ethnic minorities (for example, Dombrowski et al. 2013) and recent immigrants (for example, Montealegre et al. 2013). Not only in the scientific communities is RDS popular, but also in government statistical systems. RDS is practiced by the Centers of Disease Control and Prevention in

[1] Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48104, U.S.A. Emails: sungheel@umich.edu, tsuzer@umich.edu, jameswag@umich.edu, and rvallian@umd.edu.
Acknowledgment: This research was supported by the National Science Foundation [grant number SES-1461470].

the United States (Lansky et al. 2007; Centers for Disease Control and Prevention (CDC) 2009, 2013; Lin et al. 2013).

While there is an attempt to improve analytic aspects of RDS (for example, Salganik and Heckathorn 2004; Volz and Heckathorn 2008; Gile 2011), design aspects aligned with the realities of data collection remain largely unexamined. The TSE framework allows a systematic examination of errors, which, in turn, further informs assessing design and analytic aspects and refining them to reduce overall error. This study examines TSE as a new framework for a systematic assessment of RDS errors by using two publicly available data sets on HIV-at-risk persons. Section 2 provides an overview of RDS by comparing its theoretical development and current practice and then turns to a set of assumptions in RDS and their relevance to TSE. Data sources and methods used in this study are introduced in Section 3. Sections 4 and 5 report results from the analysis. We offer a summary of this study and open questions in Section 6.

2. Respondent Driven Sampling

2.1. *Overview of RDS*

While rare in the general population, some population subgroups are interlinked. For instance, use of injection drugs often involves others who also inject drugs, and this connectedness directly forms informal social networks among Injection Drug Users (IDUs). Although rare and hidden from the outsiders, IDUs may be easily located within these networks. RDS attempts to locate these social networks and exploit them to generate samples.

RDS roughly follows these steps in practice: First, researchers recruit a few members of the target group typically through some type of convenience sampling and collect data from them. While data collection ends at this point in traditional sample surveys, these respondents in RDS are asked to recruit their peers in their social networks. Recruited peers become respondents as well as recruiters for further recruitment. Data collection and recruitment proceed in "waves", as seen in Figure 1, until the cumulative sample size reaches the target sample size or some other criteria set in respective studies (for example, available resources, timeline). Respondents in RDS are not only the source of data, but also recruiters for participants in the immediately subsequent wave (hence, respondent *driven* sampling). For this reason, those recruited initially by researchers are called *seeds*. As noted in Figure 1, recruitment chains are formed from each seed. Under a set of assumptions examined in Subsection 2.2., these chains are regarded as Markov chains, necessitating the chain length to be reasonably long. This process then leads RDS to stationary probabilities (or equilibrium), where the characteristics of the cumulative sample become independent of seeds' characteristics. This is also a point at which the sample is assumed to become unbiased (Heckathorn 1997, 2002).

A distinctive feature of RDS recruitment is the usage of recruitment coupons. In practice, a predetermined number of coupons are given to recruiters, who then distribute the coupons to their peers. These recruits need to redeem the coupons in order to participate in the study. Once they participate, they are given coupons to distribute to their peers. (Naturally, seeds participate in the study without coupons, and, hence, seeds only distribute coupons.) With serial numbers on the coupons, the link between recruiters and

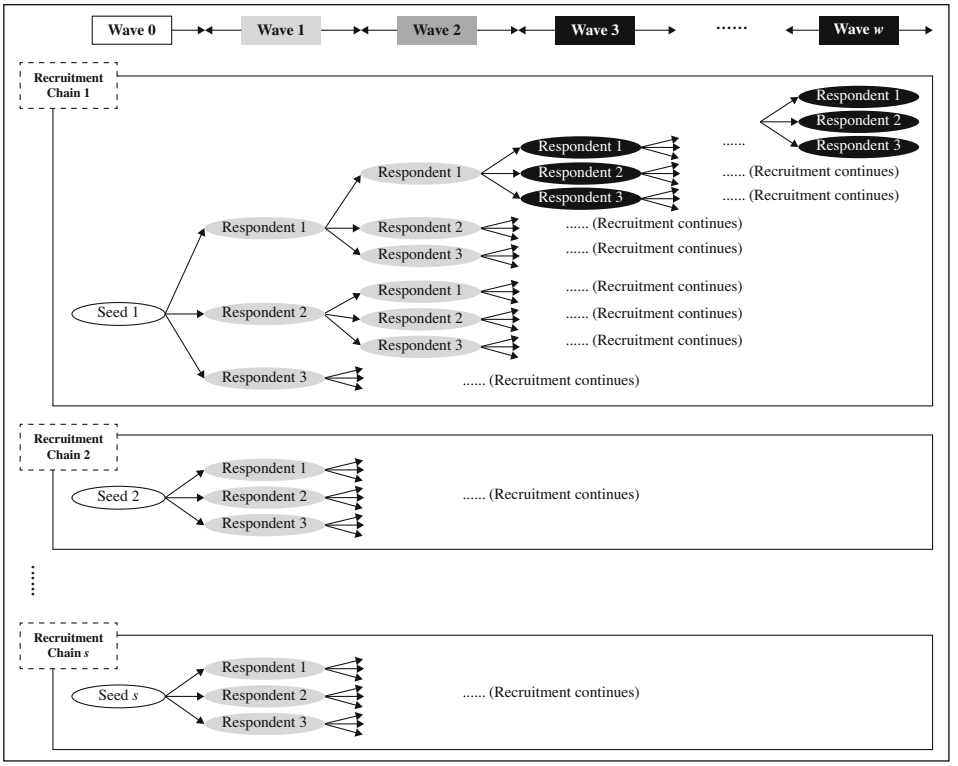

Fig. 1. Respondent-driven sampling recruitment process in theory.

their recruits can be traced. As recruitment is done through coupons, RDS does not require collecting any personal information about respondents' peers. Coupons also play an important role in incentivizing participation, which requires redeeming coupons, *and* recruitment efforts, which is reflected in the number of redeemed coupons, equating to the number of recruits. Coupons not only decrease the data collection costs but also eliminate concerns with privacy, which led some to consider RDS innovative (for example, Baker et al. 2013) and to advocate RDS consistent with the "voluntariness" spirit of the research participants, while criticizing probability sampling as being intrusive (for example, Constantine 2010). On the other hand, others have raised concerns about bias in RDS due to the incentivized nature of recruitment and the potential for unwarranted influence or coercion in the recruitment process (for example, Phillips 2010; Simon and Mosavel 2010).

One piece of critical information in RDS is the number of peers in respondents' networks, termed as *degree*, as it is a key element of RDS estimators (for example, Salganik and Heckathorn 2004; Volz and Heckathorn 2008; Gile 2011). For instance, the RDS-II estimator (Volz and Heckathorn 2008), takes the form of the Horvitz-Thompson or Hájek estimator as follows:

$$\hat{y} = \sum_{i \in S} \left(y_i d_i^{-1} \right) \Big/ \sum_{i \in S} \left(d_i^{-1} \right), \tag{1}$$

where y_i is a variable of interest measured on person i in the sample S and d_i is the network size of person i. Essentially, d_i^{-1} is used as an adjustment factor on the assumption that

persons with larger networks have higher chances of being sampled. This leads d_i^{-1} to be called a "weight" in the RDS literature. For instance, a weight of one will be given if a respondent has one peer and a weight of 0.01 if a respondent has 100 peers. By using this weight, estimation considers the characteristics of respondents with larger networks at a lower level than those with smaller networks. It should be noted that d_i^{-1} is different than weights in probability sampling, which are used to estimate population totals as incorporating population-level information. Weights in RDS simply rearrange the sample distribution by network sizes without incorporating population-level information and, hence, are irrelevant for estimating population totals. In fact, estimating population totals using RDS data requires intensive computing work (Handcock 2012).

2.2. *Assumptions in Respondent Driven Sampling*

Theoretical developments of RDS are based on a set of assumptions (Heimer 2005; Gile and Handcock 2010). Although essential for the claimed unbiasedness, these assumptions are strong, unrealistic and often difficult to verify, and violations are ignored in the inference. We discuss six assumptions. Note that the last two (Assumptions E and F) are not explicitly discussed in the RDS literature, but are critical for using existing RDS estimators.

A. Network Structure: RDS assumes that there is only one network that covers the entire population of interest. That is, everyone in the population can be traced from any starting point. This assumption requires a dense network with a single component. If the population includes multiple non-linked or loosely-linked networks, this assumption is violated. It was shown that estimates were sensitive to such a violation (Lu et al. 2012).

B. Equilibrium Condition: Let vector Q be the successive indices of units sampled by the random walk process and Q_k be the index of units sampled at the k^{th} wave. This follows the Markov process with a transition matrix,

$$P(Q_{k+1} = j | Q_k = i) = \begin{cases} 1/d_i, & \text{if } d_{ij} = 1 \\ 0, & \text{if } d_{ij} = 0 \end{cases}, \tag{2}$$

where d_{ij} is a link function in the sociomatrix of relations between unit i and unit j, defined as

$$d_{ij} = \begin{cases} 1, & \text{if there is a link between unit } i \text{ and unit } j \\ 0, & \text{if there is no link between unit } i \text{ and unit } j \end{cases}, \tag{3}$$

and $d_i = \sum_{j \neq i} d_{ij}$ in (1). Equilibrium assumption is that, as recruitment waves continue, the characteristics of recruits become independent of the seeds' characteristics. In other words, selection of seeds is not critical in the overall inference (Heckathorn 1997). This is directly related to the memorylessness of Markov chains, where the future and past states are independent given the present state. This allows us to rewrite (2) as $P(Q_{k+1} = j | Q_k = i) = P(Q_{k+1} = j | Q_k = i, Q_{k-1} = i - 1, \ldots, Q_1 = 1)$. Further, the equilibrium state in RDS is assumed to be approached at a geometric rate.

C. Random Recruitment: RDS lets respondents control the recruitment process (Frost et al. 2006). The assumption here is that recruitment is done at random, implying that

recruiters do not use systematic criteria for selecting their recruits. The transition matrix in (2) is achieved only when any given unit j ($j = 1, \ldots, d_i$) within the network of unit i selected at the kth wave has the equal probability to be selected into the $(k+1)^{\text{th}}$ wave. Lu et al. (2012) showed that systematic recruitment results in a large bias and variance.

D. Equal Homophily: Homophily is the tendency of individuals of similar characteristics to associate with one another. RDS assumes that homophily rate in the recruitment is equal across subgroups: the tendency of a member of group G to recruit other members of group G is the same as a group H member recruiting group G members.

E. Complete Response: Unlike traditional surveys, where nonresponse occurs at the time of an interview attempt, nonresponse in RDS occurs in four stages:

(1) whether respondents take coupons,
(2) whether those who take coupons actually distribute them to their peers,
(3) whether their peers accept coupons, and
(4) whether the peers who accept coupons actually participate in the study. Overall, these stages can be expressed with a vector of 0/1 response indicators, $\boldsymbol{r} = (r_1, r_2, r_3, r_4)$.

Note that nonresponse on the last two stages implies not only their own nonresponse, but also nonresponse by their peers who may, otherwise, be open to accepting coupons and participating in data collection. Current RDS practice assumes a 100 percent response rate for all stages (i.e., $r_1 = r_2 = r_3 = r_4 = 1$). This compound nature of nonresponse has been recognized only very recently (for example, Lee et al. 2012; Gile et al. 2015).

This assumption affects error properties through nonresponse bias *and* overall sampling productivity in RDS. On the error properties, with the multiple stages of nonresponse introduced above, RDS is subject to a larger scope for nonresponse bias than traditional probability sampling. Under complete response, RDS sample sizes over waves should grow exponentially. However, in the presence of nonresponse at any of the four stages, this exponential growth becomes unlikely. This further leads to slow sample size growth or smaller sample sizes than expected, and short recruitment chains, breaking the Markov process, which is required for Assumptions A, B, and C. Hence, slow sample size growth, small sample sizes, or short recruitment chains may serve as evidence for equilibrium being not realized. It should be noted that, as nonresponse affects RDS sampling productivity, the number of seeds and the chain length are unascertainable during design stages.

F. Accurate and Complete Network Size Measures: In practice, RDS estimators use $\tilde{d}_i$, a self-reported network size, instead of d_i. For example, the RDS-II estimator in (1) becomes $\tilde{y} = \sum_{i \in S}\left(y_i \tilde{d}_i^{-1}\right) / \sum_{i \in S}\left(\tilde{d}_i^{-1}\right)$. This implicitly assumes that $\tilde{d}_i$ either is error free (i.e., $\tilde{d}_i = d_i$) or has a fixed error rate across i (for example, $\tilde{d}_i = pd_i$, where p is a constant, $0 < p \leq 1$). Notably, $\tilde{d}_i$ is self-reported, subject to measurement error. The social network literature clearly indicates that obtaining an accurate network size is challenging because the scope and the nature of the networks are not standardized and that, even when the networks are defined narrowly, it is still found to be difficult (Laumann et al. 1983; Marsden 1990). This makes the error-free assumption or the fixed error rate assumption an unlikely scenario. Additionally, $\tilde{d}_i$ is subject to item nonresponse as well as zero network size reports, posing additional difficulties in using $\tilde{d}_i^{-1}$ as a weight variable.

2.3. Total Survey Error and Respondent Driven Sampling

Because RDS is a sampling method, one may conclude that RDS is subject only to sampling error. However, the chain referral in RDS affects all error types under the TSE framework. These errors are related to all assumptions in Subsection 2.2. In an attempt to frame these errors, we discuss each component of TSE in relation to the RDS assumptions below.

A. Coverage Error: While RDS does not use frames directly, obviously, networks with multiple components or with loose connections result in coverage error. Moreover, people's perceived social network structure determines coverage, making the network structure assumption relevant. Recruiters' understanding of the target population (for example, jazz musicians in Heckathorn and Jeffri 2001) is critical. This equates to the boundary specification, which has long been acknowledged as a problem in the social network literature (Laumann et al. 1983).

B. Sampling Error: Sampling error results from using a sample for inference rather than the entire population for inferences. With probability sampling, estimates are (approximately) unbiased in expectation, and the sampling variance is the sole source of sampling error. However, both sampling bias and variance come into play for nonprobability samples, including RDS. Assumptions about random recruitment and equal homophily directly influence the sampling bias through the unmet equilibrium assumption. If recruitment is done systematically, then the assumptions are violated and sampling bias is likely.

C. Nonresponse Error: A violation of the complete response assumption in RDS is the same as nonresponse error in TSE. While previous research recognizes this as an uncontrollable aspect of recruitment (for example, Gile and Handcock 2010), it is addressed as a sampling issue rather than a nonresponse issue. Understanding nonresponse in RDS is a complex task. Even calculating response rates is difficult, if not impossible. This is because the denominator required for calculating response rates includes all eligible peers in a participant's network to whom recruitment is attempted. Of course, this is not the same as the number of distributed coupons, because participants may attempt to recruit without involving coupons. If coupons are not involved, the number of unsuccessful recruitment attempts is unknown.

Moreover, of the four nonresponse stages ($\boldsymbol{r}$) in Subsection 2.2.E, the current RDS practice captures information about Stages 1 and 4 only (whether participants take coupons and whether coupons are redeemed by their peers), providing partial information about nonresponse. Without monitoring the entire recruitment process, the magnitude and the effect of nonresponse cannot be ascertained.

Despite nonresponse compounded of multiple stages, there is little effort to understand nonresponse in RDS. For probability sample surveys, covariates of nonresponse have been studied extensively (for example, Groves and Couper 1998) and are incorporated through post-survey adjustments. Nonresponse follow-up studies have been recommended for RDS and implemented (for example, Gile et al. 2015), but as discussed in Section 6, the design is yet to be established to generate useful data.

D. Measurement Error: While participants being recruiters is a unique feature of RDS, they are also a unique source of measurement error, which also affects other types of error. First, their social network structure (for example, density, single vs. multiple components) and their understanding of the target population definition (that is, the boundary specification

problems in Laumann et al. 1983) both affect noncoverage error. Second, criteria recruiters use for selecting their recruits determine the selection mechanism, influencing sampling error. More importantly, $\tilde{d}_i$ is subject to measurement error, potentially affecting overall inferences. Unreasonable network size reports have also been noted (for example, Wejnert and Heckathorn 2008; Schonlau 2014) with evidence that the report of network size is sensitive to question wording (McCreesh et al. 2012; Schonlau 2014). Measurement error in $\tilde{d}_i$ has implication for the bias of $\tilde{y}$ with an unclear direction. Also, $\tilde{d}_i$ influences the variance of $\tilde{y}$: the larger the variation of $\tilde{d}_i$, the larger the variance of $\tilde{y}$.

In summary, what sets RDS apart from traditional probability or adaptive sampling is who has the control over sample selection (Frost et al. 2006). As sample selection is controlled by participants, not by researchers in RDS, statistical inferences are challenging (Frost et al. 2006), requiring a set of strong assumptions (Heimer 2005; Gile and Handcock 2010). Some studies provide cautionary remarks about RDS in terms of bias (for example, Martin et al. 2003; Wejnert and Heckathorn 2008; McCreesh et al. 2012) and variance (for example, Goel and Salganik 2010; Verdery and Mouw 2012; Verdery et al. 2015) and call for its evaluation on empirical data (for example, Heimer 2005; Burt et al. 2010; Simon and Mosavel 2010; Lu et al. 2012, Salganik 2012; Gile et al. 2015). Still, others use RDS data without considering or acknowledging potential limitations. For instance, Lee and colleagues (2011) asserted for replacing probability sampling with RDS entirely, even for general population studies.

It is important to note two clear differences between RDS and the network sampling by Sirken (1972, 1975, 1997). First, Sirken's network sampling uses well-specified networks, such as direct family members and biological siblings, whereas RDS uses loosely defined networks, such as acquaintances and friends. Second, in Sirken's network sampling, respondents provide the information about their peers that researchers use for drawing a sample. On the other hand, in RDS, participants sample on their own. Therefore, who controls the sampling process is completely different, although the word "network" may appear to suggest similarities.

Reflecting the recency of its introduction, the realities of data collection using RDS remain to be scrutinized. The scarcity of publicly available RDS data is a further impediment to methodological assessments (Salganik 2012). See Appendix Table A1 for a list of publicly available RDS data sets. This study uses two publicly available RDS data sets with recruitment information on similar topics (for example, HIV), in similar locales (for example, Chicago), and examines the realities of RDS data using the TSE framework on two specific errors: nonresponse error arising in the recruitment process and measurement error in the network size reports. We focus on these two errors, because the current practice of RDS does not provide adequate data for assessing remaining errors.

3. Data and Methods

3.1. Data

3.1.1. Overview

We use data sets from two RDS studies available from the Inter-University Consortium for Political and Social Research (ICPSR): the Sexual Acquisition and Transmission of HIV

Cooperative Agreement Program (SATHCAP) and the Latino MSM Community Involvement (LMSM). SATHCAP targeted those at high risk of HIV/AIDS (for example, IDUs, men who have sex with men) and their sexual partners in four cities: Los Angeles (LA), California; Chicago, Illinois; Raleigh-Durham, North Carolina; and St. Petersburg, Russia and was conducted in two phases using independent samples (Compton et al. 2009; Iguchi et al. 2009). SATHCAP data from ICPSR included the three US cities between November 2006 and August 2008 (Iguchi et al. 2010). LMSM was conducted in San Francisco (SF), California and Chicago, Illinois through 2003 and 2004, targeting Latino gay or bisexual men and transgenders (Ramirez-Valles 2013).

The reasons for using these data sets are three-fold. First, to the best of our knowledge, SATHCAP and LMSM are the only publicly available RDS data sources with coupon distribution information, which is necessary to trace the link between recruiters and recruits and to ascertain the recruitment process, including how many coupons were given to each recruiter and how many were redeemed by his/her recruits. Second, using two independent RDS studies on similar topics allows us to examine whether the errors and their impact replicate across studies. Third, as these studies include roughly consistent study sites, the effect of geography, that may, otherwise, confound the results, can be minimized. For geographical consistency, this study included LA and Chicago from SATHCAP and SF and Chicago from LMSM, resulting in the sample of 3,584 for SATHCAP (845 for LA and 2,739 for Chicago) and 643 for LMSM (323 for SF and 320 for Chicago). However, it should be noted that information about these studies is limited to what is publicly available. Information about, for example, decisions around incentives and sample sizes could not be verified.

3.1.2. Nonresponse Follow-Up Study

In addition to the main data collection, SATHCAP conducted a follow-up study at the time of their return visit to study sites to obtain recruitment incentives. It included questions ascertaining the number and characteristics of peers who had accepted ("accepters") and refused ("refusers") coupons from participants: for example, "How many people accepted study coupons from you?", "How many people refused to accept study coupons from you?", and "Of the [reported number] people who accepted study coupons from you, how many are friends of yours?" With this data set, accepters and refusers can be compared on various characteristics. The follow-up study participation rate was 45.2% ($n = 382$) for LA and 56.1% ($n = 1{,}537$) for Chicago.

3.1.3. Measurement of Network Size

The network size in SATHCAP was measured by combining information from the following three questions: 1) "How many people do you know personally (that is, you know their name, you know who they are, and they know you, and you have seen them in the last six months) who use heroin, methamphetamines, and/or powder or crack cocaine or who inject some other drug?"; 2) "How many people do you know personally (that is, you know their name, you know who they are and they know you and you have seen them in the last six months) who are men who have sex with men?"; and 3) "How many of the men who have sex with men that you know use heroin, methamphetamines, crack and/or powder cocaine or inject some other drug?" The network size in LMSM was based on the

answer to the question "how many Latino gay, bisexual and transgenders over 18 years old in San Francisco/Chicago do you know?"

3.2. *Analysis Procedure*

3.2.1. Nonresponse Error

Nonresponse error was first examined using coupon distribution data. By linking recruiters and their recruits, we assessed nonresponse at Stages 1 and 4 discussed in Subsection 2.2.E. We then connected nonresponse with the sample size growth and the recruitment chain length. We also examined the potential correlates of nonresponse. Specifically, we considered recruiters' age, race/ethnicity, nativity, education, income, living arrangement, HIV status, substance use, sexual behavior, incarceration, and network size and their relationship with the number of successful recruits in Poisson regression to reflect the distribution of the dependent variable. Recruitment chains are partly affected by nonresponse and could be considered as clusters. Hence, we also examined Intra-Chain Correlation (ICC) to assess homogeneity within chain.

Further, using the SATHCAP follow-up study, we examined the Stage 3 nonresponse pattern. Specifically, we compared those who accepted versus refused coupons. Usefulness of the follow-up study was assessed with respect to its own nonresponse and measurement issues.

3.2.2. Measurement Error

Measurement error of the reported network size was first examined through basic data checks. In addition to the standard weight ($\tilde{d}_i^{-1}$), we used smoothed weights by top and bottom coding the network size at its 10th and 90th percentile, adopting the idea of weight trimming routinely performed in survey sampling to minimize the effect of extreme weights (Potter 1988; Little et al. 1997). Although large weights are often discussed for trimming (for example, Elliott 2009), extreme weights include both small and large weights as they both increase variability in estimates (see Valliant et al. 2013, p. 388). In fact, some consider both small and large weights for trimming (for example, Cole and Hernan 2008; Izrael et al. 2009). As shown in Table 6, the 10th and 90th percentiles equated to a network size of 2 and 50 for LA and 3 and 50 Chicago in SATHCAP and 3 and 75 for SF and 3 and 40 for Chicago in LMSM. Weights cannot be ascertained for cases with missing or zero network sizes. However, these occurred infrequently and imputed weights on these cases made no difference in analytic results. Hence, these cases were not assigned with weights. We compared the Unequal Weighting Effect (UWE, Kish 1992) between standard and smoothed weights and examined the relationship between respondent characteristics and their weights.

We then examined the effect of weighting on estimation by comparing unweighted estimates and estimates weighted by standard and smoothed weights. We considered both univariate and bivariate statistics. For univariate statistics, we examined proportions of various sociodemographics, health status, and risk behaviors. For bivariate statistics, the associations between HIV status and characteristics known to be related to HIV (for example, substance use, sexual behavior), as well as characteristics known to be unrelated

to HIV (for example, network size) were examined through simple logistic regression that modelled HIV status on these characteristics one by one.

There are very few verified estimators and software options for RDS. Some software (for example, RDS Analyst introduced shortly) requires entire coupon distribution information and accommodates standard weights only. This study used:

(1) an unweighted naïve estimator that incorrectly assumed simple random sampling,
(2) the RDS-II estimator (Volz and Heckathorn 2008) with standard weights, and
(3) the RDS-II estimator with smoothed weights for univariate statistics.

Their standard errors were calculated using the bootstrap method in Salganik and Heckathorn (2004) and Salganik (2006). Note that, although Volz and Heckathorn (2008) introduced a variance estimator for the RDS-II (equation 17 in their article), it requires information about all network members in the population and, hence, cannot be used for sample data. Unweighted proportions and their standard errors were computed in SAS. An R package RDS by Handcock and his colleagues (2014) was used for the RDS-II. We also used RDS Analyst, a software by Hard-to-Reach Population Methods Research Group (http://www.hpmrg.org/) for the RDS-II with standard weights. The results when using the weights from RDS Analyst were virtually the same as estimates using standard weights and, hence, not presented in this article.

For bivariate statistics, there are no known or suggested estimators for logistic regression model parameters in the literature. Given this, we used proc surveylogistic in SAS with and without weights, focusing on the estimated coefficients and their significance.

4. Nonresponse Error

4.1. Recruitment Process Through Coupon Distribution and Redemption

Within each study, the recruitment started with similar numbers of seeds across cities. For SATHCAP, there were 117 seeds for LA and 132 for Chicago. LA seeds were recruited using passive recruitment (for example, flyers and advertisements) while Chicago added active recruitment (for example, study staff approaching potentially eligible community members) (Iguchi et al. 2009).

From these seeds, a total sample size of 845 was generated over 19 waves in LA and 2,739 in Chicago over 45 waves. Seeds in LMSM were recruited actively using prespecified sociodemographics criteria: country of origin, main language spoken, HIV status, gender, and sexual orientation (that is, gay, bisexual, transgender) (Ramirez-Valles et al. 2005). In LMSM, 17 seeds generated a total sample size of 323 over twelve waves in SF and 13 seeds generated 320 over nine waves in Chicago. Rows A through C in Table 1 summarize the recruitment process.

When examining the recruitment process at the recruiter level, there were 842 (row D) potential recruiters who could have taken and distributed coupons, for example, in SATHCAP LA. Among them, 769 (row E) actually took coupons and 410 (row L) generated actual recruits. This actual recruitment rate (row M) was 48.7% for LA and 55.8% for Chicago in SATHCAP and 50.0% for SF and 49.0% for Chicago in LMSM.

Up to six coupons were given to all respondents, except for those in the last wave in SATHCAP and three coupons in LMSM. However, not all potential recruiters took coupons for distribution. Rows D and E in Table 1 compare the number of potential recruiters (that is, who could have taken coupons) versus actual recruiters (that is, who took coupons). This equates to nonresponse Stage 1. The rate of actual recruiter (row F) ranged from 89.1% for LMSM SF to 96.9% for SATHCAP Chicago.

A total of 3,140, 8,245, 854, and 917 coupons were distributed in SATHCAP LA, SATHCAP Chicago, LMSM SF and LMSM Chicago (row G), equating to an average of 3.73, 3.02, 2.67, and 2.96 coupons taken by potential recruiters (row H). As expected, not all coupons were redeemed. The number of redeemed coupon per potential recruiter ranged from 0.87 for SATHCAP LA to 0.99 LMSM Chicago (row K), meaning that less than one recruit was generated per potential recruiter. Recall that each potential recruiter could have generated up to six additional recruits in the immediately subsequent wave in SATHAP and up to three in LMSM.

Table 1. Summary of recruitment process by city, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) and the Latino MSM Community Involvement (LMSM).

	SATHCAP		LMSM	
	LA	Chicago	SF	Chicago
Overall recruitment results				
A. No. of seeds	117	132	17	13
B. No. of total data collection waves	19	45	12	9
C. Total sample size (i.e., all respondents, including seeds)	845	2,739	323	320
D. No. of potential recruiters (= C – no. of last wave respondents)	842	2,735	320	310
Coupon distribution				
E. No. of actual recruiters (i.e., those who took coupons)	769	2,650	285	299
F. Rate of actual recruiters (= E/D)	91.3%	96.9%	89.1%	96.5%
G. No. of coupons taken by potential recruiters	3,140	8,245	854	917
H. Average no. of coupons taken by potential recruiters (= G/D)	3.73	3.02	2.67	2.96
Coupon redemption				
I. No. of recruits (i.e., redeemed coupons)	728	2,607	306	307
J. Coupon redemption rate (= I/G)	23.2%	31.6%	35.8%	33.5%
K. Average no. of recruits generated by potential recruiters (= I/D)	0.87	0.95	0.96	0.99
L. No. of recruiters generating recruits (i.e., those whose coupons were redeemed by peers)	410	1,526	160	152
M. Actual recruitment rate (= L/D)	48.7%	55.8%	50.0%	49.0%

4.2. Coupon Redemption Rates, Sample Sizes, and Recruitment Chain Length

As there is no viable way of measuring response rates for RDS, we used coupon redemption rates (row J of Table 1) as a proxy. Although neither complete nor perfect, this was the only measure that reflected Stage 4 nonresponse. Because there are three other stages in RDS nonresponse, coupon redemption rates reported here indicate an upper bound for the true response rates. This rate ranged from 23.2% in SATHCAP LA to 35.8% in LMSM SF.

If all potential recruiters took coupons and their peers accepted and redeemed coupons, the cumulative sample size over recruitment waves would grow exponentially and all recruitment chains would reach the same length. However, with low coupon redemption rates and a small number of recruits per potential recruiter, cumulative sample sizes in Figure 2 grew in a quadratic, not the assumed exponential, pattern and approached a stationary phase rather rapidly. This was true across cities and studies.

At the recruitment chain level, nonresponse occurred differently, resulting in differential lengths as summarized in Table 2. On average, after seeds, chains lasted for as short as 1.56 waves in SATHCAP LA and as long as 4.38 waves in LMSM Chicago. Chains lasted longer in LMSM than in SATHCAP and in Chicago than in California cities. The distribution of chain lengths of SATHCAP was highly skewed, with the medians far

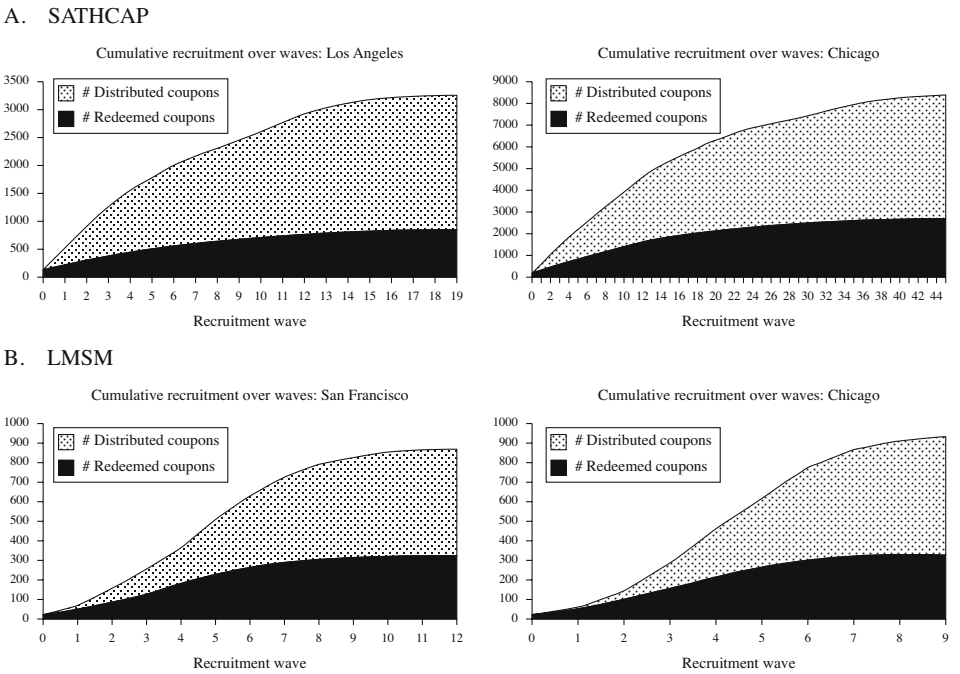

Fig. 2. Cumulative sample sizes (number of distributed coupons and redeemed coupons) by city, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) and the Latino MSM community involvement (LMSM). Note. The number of redeemed coupon equals to the sample size; Wave 0 is consisted of seeds only who did not have to redeem coupon to participate. Hence, the coupon redemption rates are inapplicable.

Table 2. Distribution of recruitment chain lengths, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) and the Latino MSM community involvement (LMSM).

	SATHCAP		LMSM	
	LA	Chicago	SF	Chicago
No. of recruitment chains (i.e., no. of seeds)	117	132	17	13
Chain length				
Average	1.56	3.39	3.76	4.38
Standard deviation	2.33	5.52	2.25	1.93
Maximum	18	44	11	8
90th percentile	4	7	9	8
75th percentile	2	4	5	7
Median	0	1	4	5
25th percentile	0	0	2	2
10th percentile	0	0	0	0
Minimum	0	0	0	0
Mode	0	0	0	0
% of chains died after seed (i.e., chain length = 0)	58.1	32.6	23.5	15.4

below the means. In fact, 58.1% of chains in SATHCAP LA died immediately after seeds without generating recruits (that is, chain length = 0), meaning no chance for incorporating respondent-driven participant selection into the sample. This rate was 32.6% for SATHCAP Chicago 23.5% and 15.4% for LMSM SF and Chicago. The length varied widely across chains; for example, chains in SATHCAP Chicago lasted anywhere from 0 to 44 waves after seeds. The smallest variation was observed for LMSM Chicago with a range of 0 to 8. While this small variation in chain lengths for LMSM Chicago indicated that individual chains made similar contributions to the overall data, the relatively short maximum chain length suggested an issue for the memorylessness of the Markov chain.

4.3. Association Between Recruiter's Characteristics and the Number of Recruits

In order to further understand the recruitment process, we examined whether characteristics of potential recruiters were associated with the number of recruits they generated in Table 3. In SATHCAP LA, younger recruiters, those with lower income (less than USD 500 a month) and men who had sex with men generated more recruits than their counterparts, while in Chicago, it was IDUs and those who had ever been incarcerated that generated more recruits. In LMSM, foreign-borns in SF and those with lower income (less than USD 15,000 a year) in Chicago generated more recruits. It was notable that recruiter's network size had virtually no effect on recruitment across studies and cities. This contradicts the view by Johnston and Sabin (2010) that seeds with large and dense networks generate more recruits. Rather, socioeconomics (for example, income) and risk behaviors (for example, MSMs) of the recruiters made a difference in recruitment. Note that these characteristics were significantly related to HIV status, one of the key outcomes in these studies (results not shown), further suggesting nonresponse bias.

Table 3. Poisson regression of number of recruits on recruiters' characteristics, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) and the Latino MSM community involvement (LMSM).

	SATHCAP			LMSM		
		LA	Chicago		SF	Chicago
Recruiter characteristics		Est.	Est.		Est.	Est.
Intercept		-0.551^{**}	-0.289^{**}		$-0.323^{\#}$	-0.183
Age	>45 vs. ≤ 45 yrs	-0.230^{*}	-0.023	>35 vs. ≤ 35 yrs	0.049	-0.202
Race/nativity	Black vs. No	0.143	-0.051	US Born vs. No	-0.531^{*}	-0.235
Education	$\leq$High school vs. $>$ HS	0.104	-0.022	$\leq$High school vs. $>$ HS	0.141	-0.169
Income	$<\$500$/mo vs. $\geq\$500$/mo	0.290^{**}	0.066	$<\$15$K/yr vs. $\geq\$15$K/yr	0.026	0.423^{**}
Living arrangement	Homeless vs. No	0.024	0.003	Live alone vs. No	0.092	-0.066
HIV+	Yes vs. No	-0.086	-0.052	Yes vs. No	0.210	0.142
Substance use	Injection drug ever vs. No	0.080	0.165^{***}	Substance use 6 mos vs. No	0.190	0.301
Sexual behavior	MSM vs. No	0.222^{*}	0.081	Bi/Transgender vs. No	0.110	-0.281
Incarcerated	Ever vs. No	0.040	0.219^{***}	–	–	–
Network size		0.001	$0.001^{\#}$		0.002	0.002

$^{\#}$Significant at $p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

4.4. Recruitment Chain Homogeneity

ICCs are reported in Table 4. Recall that ICCs are evidence of correlated responses among respondents from the same recruitment chain, which further is not consistent with the equilibrium assumption. Overall, ICCs were sizable, indicating homogeneity within chain and heterogeneity between chains. Within-chain homogeneity was larger for SATHCAP than LMSM. ICC was notably large for race in SATHCAP Chicago at 0.619 and for HIV status in SATHCAP LA at 0.490, indicating that 61.9% and 49.0% of the overall variance in these variables were due to between-chain variance.

4.5. Nonresponse Follow-Up Study

In the SATHCAP follow-up study, recruiters were asked the number of coupon accepters and refusers. On average, follow-up respondents in LA reported 3.10 peers accepting and 1.60 refusing coupons; in Chicago 2.17 accepting and 1.06 refusing. If to examine any incidence of coupon being refused or accepted reported in the follow-up study, 46.2% of follow-up respondents in LA and 31.1% in Chicago reported any of their coupons being refused by the peers, while over 97% of respondents reported any of their coupons being accepted. The fact that coupon refusal was reported implies that nonresponse did arise at this stage and the true response rates were lower than the coupon redemption rates in Subsection 4.2.

Recruiters who reported any coupons being accepted or refused were asked about the characteristics of accepters and refusers separately. Their characteristics are listed in Table 5. In both cities, the proportions of friends and sex partners were significantly higher among accepters than among refusers. For example, 87.6% of the coupon accepters were friends of recruiters, while 78.59% of the refusers were so in Chicago, a significant difference at $p < 0.001$. Coupon accepters in Chicago were less likely to be homeless and more likely to be IDUs, compared to coupon refusers (both at $p < 0.05$).

The follow-up study itself was subject to own nonresponse and measurement errors. As noted previously, about 53.5% of the potential recruiters participated in the follow-up. Given that this was conducted at the time of recruitment incentive payment, it is not surprising that follow-up study respondents had distributed more coupons than nonrespondents (4.40 vs. 3.15 for LA and 3.32 vs. 2.61 for Chicago) and were associated with a larger number of recruits (1.75 vs. 0.13 for LA and 1.66 vs. 0.04 for Chicago), all significant at $p < 0.001$, results similar to Gile et al. (2015). With logistic regression, we examined whether recruiters' characteristics beyond the number of coupons they took affected their follow-up study participation. The results in Appendix Table A2 suggested that those with lower income or with HIV were more likely to participate in the follow-up study than their counterparts in LA, while it was Black recruiters who were more likely to participate in Chicago.

Additionally, the number of accepted coupons reported by the recruiters in the follow-up study matched neither the number of coupons they took nor the number of coupons redeemed by their peers in the coupon distribution data. While recruiters in the follow-up study reported 3.10 and 2.17 coupon accepters in LA and Chicago, respectively, their coupon distribution data showed that they took 4.40 and 3.32 coupons in LA and Chicago and that 1.75 and 1.66 coupons were redeemed in LA and Chicago.

Table 4. Intra − chain correlation on respondent characteristics, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) and the Latino MSM community involvement (LMSM).

	SATHCAP			LMSM		
		LA	Chicago		LA	Chicago
Respondent characteristics		Est.	Est.		Est.	Est.
Age	>45 vs. ≤45 yrs	0.102	0.117	>35 vs. ≤35 yrs	0.117	0.133
Race/nativity	Black vs. No	0.176	0.619	US Born vs. No	0.026	0.236
Education	≤High school vs. > HS	0.108	0.017	≤High school vs. > HS	0.101	0.103
Income	<$500/mo vs. ≥$500/mo	0.124	0.000	<$15K/yr vs. ≥$15K/yr	0.114	0.099
Living arrangement	Homeless vs. No	0.171	0.064	Live alone vs. No	0.000	0.017
HIV+	Yes vs. No	0.490	0.192	Yes vs. No	0.110	0.152
Substance use	Injection drug ever vs. No	0.158	0.199	Substance use 6 mos vs. No	0.141	0.079
Sexual behavior	MSM vs. No	0.114	0.182	Bi/Transgender vs. No	0.077	0.060
Incarcerated	Ever vs. No	0.134	0.001	–	–	–

Table 5. Comparison of characteristics between coupon accepters and refusers, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP).

	LA				Chicago			
	Coupon accepters		Coupon refusers		Coupon accepters		Coupon refusers	
Characteristics	*n*	Mean (SE)	*n*	Mean (SE)	*n*	Mean (SE)	*n*	Mean (SE)
Friend (%)	351	76.5 (1.8)	163	69.3 (3.2)$^{\#}$	1,433	87.6 (0.7)	457	78.6 (1.6)***
Sex partner (%)	349	36.1 (2.1)	162	29.5 (3.1)$^{\#}$	1,428	48.5 (1.1)	453	39.8 (2.0)***
Known for 6+ months (%)	349	70.0 (2.0)	164	66.8 (3.3)	1,430	89.1 (0.7)	456	86.6 (1.4)
See daily (%)	348	60.1 (2.1)	164	57.7 (3.3)	1,429	74.4 (1.0)	455	72.0 (1.7)
Live in the same city (%)	351	76.1 (2.1)	164	74.4 (3.2)	1,429	96.5 (0.4)	452	95.5 (0.9)
Male (%)	351	75.4 (1.6)	137	74.6 (2.9)	1,430	61.5 (1.0)	382	61.2 (2.0)
Black (%)	350	30.2 (2.0)	164	26.3 (2.9)	1,431	48.9 (1.3)	456	49.1 (2.2)
Homeless (%)	346	36.1 (2.3)	160	35.0 (3.4)	1,425	18.7 (0.9)	447	23.8 (1.8)*
Injected drug together (%)	345	50.3 (2.4)	162	51.4 (3.5)	1,430	81.4 (0.9)	455	76.9 (1.7)*

$^{\#}$ Significantly different from coupon accepters at $p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

5. Measurement Error

5.1. Reported Network Size

We examined the distribution of reported network sizes in Table 6. First, in a small number of cases, networks sizes were not reported. On average, respondents reported their network sizes being in the neighborhood of 20: in SATHCAP 17.5 for LA and 21.1 for Chicago; and in LMSM 23.9 for SF and 36.7 for Chicago. The network size showed a wide variation, as small as zero and as large as 2,100; however, the median was modest at 7, 10, 10, and 11 for across cities and studies, resulting in large positive skewness. In fact, 90% of the respondents reported network sizes smaller than 50 for both cities in SATHCAP and 40 and 75 for LMSM SF and Chicago.

By default in the RDS recruitment and by the reciprocal nature of social networks, non-seed respondents should report at least one network member. This is because their recruiters considered them as a network member, and so should they. However, 43 non-seeds (5.9%) in SATHCAP LA reported zero network size, a problem reported by McCreesh et al. (2012). However, this zero network size reported by non-seeds occurred infrequently for SATHCAP Chicago and both cities in LMSM at 19 (0.7%), 3 (1.0%), and 2 (0.7%).

5.2. Reported Network Sizes for Weights

The maximum network size respondents reported was 400 and 591 for SATHCAP LA and Chicago and 1,000 and 2,100 for LMSM SF and Chicago. While not impossible, it is difficult to imagine a LMSM respondent in Chicago knowing exactly 2,100 Latino gay, bisexual, and transgenders over 18 years old in Chicago. This observation had an implication for inference because the inverse of network size was used as weights, which

Table 6. Distribution of reported network size, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) and the Latino MSM community involvement (LMSM).

	SATHCAP		LMSM	
	LA	Chicago	SF	Chicago
n	845	2,739	323	320
No. of cases with missing network size	12	12	0	0
Reported network size				
Average	17.5	21.1	23.9	36.7
Standard deviation	33.6	36.4	69.2	128.2
Maximum	400	591	1,000	2,100
90th percentile	50	50	40	75
75th percentile	16	25	20	30
Median	7	10	10	11
25th percentile	3	6	5	5
10th percentile	2	3	3	3
Minimum	0	0	0	0
Mode	2	10	2	10
No. of non-seeds with 0 network size	43	19	3	2

ranged from 0.0005 (= 1/2,100) to 1.0000 (= 1/1) in LMSM Chicago. Weight dispersion resulted in UWEs of 2.19 and 2.29 for SATHCAP LA and Chicago and 2.21 and 2.57 for LMSM SF and Chicago. Weight smoothing reduced UWEs substantially to 1.73, 1.62, 1.61, and 1.76 for respective study and city. This is not surprising given that the weights in LMSM Chicago, for example, varied from 0.0005 to 1.000 without smoothing but in a smaller range, from 0.0133 (= 1/75) to 0.3333 (= 1/3), with smoothing. In Appendix Table A3, we examined respondents' characteristics associated with both weights. Overall, weights were related to certain respondent characteristics, and this relationship persisted regardless of weight smoothing.

5.3. Effects of Weights in Estimation

We used three types of estimation approaches: 1) unweighted; 2) weighted with standard weights; and 3) weighted with smoothed weights. The focus of this section is on whether the weights affected estimates and their variabilities or significance. Table 7A includes estimated proportions for various sociodemographic and health risk variables and their standard errors. Table 7B includes estimates of coefficients in simple logistic regression of HIV status and their *p*-values.

In Table 7A, weights changed univariate statistics. Estimates affected the most by weights were characteristics that were significantly related to weights in Appendix Table A3. For instance, incarceration, a significant covariate of weights in SATHCAP LA, changed from 66.4% (unweighted) to 61.6% (with standard weights), and to 61.1% (with smoothed weights). For LMSM, HIV status was a significant covariate of weights in Chicago, and the HIV+ rate decreased from 21.5% (unweighted) to 17.1% (with smoothed weights), and to 14.3% (with standard weights). Not surprisingly, with weights, standard errors increased by a factor of 1.5 to 2.

Weights affected logistic regression coefficients in Table 7B both substantively and statistically. Risk factors known to be highly related to HIV status showed mixed results depending on the estimation approaches. For example, injection drug use was estimated to be significant in SATHCAP Chicago regardless of weights. However, in SATHCAP LA, it was not significant when unweighted or with smoothed weights but marginally significant with standard weights in the direction opposite of what one would expect: injection drug use was negatively related to HIV+. In LMSM, regardless of the estimation approaches, substance use was not a significant predictor in SF, but was a significant predictor in Chicago without weights or smoothed weights. MSM in SATHCAP was a significant predictor regardless of weights in both cities. Whether someone had STD in LMSM was a significant predictor of HIV+ for Chicago consistently across approaches, but was not so in SF when applying standard weights. Significance of network size in SATHCAP varied depending on the approaches. In LMSM, network size was a consistently significant predictor in Chicago, but insignificant in SF.

6. Discussion

Our study showed 1) that there existed nonresponse and measurement errors pertinent to the assumptions and practices of RDS; 2) that these errors had implications for inferences; and 3) that this was observed commonly in two independent RDS studies.

Table 7. Comparison of unweighted and weighted estimates using standard and smoothed weights, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) and the Latino MSM community involvement (LMSM).

A. Univariate statistics: proportions.

	SATHCAP: LA						SATHCAP: Chicago					
	Unweighted ($n = 845$)		Weighted, standard ($n = 802$)		Weighted, smoothed ($n = 845$)		Unweighted ($n = 2{,}739$)		Weighted, standard ($n = 2{,}720$)		Weighted, smoothed ($n = 2{,}739$)	
Characteristics	%	SE (%)	%	SE (%)	%	SE (%)	%	SE (%)	%	SE (%)	%	SE (%)
Age $<$45 yrs old	54.3	1.7	51.7	3.1	50.8	2.8	47.1	1.0	48.2	1.7	47.1	1.5
Race: Black	48.4	1.7	46.0	3.5	47.4	3.5	81.5	0.7	79.4	1.7	79.8	1.5
$\leq$ High school	59.5	1.7	64.9	3.1	63.0	2.6	72.6	0.9	75.7	1.4	75.4	1.1
Homeless	56.2	1.7	52.5	3.3	53.8	2.9	38.6	0.9	36.1	1.4	36.1	1.3
Income $<$\$500/ mo	61.9	1.7	62.8	3.0	63.1	2.6	70.3	0.9	71.8	1.4	71.2	1.1
HIV+	30.1	1.6	29.3	3.7	28.1	2.8	9.0	0.5	8.0	0.8	8.8	0.8
MSM	59.9	1.7	56.3	3.2	55.2	2.9	21.0	0.8	19.2	1.4	19.8	1.2
Inject drug ever	47.7	1.7	42.0	3.4	42.3	2.7	41.3	0.9	39.9	1.7	40.8	1.4
Ever prison	66.4	1.6	61.6	2.4	61.1	2.1	75.3	0.8	72.6	1.2	72.7	1.0
	LMMS: SF						LMMS: Chicago					
	Unweighted ($n = 323$)		Weighted, standard ($n = 320$)		Weighted, smoothed ($n = 323$)		Unweighted ($n = 320$)		Weighted, standard ($n = 318$)		Weighted, smoothed ($n = 320$)	
Characteristics	%	SE (%)	%	SE (%)	%	SE (%)	%	SE (%)	%	SE (%)	%	SE (%)
Age $<$35 yrs old	41.2	2.7	42.4	4.9	40.7	4.7	60.0	2.7	57.6	5.1	58.0	5.0
US Born	14.2	1.9	19.7	2.7	16.8	2.5	30.9	2.6	30.5	5.0	32.0	4.6
$\leq$ High school	47.1	2.8	48.9	4.4	49.7	4.0	52.8	2.8	54.4	5.1	54.9	4.5
Income $<$\$15K/yr	62.2	2.7	67.6	4.1	65.7	3.5	50.0	2.8	52.3	4.6	51.1	4.2
Live alone	22.6	2.3	19.1	3.3	20.3	2.9	26.3	2.5	23.3	4.1	23.8	3.8
HIV+	38.0	2.8	37.3	4.9	37.0	4.2	21.5	2.5	14.3	6.2	17.1	4.6
Bi/Transgender	33.5	2.6	40.5	4.4	39.6	4.3	27.2	2.5	31.4	4.6	32.7	4.0
STD	9.9	1.7	7.8	1.7	8.7	1.8	13.8	1.9	14.2	3.7	13.3	3.0
Substance 6 mos	42.7	2.8	40.8	3.9	40.8	3.7	52.8	2.8	56.0	4.8	54.9	4.0

B. Bivariate statistics: coefficients of simple logistic regression (dependent variable: HIV+).

	SATHCAP: LA						SATHCAP: Chicago					
	Unweighted		Weighted, standard		Weighted, smoothed		Unweighted		Weighted, standard		Weighted, smoothed	
Independent variable	Est.	*p-val.*	Est.	*p-val..*	Est.	*p-val.*	Est.	*p-val.*	Est.	*p-val.*	Est.	*p-val.*
Inject drug vs. No	0.065	0.668	−0.385	0.084	−0.261	0.187	0.416	0.002	0.503	0.007	0.393	0.021
MSM vs. No	1.426	<.001	1.143	<.001	1.436	<.001	1.442	<.001	1.692	<.001	1.611	<.001
Network size	0.004	0.081	0.005	0.404	0.006	0.037	0.001	0.436	0.007	0.031	0.002	0.191
	LMMS: SF						LMMS: Chicago					
	Unweighted		Weighted, standard		Weighted, smoothed		Unweighted		Weighted, standard		Weighted, smoothed	
Independent variable	Est.	*p-val.*	Est.	*p-val.*	Est.	*p-val.*	Est.	*p-val.*	Est.	*p-val.*	Est.	*p-val.*
Substance 6 mos vs. No	0.225	0.350	0.083	0.817	0.187	0.539	0.790	0.011	0.624	0.199	0.741	0.088
STD vs. No	0.651	0.009	0.607	0.144	0.967	0.003	1.432	<.0001	1.547	0.006	1.717	0.001
Network size	−0.002	0.280	0.001	0.847	−0.001	0.546	0.006	0.015	0.017	0.023	0.007	0.032

6.1. Summary

Nonresponse in the recruitment process impacted not only the sample size growth but also the recruitment chain length. The assumed exponential growth was far from the reality, and a substantial proportion of chains died immediately after seeds. Moreover, coupon distribution data as well as follow-up data suggested that nonresponse did not occur at random. First, closeness of the relationship between participants and their peers influenced peers' coupon acceptance. Proportions of friends and sex partners were significantly larger among coupon accepters than among refusers by about ten percent points. Second, participants with certain characteristics, most notably lower income, generated more recruits than the counterpart. This systematic nonresponse, nonrandom recruitment pattern and unequal chain length, when combined with large ICCs, further suggest that the Markov chain is not achieved in the practice of RDS.

Self-reported network sizes showed a wide variation with some unrealistic extreme values, strong evidence for measurement error. This measurement error is of concern on its own, of course. In RDS, this is also of concern for inference: as the inverse of network sizes is used as weights, the accuracy of the report matters. In particular, their variability means variability in weights, which, in turn, decreases efficiency of estimates shown through UWE that ranged around two in our analysis. Weights changed estimated prevalence in directions that were not entirely explainable. It is true that whether a person reports 2,100 or 2,150 for the network size has a little effect on the weight assigned to this person, with both resulting in a weight of 0.0005. However, whether a person reports 1 versus 50 does have an effect on the weight, with the weight being 1 versus 0.02. Moreover, in principle, while the current RDS estimators of prevalence attempt to take a form of model-based estimation (Valliant 2013), the information used in the estimation is subject to measurement error, making the estimators inadequate to account for such an error. These may hamper inferences in an unknown direction.

One may argue that the purpose of RDS is to study relationships between variables, not to estimate prevalence. Our analysis of simple models that regressed HIV status on various characteristics with different applications of weights (for example, unweighted, standard weights, smoothed weights) showed unexplainable patterns. For instance, in SATHCAP, injection drug use was a significant and positive predictor of HIV status regardless of weights in Chicago; but in LA, it was not a significant predictor without weights and with smoothed weights, and was a marginally significant and *negative* predictor with standard weights. While one may suspect that applying weights, particularly standard weights, would decrease the significance of covariates due to increased variability of their estimates, this was not always the case. Overall, weights did affect the inferences about bivariate relationships, but in a yet unexplainable and, in some cases, unreasonable way.

Moreover, participants' network sizes played no role in their recruitment success (Table 3), providing no support for the theoretical rationale of using them as weights in RDS to account for unequal selection probabilities. The significant relationships between the weights and respondent sociodemographic characteristics (for example, younger age), as well as outcome variables (for example, HIV+) in Appendix Table A3 make it very difficult to understand what these weights are adjusting for.

While errors examined in this article are important for understanding the key assumptions in RDS that further affect sampling productivity and inference, they are considered in neither the data collection nor the inferences. It is true that there is no practical and clear solution for sampling rare, hidden, and elusive populations. However, with obvious violations of these assumptions shown in this study, it is questionable how long the lack of practical solutions for sampling rare populations can be used as a justification of practicing RDS without improving design features that may minimize the effects of these breakdowns or accounting for them.

Undoubtedly, this study is limited in a number of ways. First, it addressed only two of four components of TSE, because existing RDS data do not provide information about remaining errors. Even with the two errors examined in this study, the breadth of examination was bounded by data availability. While the implications of nonresponse patterns and the effect of potential measurement error in network sizes on inferences were consistent between the two data sources, they may be specific only to these two studies. It should, however, be noted that the majority of methodological studies of RDS rely on a single data source or data that are not publicly available (for example, Wejnert and Heckathorn, 2008; McCreesh et al. 2012; Gile et al. 2015), making replication difficult, if not impossible. Rather than using the findings from this article against RDS, it would be productive to take them to develop a new framework for evaluating and improving RDS data collection practices and inferences.

6.2. *Open Questions*

With RDS, until a clear guidance is developed for assessing errors in RDS and for improving inferences, we may run the risk of mischaracterizing the hidden, rare and elusive populations, unintentionally negatively impacting these groups. In this section, we pose questions about diagnostics and estimations that may be considered in improving RDS.

On diagnostics, a recent study by Gile and her colleagues (2015) provides a set of approaches for examining RDS assumptions. As one of the first focusing on diagnostics, their study is innovative. However, their approaches rely heavily on follow-up interviews, which our study found not free from own nonresponse and measurement errors. For example, participants who did not recruit their peers were less likely to participate in the follow-up study than the counterpart. Questions used in their follow-up study were difficult to answer. For example, a question, "How many people did you try to give a coupon but they had already participated in the study?" was used to study failed recruitment attempts. This question assumes that participants are familiar with the recruitment status of their peers and/or are able to recall the number of own recruitment attempts. Other questions in their study include, "How many people do you know who have used illegal drugs in the past three months?", "If we were to give you as many coupons as you wanted, how many of these drug users do you think you could give a coupon to by this time tomorrow?" and "What is the principal reason why these persons did not accept a coupon?" Undoubtedly, these questions are difficult as respondents simply may not have information for them (for example, peers' illegal drug use). Data from such questions are not free from measurement error and may not provide meaningful information for understanding the recruitment process.

While follow-up interviews are a logical and attractive option for studying nonresponse error, their design cannot be taken lightly with respect to the types of questions and the timing of the follow-up interview. It would be advantageous to consider questions that provide meaningful data for investigating nonresponse, yet with little room for measurement error. For the timing, it would be ideal to pick a time that is reasonably long after the main interview so that all recruitment efforts can be captured, yet reasonably short so that recall does not become overly demanding. While the follow-up study in Gile et al. (2015) was conducted within one week after the main interview, our analysis of SATHCAP and LMSM showed that the average time gap between participants' main interview and their peers' interview was 14 to 46 days. Timing of the follow-up study should be informed either by the time gap observed in the field or by recruitment protocol designs (for example, assigning expiration dates to the coupons and conducting follow-up studies shortly after the expiration).

It would be ideal to account for these errors in inference. Despite the systematic nature of nonresponse examined in this article, there were no variables that explained nonresponse commonly across cities and studies. Hence, more organized efforts should be made to understand this mechanism. While the idea of accounting for unequal selection probabilities through weighting by network sizes, their measurement error needs to be addressed. One may consider using estimated network sizes through appropriate models, such as variants of the Fay-Herriot model (Fay and Herriot 1979) or those used by Beaumont (2008). Additionally, for increasing accuracy in network size measurements, the scale-up method for estimating network structures through specific questions (McCarthy et al. 2001; Zheng et al. 2006) may serve as a reasonable approach.

Appendix

Table A1. Publicly available respondent driven sampling data sets through Interuniversity Consortium for Political and Social Research (ICPSR)[§] as of January, 2016.

Study name (ICPSR study number)	Year of data collection	Year of data release	Coupon information
1. Information on Artists (ICPSR 35585)	1989, 1997, 2004, 2006–2007, 2009–2010,	2015	Not available
2. Study of Jazz Artists [United States] (ICPSR 35593)	2001	2015	Not available
3. Latino MSM Community Involvement: HIV Protective Effects (ICPSR 34385)	2003–2004	2013	Available
4. Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) [United States] (ICPSR 29181)	2005–2006, 2006–2008	2010	Available
5. The Commercial Sexual Exploitation of Children in New York City, 1982–2007 (ICPSR 34657)	2006–2007	2015	Available
6. Dynamics of Retail Methamphetamine Markets in New York City (ICPSR 29821)	2007–2009	2014	Not available
7. Health Consequences of Long-Term Injection Heroin Use Among Aging Mexican American Men in Houston, Texas (ICPSR 34896)	2008–2011	2014	Not available
8. Social Justice Sexuality Project: 2010 National Survey, including Puerto Rico (ICPSR 34363)	2010	2013	Not available

[§] Interuniversity Consortium for Political and Social Research (ICPSR) is a major data archive for social science research (*https://www.icpsr.umich.edu/icpsrweb/landing.jsp*). To our best knowledge, there are no other publicly available data using RDS located outside of ICPSR.

Table A2. Logistic regression of follow-up study participation on respondent characteristics, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP).

		LA	Chicago
Respondent characteristics		Est.	Est.
Intercept		-2.133^{***}	-1.883^{***}
Age	>45 vs. ≤45 yrs	−0.010	−0.069
Race/nativity	Black vs. No	0.190	0.365^{*}
Education	≤High school vs. > HS	−0.143	0.054
Income	<$500/mo vs. ≥$500/mo	$0.384^{\#}$	0.120
Living arrangement	Homeless vs. No	0.078	0.023
HIV+	Yes vs. No	0.735^{***}	0.231
Substance use	Injection drug ever vs. No	−0.024	0.133
Sexual behavior	MSM vs. No	−0.140	−0.024
Incarcerated	Ever vs. No	−0.068	−0.006
Network size		0.004	0.001
No. coupons		0.385^{***}	0.555^{***}

[#] Significant at $p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table A3. Linear regression of log transformed standard and smoothed weights on respondent characteristics, the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) and the Latino MSM Community Involvement (LMSM).

A. Standard and smoothed weights in SATHCAP.

		Dependent variable: standard weight		Dependent variable: smoothed weight	
		LA	Chicago	LA	Chicago
Respondent Characteristics		Est.	Est.	Est.	Est.
Intercept		-1.605^{***}	-1.995^{***}	-1.633^{***}	-2.045^{***}
Age	>45 vs. ≤ 45 yrs	-0.411^{***}	-0.069	-0.365^{***}	$-0.072^{\#}$
Race/nativity	Black vs. No	-0.220^{*}	-0.322^{***}	-0.201^{*}	-0.279^{***}
Education	$\leq$ High school vs. $>$ HS	0.246^{*}	0.137^{**}	0.210^{*}	0.133^{**}
Income	$<$\$500/mo vs. $\geq$\$500/mo	0.251^{*}	0.059	0.228^{*}	0.054
Living arrangement	Homeless vs. No	-0.164	-0.144^{**}	-0.120	-0.126^{**}
HIV+	Yes vs. No	0.071	0.014	0.081	0.046
Substance use	Injection drug ever vs. No	-0.228^{*}	-0.068	-0.200^{*}	-0.054
Sexual behavior	MSM vs. No	-0.271^{*}	-0.006	-0.263^{**}	-0.004
Incarcerated	Ever vs. No	-0.426^{***}	-0.201^{***}	-0.371^{***}	-0.176^{***}
No. distributed coupons		0.039	-0.028	0.029	-0.026

$^{\#}$Significant at $p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

B. Standard and smoothed weights in LMSM.

Respondent characteristics		Dependent variable: standard weight		Dependent variable: smoothed weight	
		SF	Chicago	SF	Chicago
		Est.	Est.	Est.	Est.
Intercept		−2.768***	−1.720***	−2.706***	−1.704***
Age	>35 vs. ≤35 yrs	−0.051	−0.347*	−0.072	−0.331*
Race/nativity	US Born vs. No	0.290	0.128	0.201	0.067
Education	≤High school vs. > HS	0.095	0.292#	0.061	0.254#
Income	<$15K/yr vs. ≥$15K/yr	0.203	−0.051	0.177	−0.036
Living arrangement	Live alone vs. No	−0.213	−0.120	−0.114	−0.081
HIV+	Yes vs. No	0.057	−0.768***	0.017	−0.628***
Substance use	Substance use 6 mos vs. No	−0.031	−0.134	0.016	−0.138
Sexual behavior	Bi/Transgender vs. No	0.157	0.165	0.139	0.142
No. distributed coupons		0.044	−0.248	0.051	−0.247#

#Significant at $p < 0.1$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

7. References

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1(2): 90–143. Doi: https://doi.org/10.1093/jssam/smt008.

Beaumont, J.-F. 2008. "A New Approach to Weighting and Inference in Sample Surveys." *Biometrika* 95(3): 539–553. Doi: https://doi.org/10.1093/biomet/asn028.

Burt, R.D., H. Hagan, K. Sabin, and H. Thiede. 2010. "Evaluating Respondent-Driven Sampling in a Major Metropolitan Area: Comparing Injection Drug Users in the 2005 Seattle Area National HIV Behavioral Surveillance System Survey with Participants in the RACEN and Kiwi Studies." *Annals of Epidemiology* 20(2): 159–167. Doi: https://doi.org/10.1016/j.annepidem.2009.10.002.

Centers for Disease Control and Prevention (CDC). 2009. HIV-Associated Behaviors Among Injecting-Drug Users—23 Cities, United States, May 2005–February 2006. Morbidity and Mortality Weekly Report, 58, 329–332. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5813a1.htm (accessed September 2015).

Centers for Disease Control and Prevention (CDC). 2013. *National HIV Behavioral Surveillance System Round 4: Model Surveillance Protocol*. Available at: http://www.cdc.gov/hiv/pdf/NHBS_Round4ModelSurveillanceProtocol.pdf (accessed September 2015).

Cole, S.R. and M.A. Hernan. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168: 656–664. Doi: https://doi.org/10.1093/aje/kwn164.

Compton, W., J. Normand, and E. Lambert. 2009. "Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP)." *Journal of Urban Health* 86(1): 1–4. Doi: https://doi.org/10.1007/s11524-009-9373-4.

Constantine, M. 2010. "Disentangling Methodologies: The Ethics of Traditional Sampling Methodologies, Community-Based Participatory Research, and Respondent-Drive Sampling." *American Journal of Bioethics* 10(3): 22–24. Doi: https://doi.org/10.1080/15265160903585628.

Dombrowski, R., B. Khan, J. Moses, E. Channell, and E. Misshula. 2013. "Assessing Respondent Driven Sampling for Network Studies in Ethnographic Contexts." *Advances in Anthropology* 3(1): 1–9. Doi: https://doi.org/10.4236/aa.2013.31001.

Elliott, M.R. 2009. "Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models." *Journal of Official Statistics* 25(1): 1–21. Doi: https://doi.org/10.1.1.552.9050.

Fay, R.E. and R.A. Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of American Statistical Association* 74: 269–277. Doi: https://doi.org/10.2307/2286322.

Frost, S.D.W., K.C. Brouwer, M.A.F. Cruz, R. Ramos, M.E. Ramos, R.M. Lozada, C. Magis-Rodriguez, and S.A. Strathdee. 2006. "Respondent-Driven Sampling of Injection Drug Users in Two U.S.-Mexico Border Cities: Recruitment Dynamics and Impact on Estimates of HIV and Syphilis Prevalence." *Journal of Urban Health* 83(1): 83–97. Doi: https://doi.org/10.1007/s11524-006-9104-z.

Gile, K.J. 2011. "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation." *Journal of American Statistical Association* 106(493): 135–146. Doi: https://doi.org/10.1198/jasa.2011.ap09475.

Gile, K.J. and M.S. Handcock. 2010. "Respondent-Driven Sampling: An Assessment of Current Methodology." *Sociological Methodology* 40(1): 286–327. Doi: https://doi.org/10.1111/j.1467-9531.2010.01223.x.

Gile, K.J., L.G. Johnston, and M.J. Salganik. 2015. "Diagnostics for Respondent-Driven Sampling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(1): 241–269. Doi: https://doi.org/10.1111/rssa.12059.

Goel, S. and M.J. Salganik. 2010. "Assessing Respondent-Driven Sampling." *Proceedings of the National Academy of Sciences of the United States of America* 107(15): 6743–6747. Doi: https://doi.org/10.1073/pnas.1000261107. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2872407/ (accessed September 2015).

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

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Surveys*. New York: Wiley.

Handcock, M.S. 2012. *Estimating the Size of Hard-to-Reach Populations Using Respondent-Driven Sampling Data*. Paper for the International Conference on Methods for Surveying and Enumerating Hard–to-Reach Populations, October 31–November 3, New Orleans, LA.

Handcock, M.S., K.J. Gile, I.E. Fellows, and W.W. Neeley. 2014. *Package 'RDS.'* Available at: http://cran.r-project.org/web/packages/RDS/RDS.pdf (accessed September 2015).

Heckathorn, D.D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Society for the Study of Social Problems* 44(2): 174–199. Doi: https://doi.org/10.2307/3096941.

Heckathorn, D.D. 2002. "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations." *Social Problems* 49(1): 11–34. Doi: https://doi.org/10.1525/sp. 2002.49.1.11.

Heckathorn, D.D. and J. Jeffri. 2001. "Finding the Beat: Using Respondent-Driven Sampling to Study Jazz Musicians." *Poetics* 28(4): 307–329. Doi: https://doi.org/10.1016/S0304-422X(01)80006-1.

Heimer, R. 2005. "Critical Issues and Further Questions About Respondent-Driven Sampling: Comment on Ramierz-Valles et al. (2005)." *AIDS and Behavior* 9(4): 403–408. Doi: https://doi.org/10.1007/s10461-005-9030-1.

Iguchi, M.Y., S.H. Berry, A.J. Ober, T. Fain, D.D. Heckathorn, P.M. Gorbach, R. Heimer, A. Kozlov, L.J. Ouellet, S. Shoptaw, and W. Zule. 2010. *Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) 2006–2008 [United States]*. ICPSR29181-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. Doi: https://doi.org/10.3886/ICPSR29181.

Iguchi, M.Y., A.J. Ober, S.H. Berry, T. Fain, D.D. Heckathorn, P.M. Gorbach, R. Heimer, A. Kozlov, L.J. Ouellet, S. Shoptaw, and W.A. Zule. 2009. "Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia Using Respondent-Driven Sampling: Sampling Methods and Implications." *Journal of Urban Health* 88(1): 5–31. Doi: https://doi.org/10.1007/s11524-009-9365-4.

Izrael, D., M. Battaglia, and M. Frankel. 2009. "Extreme Survey Weight Adjustment as a Component of Sample Balancing (a.k.a. Raking)." *Proceedings from the 2009 SAS Global Forum*. Cary, NC: SAS Institute. Available at: http://abtassociates.com/AbtAssociates/files/c1/c1bc376c-1931-4721-b71c-cb823a0fe809.pdf (accessed January 2017).

Johnston, L.G. and K. Sabin. 2010. "Sampling Hard-to-Reach Populations with Respondent Driven Sampling." *Methodological Innovations Online* 5(2): 38–48. Doi: https://doi.org/10.4256/mio.2010.0017.

Kish, L. 1992. "Weighting for unequal Pi." *Journal of Official Statistics* 8(2): 183–200.

Lansky, A., A. Abdul-Quader, M. Cribbin, T. Hall, T.J. Finlayson, R.S. Garfein, L.S. Lin, and P.S. Sullivan. 2007. "Developing an HIV Behavioral Surveillance System for Injecting Drug Users: The National HIV Behavioral Surveillance System." *Public Health Reports* 122: 48–55. Doi: https://doi.org/10.1177/00333549071220S108.

Laumann, E.O., P.V. Marsden, and D. Prensky. 1983. "The Boundary Specification Problem in Network Analysis." In *Applied Network Analysis. A Methodological Introduction*, edited by R.S. Burt and M.J. Minor. 18–34. Beverly Hills, CA: Sage.

Lee, S. 2009. "Understanding Respondent Driven Sampling from a Total Survey Error Perspective." *Survey Practice*. Available at: http://www.surveypractice.org/index.php/SurveyPractice/article/view/187/html (accessed September 2015).

Lee, R., J. Ranaldi, M. Cummings, J.N. Crucetti, H. Stratton, and L.-A. McNutt. 2011. "Given the Increasing Bias in Random Digit Dial Sampling, Could Respondent-Driven Sampling be a Practical Alternative?" *Annals of Epidemiology* 21(4): 272–279. Doi: https://doi.org/10.1016/j.annepidem.2010.11.018.

Lee, S., Z.T. Suzer-Gurtekin, J. Wagner, and R. Valliant. 2012. *Exploring Error Properties of Respondent Driven Sampling*. Paper presented at the Joint Statistical Meeting, July 28–August 2, San Diego, CA.

Lin, L., T. Finlayson, R. Iachan, M.C.B. Mendoza, and C. Wejnert. 2013. "Sampling Designs for Populations at High Risk for HIV." Paper presented at the Joint Statistical Meeting, August 3–August 8, Montréal, Canada.

Little, R.J.A., S. Lewitzky, S. Heeringa, J. Lepkowski, and R.C. Kessler. 1997. "Assessment of Weighting Methodology of the National Comorbidity Survey." *American Journal of Epidemiology* 146(5): 439–449. Doi: https://doi.org/10.1093/oxfordjournals.aje.a009297.

Lu, X., L. Bengtsson, T. Britton, M. Camitz, B.J. Kim, A. Thorson, and F. Liljeros. 2012. "The Sensitivity of Respondent-Driven Sampling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175(1): 1–26. Doi: https://doi.org/10.1111/j.1467-985X.2011.00711.x.

Marsden, P.V. 1990. "Network Data and Measurement." *Annual Review Sociology* 16: 435–463. Doi: https://doi.org/10.1146/annurev.so.16.080190.002251.

Martin, J.L., J. Wiley, and D. Osmond. 2003. "Social Networks and Unobserved Heterogeneity in Risk for AIDS." *Population Research and Policy Review* 22(1): 65–90. Doi: https://doi.org/10.1023/A:1023509211339.

McCarty, C., P.D. Killworth, H.R. Bernard, E.C. Johnsen, and G.A. Shelley. 2001. "Comparing Two Methods for Estimating Network Size." *Human Organization* 60: 28–39. Doi: https://doi.org/10.17730/humo.60.1.efx5t9gjtgmga73y.

McCreesh, N., S.D. Frost, J. Seeley, J. Katongole, M.N. Tarsh, R. Ndunguse, F. Jichi, N.L. Lunel, D. Maher, L.G. Johnston, P. Sonnenberg, A.J. Copas, R.J. Hayes, and R.G. White. 2012. "Evaluation of Respondent-Driven Sampling." *Epidemiology* 23(1): 138–147. Doi: https://doi.org/10.1097/EDE.0b013e31823ac17c.

Montealegre, J.R., J.M. Risser, B.J. Selwyn, S.A. McCurdy, and K. Sabin. 2013. "Effectiveness of Respondent Driven Sampling to Recruit Undocumented Central American Immigrant Women in Houston, Texas for an HIV Behavioral Survey." *AIDS and Behavior* 17(2): 719–727. Doi: https://doi.org/10.1007/s10461-012-0306-y.

Phillips, T. 2010. "Protecting the Subject: PDR and the Potential for Compromised Consent." *American Journal of Bioethics* 10(3): 14–15. Doi: https://doi.org/10.1080/15265160903585602.

Potter, F. 1988. "Survey of Procedures to Control Extreme Sampling Weights." *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 453–458. Available at: http://www.websm.org/uploadi/editor/1368363852Potter_1988_Survey_of_procedures_to_control_extreme_sampling_weights.pdf (accessed January 2017).

Ramirez-Valles, J. 2013. *Latino MSM Community Involvement: HIV Protective Effects*. ICPSR34385-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. Doi: https://doi.org/10.3886/ICPSR34385.v1.

Ramirez-Valles, J., D.D. Heckathorn, R. Vázquez, R.M. Diaz, and R.T. Campbell. 2005. "From Networks to Populations: the Development and Application of Respondent-Driven Sampling Among IDUs and Latino Gay Men." *AIDS and Behavior* 9(4): 387–402. Doi: https://doi.org/10.1007/s10461-005-9012-3.

Salganik, M.J. 2006. "Variance Estimation, Design Effects and Sample Size Calculations for Respondent Driven Sampling." *Journal of Urban Health* 83(7): 98–112. Doi: https://doi.org/10.1007/s11524-006-9106-x.

Salganik, M. 2012. "Commentary: Respondent-Driven Sampling in the Real World." *Epidemiology* 23(1): 148–150. Doi: https://doi.org/10.1097/EDE.0b013e31823b6979.

Salganik, M.J. and D.D. Heckathron. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34: 193–239. Doi: https://doi.org/10.1111/j.0081-1750.2004.00152.x.

Schonlau, M. 2014. "Recruiting an Internet Panel Using Respondent Driven Sampling." *Journal of Official Statistics* 30(2): 291–310. Doi: https://doi.org/10.2478/jos-2014-0018.

Simon, C. and M. Mosavel. 2010. "Community Members as Recruiters of Human Subjects: Ethical Considerations." *American Journal of Bioethics* 10(3): 3–11. Doi: https://doi.org/10.1080/15265160903585578.

Sirken, M.G. 1972. "Stratified Sample Surveys with Multiplicity." *Journal of American Statistical Association* 67: 224–227. Doi: https://doi.org/10.1080/01621459.1972.10481236.

Sirken, M.G. 1975. "Network Surveys of Rare and Sensitive Conditions." *Advances in Health Survey Research Methods, NCHSR Research Proceedings* 31. Hyattsville, MD: National Center Health Statistics.

Sirken, M.G. 1997. "Network Sampling." In *Encyclopedia of Biostatistics*, edited by P. Armitage and T. Colton, 2977–2986. Hoboken, NJ: Wiley & Sons.

Valliant, R. 2013. "Comment." *Journal of Survey Statistics and Methodology* 1(2): 105–111. Doi: https://doi.org/10.1093/jssam/smt010.

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Verdery, A.M. and T.D. Mouw. 2012. *Estimated Sampling Variance in Respondent Driven Sampling Data: Mathematical Derivations, Simulated Tests on Empirical Data, and Evidence from Other Forms of Chain-Referral Data Collection*. Paper for the International Conference on Methods for Surveying and Enumerating Hard-to-Reach Populations, October 31–November 3, New Orleans, LA.

Verdery, A.M., T.D. Mouw, S. Bauldry, and P.J. Mucha. 2015. "Network Structure and Biased Variance Estimation in Respondent Driven Sampling." *PLOS ONE* 10(12): e0145296. Doi: http://dx.doi.org/10.1371/journal.pone.0145296.

Volz, E. and D.D. Heckathorn. 2008. "Probability Based Estimation Theory for Respondent Driven Sampling." *Journal of Official Statistics* 24(1): 79–97.

Wejnert, C. and D.D. Heckathorn. 2008. "Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research." *Sociological Methods and Research* 37: 105–134. Doi: https://doi.org/10.1177/0049124108318333.

Zheng, T., M.J. Salganik, and A. Gelman. 2006. "How Many People Do You Know in Prison? Using Overdispersion in Count Data to Estimate Social Structure in Networks" *Journal of American Statistical Association* 101(474): 409–423. Doi: https://doi.org/10.1198/016214505000000116.

Received January 2016
Revised February 2017
Accepted March 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 367–383, http://dx.doi.org/10.1515/JOS-2017-0018

Using Linked Survey Paradata to Improve Sampling Strategies in the Medical Expenditure Panel Survey

Lisa B. Mirel[1] *and Sadeq R. Chowdhury*[2]

Using paradata from a prior survey that is linked to a new survey can help a survey organization develop more effective sampling strategies. One example of this type of linkage or subsampling is between the National Health Interview Survey (NHIS) and the Medical Expenditure Panel Survey (MEPS). MEPS is a nationally representative sample of the U.S. civilian, noninstitutionalized population based on a complex multi-stage sample design. Each year a new sample is drawn as a subsample of households from the prior year's NHIS. The main objective of this article is to examine how paradata from a prior survey can be used in developing a sampling scheme in a subsequent survey. A framework for optimal allocation of the sample in substrata formed for this purpose is presented and evaluated for the relative effectiveness of alternative substratification schemes. The framework is applied, using real MEPS data, to illustrate how utilizing paradata from the linked survey offers the possibility of making improvements to the sampling scheme for the subsequent survey. The improvements aim to reduce the data collection costs while maintaining or increasing effective responding sample sizes and response rates for a harder to reach population.

Key words: Sampling; response propensity; paradata; Medical Expenditure Panel Survey; National Health Interview Survey; interviewer observations.

1. Introduction

Costs of conducting surveys are increasing along with a growing reluctance among respondents to participate in surveys. Survey statisticians are exploring innovative ways to improve data collection efforts while minimizing costs through the use of paradata in adaptive/responsive design frameworks. In the 1940's, Hansen and Hurwitz first introduced concepts similar to adaptive/responsive design sampling schemes (Hansen and Hurwitz 1946). Groves and Heeringa (2006) defined responsive design and discussed the use of paradata to develop responsive designs to control survey costs, nonresponse, and improve the precision of the survey estimates. In recent years paradata are increasingly

[1] Office of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A. Email: LMirel@cdc.gov
[2] Center for Financing, Access, and Cost Trends, Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20857, U.S.A. Email: Sadeq.Chowdhury@ahrq.hhs.gov
Disclaimer and Acknowledgments: The views expressed in this article are those of the authors and no official endorsement by the Department of Health and Human Services, the Agency for Healthcare Research and Quality (AHRQ), or the Centers for Disease Control and Prevention (CDC). The authors would like to thank Steve Machlin, Sam Zuvekas, Fred Rohde, and Joel Cohen for their helpful feedback and comments on the topics discussed in this article. In addition, the authors would like to thank the editors of this special edition for their thoughtful and valuable comments. This research was conducted while Lisa B. Mirel was employed at AHRQ prior to becoming an employee at CDC.

being used for that purpose (Durrant et al. 2014; Durrant et al. 2015; Kreuter 2013; Wagner 2013; Groves et al. 2009).

Using paradata from a larger survey that is used for subsampling or linked to the sampling frame of a new survey, can also help develop tailored design sampling strategies to reduce data collection effort. One example of how this is being done is in the Medical Expenditure Panel Survey Household Component (MEPS-HC). MEPS-HC is the main component of MEPS and will be referred to as MEPS hereafter. MEPS, administered by the Agency for Healthcare Research and Quality, is a complex, multi-stage, nationally representative sample of the U.S. civilian, noninstitutionalized population. Each year a sample is drawn for MEPS as a subsample of responding households from the prior year's National Health Interview Survey (NHIS). The linkage or connection of these surveys offers a unique opportunity to use paradata from NHIS to inform sampling strategies in MEPS. One paradata variable, whether the NHIS interview was complete or partially complete, is associated with response propensity in MEPS and is currently being used for forming sampling subdomains or substrata in MEPS. Previous research has explored disproportionate sampling as a way to lower data collection costs (Barron et al. 2015). MEPS has implemented a similar strategy where the sample is drawn at different rates in different substrata based on response propensity as a way to reduce the data collection effort while increasing the unweighted and, potentially, the weighted response rates. The main objective of this article is to examine how paradata from a prior survey can be used in developing a sampling scheme in a subsequent survey. We illustrate how innovative methods can reduce the data collection costs without affecting the precision of the survey estimates.

The results from this research are applicable to other surveys, particularly those that use information from a larger survey to plan for a subsequent survey. For example, the 2010 National Survey of College Graduates (NSCG) selected a portion of its sample from the 2009 American Community Survey (ACS) respondents who indicated they had a bachelor's degree or higher in any field of study (National Science Foundation 2016). The ACS collects substantial amounts of paradata and, as noted in the National Academies Press book, "through its paradata, the ACS can also inform the subsequent survey process in ways that would improve the efficiency and quality of the data" (National Research Council 2008, 58). As an additional example, the American Time Use Survey (ATUS) uses paradata collected in the Current Population Survey (CPS) to aid in developing sampling strategies. The ATUS is sponsored by the Bureau of Labor Statistics and is conducted by the U.S. Census Bureau to measure how respondents spend their time. In the ATUS, individuals are randomly selected from a subset of households that completed their eighth and final month of interviews for the CPS (Bureau of Labor Statistics 2016). Similarly, these methods could aid in follow up surveys, such as those being proposed by the National Health and Nutrition Examination Survey (NHANES) Longitudinal Study (Centers for Disease Control and Prevention 2016). Paradata were used in an experiment to follow up sample units from the Survey of Consumer Sentiment to "predict the contact and co-operation propensities and at-home patterns of sample units in a new wave" (Luiten and Schouten 2013, 171). While there are drawbacks in subsampling (e.g., two phases of nonresponse and potential increases in design effects from unequal weighting), having the additional information from the prior survey could help target sampling strategies that would ease burdens on both the interviewers and respondents, and it has the potential to reduce costs of data collection.

This article focuses specifically on incorporating NHIS paradata variables into the MEPS sample design and presents an evaluation of the effectiveness of using two paradata variables complete/partial interview status and the interviewers' assessment of the likelihood of response in a linked survey. To do this, we present a method for optimal allocation of the MEPS sample, using NHIS paradata. We also present an approach to evaluate the relative performance of alternative stratification and allocation schemes in terms of data collection costs, impact on design effect, and the potential for increasing the response rate. Increasing the response rate at the first round of data collection in MEPS has the potential to help the overall response rates with multiple rounds of data collection. Our evaluation is based on real cost and response propensity data collected in earlier rounds of MEPS. The framework presented for sample allocation and evaluation of alternative strategies can be applicable to other surveys in similar situations.

2. Background

MEPS is a nationally representative sample of the U.S. civilian, noninstitutionalized population. It is an annual survey of about 14,000 households and has been conducted continuously since 1996. MEPS is a panel survey, and the annual sample consists of two overlapping panels (Figure 1). A new panel of sample is selected each year. It is followed up for two consecutive years with five rounds of data collection. MEPS is an in-person survey, and the results from the survey provide national estimates on health care use, expenditures, insurance coverage, sources of payment, access to care, and health care quality (Ezzati-Rice et al. 2008).

As mentioned above, the MEPS sample is drawn as a subsample of households that participated in the prior year's NHIS conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. The NHIS is a multi-purpose health survey that serves as the principal source of information on the health status and health behaviors of the civilian, noninstitutionalized U.S. population. NHIS is based on a complex, multi-stage sample design with oversampling of Hispanics, blacks and Asians (National Center for Health Statistics 2014). The NHIS complex sample design carries over to the MEPS through the set of NHIS responding households that comprise the frame for MEPS sample selection.

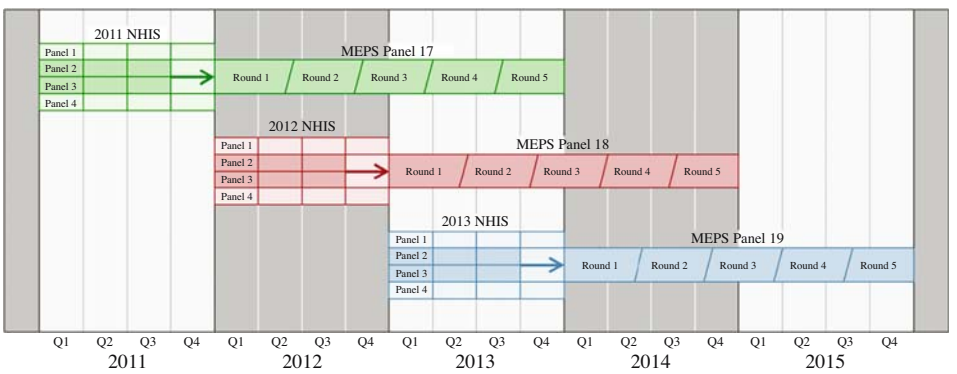

Fig. 1. NHIS-MEPS integrated overlapping panel sample design.

A disadvantage to this integrated design is that the response rates in MEPS are conditional on response rates in NHIS. More specifically, the overall response rate in MEPS is a compound response rate of NHIS and conditional MEPS response rates. For example, the MEPS Panel 19 response rate in Round 1 is 72% conditional on an NHIS response rate of 76%. Hence, the compound response rate for MEPS Round 1 is 55% (76% × 72%) which is much lower than the NHIS response rate of 76%. A detailed discussion of the calculation of MEPS response rates can be found in the MEPS public use file documentation (Agency for Healthcare Research and Quality 2016). The response rate is calculated independently for each round of data collection by considering the eligibility of a responding unit at that round. Multiplying the NHIS response rate with the product of the conditional response rates for each of the previous and current MEPS rounds produces the overall MEPS response rate up to that round.

Despite the disadvantages of the conditional response rates, there are many advantages to the integrated design. For example, the MEPS sampling frame from the NHIS contains a wealth of information collected in NHIS, including demographic and socioeconomic characteristics of responding members. The integration of the two surveys also means that MEPS does not need to screen households. This increases the efficiency of the design by eliminating the need to independently list and screen households and to locate policy-relevant subgroups of the population. Similarly, the very rich frame of auxiliary variables is used for nonresponse adjustments. The linkage of the two surveys also offers the opportunity to link MEPS data with NHIS data for longitudinal analysis. Another advantage is that the paradata collected in NHIS are used to help inform sampling strategies in the MEPS.

Our article examines the use of paradata from NHIS to inform sampling strategies in MEPS. In our study, the focus is on the non-certainty households and how sample allocation strategies within that domain can be improved, using paradata from the NHIS. The reason for this focus is as follows. Minority households are selected with certainty. Based on race and ethnicity information collected in NHIS, minorities are oversampled in MEPS to improve sample sizes for policy-relevant analyses. Minority households are defined as at least one or more people in the household who identify as Hispanic, Asian, or non-Hispanic black and are sampled with certainty for MEPS. Households that are not defined as minority households are classified as non-Hispanic white/other households. These households are the largest sampling domain in MEPS and are sampled at a non-certainty rate that balances the precision requirement of the estimates for this domain and the pre-assigned targeted sample size for allocation. Over the past several years, the overall sampling rate for the non-Hispanic white/other households has been about 61% of the households on the frame (see Table 1). Consequently, the MEPS sampling strategy can only be improved for non-certainty households that is, non-Hispanic white/other households.

Starting with Panel 16 of MEPS (2011), NHIS paradata have been used to further stratify the non-Hispanic white/other households and to help develop a tailored sampling strategy. A good predictor of response propensity is a paradata variable from NHIS that indicates if the NHIS interview was complete or partially complete. A complete interview means that the household composition, family, sample adult, and sample child (if a child was present) modules were all completed. A partial interview means that at least a

Table 1. Example of Sampling rates used in various sampling domains and subdomains starting with 2011 (Panel 16) of MEPS.

Domain	Sampling rate (%)
Hispanic	100
Asian	100
Non-Hispanic black	100
Non-Hispanic white/other	61*
NHIS Complete	63.2
NHIS Partial	49.2

*This number is a weighted average of the complete and partial sampling rates

sufficient portion of the family module was completed. Table 1 shows the sampling rates used in recent years for different domains and subdomains in MEPS.

Another paradata variable that may correlate with response propensity is the NHIS interviewer's assessment of cooperation. At the end of each NHIS interview the interviewer records an assessment of how likely she/he thinks the respondent would be to respond to a future linked survey. The interviewer can choose:

1. definitely agree to linked survey,
2. probably agree to a linked survey,
3. probably refuse a linked survey, or
4. definitely refuse a linked survey.

This NHIS variable is not currently used in the MEPS sampling but we discuss it here as a possible future enhancement to the MEPS sampling scheme.

Past research suggests that interviewer assessments can be useful for sampling and for assessing respondent burden. While interviewer assessments may not always be perfectly accurate because they are based on judgments by the interviewers which may add measurement error (West 2013; West and Kreuter 2013), they can still provide insight for sampling and estimating this burden. A recent study revealed that interviewer ratings about participation can "correlate with the cooperation rate" (Eckman et al. 2013, 1). One case study, described in Groves and Heeringa (2006) has illustrated the utility of these types of interviewer ratings: "Sample cases that interviewers expected to have low propensities achieved a second-phase response rate of 38.5%; the high propensity stratum, 73.7%" (442). Similarly in a study conducted using the Consumer Expenditure Survey, the researchers incorporated post survey questions about an interviewer's perception about a respondent's willingness to participate in the survey into their conceptual model predicting response burden in their longitudinal survey (Fricker et al. 2014). Utilizing a similar variable collected in NHIS, we examine possible improvements to sampling the non-certainty households in MEPS.

3. Methodology

We use paradata from NHIS and actual outcomes from previous MEPS fieldwork to create two alternative stratifications to form subdomains or substrata for sampling the non-certainty households. We then allocate the sample at different rates depending on the

relative cost and response propensity of a substratum. We allocate the same overall sample size for each of the stratification schemes to compare their cost-effectiveness. Since sampling at different rates in different substrata increases the variance, we try to optimize the allocation in a manner that balances cost, variance, and response rates. Given the integrated design of NHIS and MEPS, there is inherent variation of the base sampling weights for MEPS. All discussion of variability in the article reflects "additional" variation that the MEPS sampling scheme adds on top of the variation of the NHIS base weight.

We present an approach to optimally allocate the sample to minimize the data collection effort while maintaining the efficiency of the estimates. The proposed optimal allocation approach is used to allocate the sample to different substrata within a domain. We then evaluate and compare the cost effectiveness of different stratification options. Cost effectiveness is defined in the following sections.

3.1. Sample Allocation for a Cost-Effective Design

We allocate the sample to substrata by balancing data collection effort, response rate, and the variance of the estimates. The allocation of the sample is done in a two-step process. First, the sample is optimally allocated to minimize the data collection effort and then the sample size is adjusted to control the increase in variance due to the variation in sampling rates.

We use a cost function that incorporates a fixed cost and a variable cost of data collection in each substratum. The average number of contacts is used as a rough indicator for variable cost of data collection, ignoring any variation in unit cost of a contact by region or primary sampling unit. The number of contacts is affected by many factors, including, but not limited to, locating the study participants, willingness of respondents to participate in the survey, and break offs during the survey. Throughout this article, contacts include actual contacts, contact attempts and calls, but we will generally use the term contacts.

The cost function for a domain or a broad stratum can be considered as follows:

$$C = C_o + \sum C_h n_h \tag{1}$$

where C_o is the fixed cost and all other costs that are invariant to subsampling in substratum h, C_h is the average cost for completing each sampled unit in substratum h and n_h is the sample size in substratum h.

The average cost C_h in substratum h can be defined by factoring in the average number of contacts and response rate as follows:

$$C_h = Q_h/R_h = \text{Overall average number of contacts for achieving a response}, \tag{2}$$

with

$Q_h =$ average number of contacts for each selected household including both respondents and nonrespondents,

$R_h = \frac{n_{hr}}{n_h} =$ response rate, where n_{hr} is the number of respondents in substratum h.

Any other perceived or real cost component can be incorporated in deriving C_h or C. For example, any variation in the unit cost of a contact by geography or other factors can also be accounted for by computing an weighted average cost C_h.

In the absence of any attempt to reduce the number of contacts, no sampling substratum is formed and there is no need for any sample allocation. However, for a comparison at the stratum level with a stratified sampling scheme, the sample selected without stratification can be considered in expectation as proportionally allocated (i.e., the same sampling rate) in different substrata. Therefore, if NHIS paradata were not used for the subsampling of non-certainty households, then the sample in an overall draw is expected to be allocated proportionally in substrata as follows:

$$n_h = n * \frac{N_h}{\sum_h N_h} \tag{3}$$

where n is the overall sample size in the domain or the broad stratum, n_h is the expected allocated sample size in substratum h, and N_h is the frame size in substratum h.

To minimize the cost (in our example, number of contacts) for a fixed sample size n, an appropriate substratification can be formed and the sample can be allocated optimally (Neyman 1934) as follows:

$$n_h = n * \frac{N_h S_h/\sqrt{C_h}}{\sum N_h S_h/\sqrt{C_h}} = n * \frac{N_h S_h/\sqrt{Q_h/R_h}}{\sum N_h S_h/\sqrt{Q_h/R_h}} \tag{4}$$

where S_h is the standard deviation of a target variable in substratum h.

Since the interest here is to control the variance increase due to variation in weights for differential allocation or sampling rates, the variation of a target variable in different substrata within a broad stratum will be assumed the same that is, $S_h = S$. In that case, the above expression for optimal allocation will be reduced to:

$$n_h = n * \frac{N_h/\sqrt{C_h}}{\sum N_h/\sqrt{C_h}} = n * \frac{N_h/\sqrt{Q_h/R_h}}{\sum N_h/\sqrt{Q_h/R_h}} \tag{5}$$

Focusing on the objective of reducing cost, the allocation is set to sample more heavily within a substratum that has larger populations and lower costs (Lohr 2009). The above allocation will minimize costs for a fixed sample size n in a domain. However, as the sampling rate varies by substrata the variance in the domain will increase due to variation in weights. To control the variance, the stratum sample size should be adjusted by considering the higher design effect and increase in response rate.

As we deviate from the proportional allocation to the optimum allocation to minimize costs, the variation in base sampling weights (w) will increase the overall design effect ($deff_w$) due to variation in weights as follows (Kish 1965):

$$deff_w = \left(1 + CV_w^2\right) \tag{6}$$

where $CV_w = \frac{\sqrt{V(w)}}{\bar{w}}$ is the coefficient of variation of sampling weights across substrata with the variance of weight defined as,

$$V(w) = \sqrt{\frac{\sum_h n_h (w_h - \bar{w})^2}{n}}. \tag{7}$$

In our example, the variation of sampling weights is only for selection in MEPS. As noted previously, this is "additional" variation to the NHIS base weight for those participants who are selected into MEPS. For the proportional allocation, since the subsampling rate is the same in all substrata, the $CV_w = 0$ and hence $deff_w = 1$; the effective sample size will remain the same as n, where n is the realized sample size.

On the other hand, under the optimum allocation, the effective sample size will be reduced to $\frac{n}{deff_w}$.

We will consider this loss in the effective sample size when we consider the cost benefit analysis of the proportional and the optimum allocation.

Considering the increased design effect and increase in response rate, the stratum sample size n will be adjusted as follows:

$$n^* = n\frac{R\,deff_w}{R^*} \tag{8}$$

where n^* is the adjusted sample size, R is the stratum-level response rate with equal sampling rate in the stratum, R^* is the stratum-level response rate under the above allocation and $deff_w$ is the design effect for variation in sampling rate by substrata. The adjusted sample size n^* can now be used in (5) and reallocated to keep the variance fixed.

For appropriate stratification and optimal allocation, the ratio of increase in stratum-level unweighted response rate $\left(\frac{R}{R^*}\right)$ is usually higher than the increase in design effect ($deff_w$) that is, $\frac{R}{R^*} \leq deff_w$ implying $n^* \leq n$. An optimal allocation with appropriate stratification can reduce costs and also increase response rate while keeping the stratum-level variance the same or lower.

In the above adjustment of the stratum sample size, the design effect due to additional variation in the weights due to nonresponse adjustments is not considered. This is partly because the nonresponse adjustment is usually calculated by forming cells across strata or sampling domains with wide variation in base weights. A marginal increase in the variation of a stratum base weight has relatively small impact on the extra variation of the nonresponse adjusted weight within a cell or an estimation domain (Chowdhury and Baskin 2014). However, if necessary, a compensation for additional increase in the variation of the nonresponse adjusted weight can be done in two ways:

(a) by inflating $deff_w$ slightly in adjusting the stratum sample size in (8) above; and/or
(b) by reinvesting some of the cost savings into additional attempts during data collection to increase the response rate in the harder-to-reach substrata with lower sampling rates.

To illustrate the gain under the above procedure, let us consider the following example of a stratum of 10,000 households with 7,000 households in Substratum 1 and 3,000 in Substratum 2. Suppose that on average to obtain a response it requires seven contacts per household in Substratum 1 and ten contacts per household in Substratum 2. If we sample 1,000 households under the optimal allocation and adjust to keep the effective responding sample size fixed, the cost savings and the increase in the overall response rates can be seen in Figure 2 for different levels of response rate differences between Substratum 1 and Substratum 2. For a response rate of 75% in substratum 1 and 65% in Substratum 2, the cost savings compared to proportional allocation is about 2%, and the increase in the

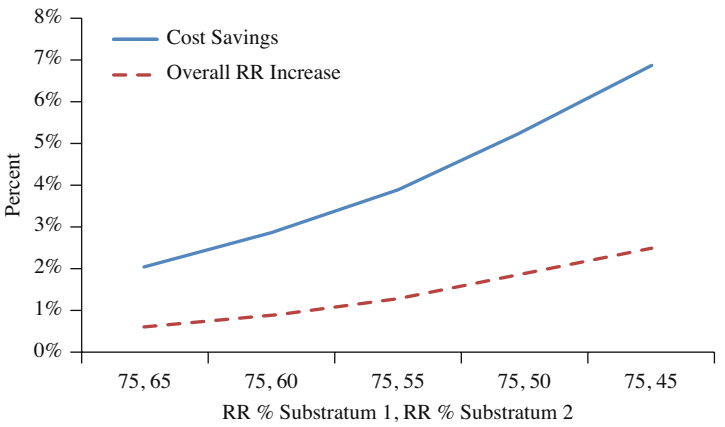

Fig. 2. Cost savings and increase in overall Response Rate (RR) by differences in response rates between substrata.

overall response rate is about 0.8%. For a response rate of 75% in Substratum 1 and 45% in Substratum 2, the increase in the overall response rate is about 2.5% and the cost saving is about 7%.

3.2. Evaluation of Alternative Stratification Schemes

We evaluate the overall savings in terms of the costs and the expected inflation in variance due to the increase in the CV of the weights. We also note the potential for an increase in response rate. To do this evaluation, we compare the cost-effectiveness of alternative schemes using the combined data from MEPS Panels 17 and 18 for the non-certainty sampling domain. We use this as a frame and select a hypothetical sample of size 4,750 households, roughly the usual sample size selected for non-certainty households in MEPS. The mean number of contacts and response rates observed in Panels 17 and 18 will be used for evaluation.

We evaluate two substratification alternatives which differ in how they separate the cases in the non-certainty domain into low and high response propensity. Scheme 1 uses the NHIS paradata variable which captures if the NHIS interview was complete or partially complete. Scheme 2 combines the interview status (complete/partial interview) and the perceived likelihood of response to a future linked survey (definitely respond, probably respond, probably refuse, likely refuse) as assessed by the NHIS interviewer. Since the stratification using the likelihood of response by itself does not offer significant gains compared to the interview completion status, this variable alone is not presented as a stratification scheme. Moreover, some of the categories of the likelihood of response are collapsed because the sample sizes are limited. The blended variable used in Scheme 2 includes three substrata that combine the two paradata variables as shown in Table 4.

The cost-effectiveness of each stratified scheme with optimal allocation is compared with the default scheme. In the default scheme no effort is made to reduce data collection costs, no substrata are formed and there is no varying of sampling rates across substrata. The default scheme without substratification is equivalent to proportional allocation or

equal sampling rate in all substrata as shown in Equation (3). We assumed proportional allocation or equal sampling rate in substrata for the default scheme just for comparison at the substratum level. It has no implication for the overall findings or conclusion of this article as we are not claiming that proportional allocation contributes to any variance improvement or cost reduction.

4. Results

4.1. Stratification Using Complete/Partial Interview Status

Table 2 shows the distribution of the non-certainty sampling domain for MEPS Panels 17 and 18 combined for Scheme 1 that is, by NHIS interview status (i.e., complete or partial) along with the corresponding response rates and mean number of contacts. Those with completed NHIS interviews have a higher response propensity and lower number of contacts on average. The unweighted MEPS response rate was 76.4% for those with a complete interview status based on NHIS compared to 58.5% for those with a partial interview status based on NHIS. The average number of contacts per response in MEPS is much lower (10.38) among NHIS completes compared to the average number of contacts (17.15) for NHIS partials.

Table 3 presents a comparison of sample allocation and the cost-benefit factors between the default scheme and the optimum allocation under Scheme 1. A sample size of 4,750 households was allocated under both schemes. Under the default scheme, the sampling rate is expected to be 61% from both the NHIS interview status of complete and partial substrata while under the stratified scheme with optimal allocation, the sampling rates are 63.2% from the complete substratum and 49.2% from the partial substratum. This difference in sampling rates is due to the higher cost in terms of the number of contacts in the partial substratum, which drives the sample allocation to be lower in the partial substratum and higher in the complete substratum.

As a result, the cost (number of contacts) is expected to decrease by 0.75% under the stratified sampling with optimal allocation. Similarly, since the response rate is lower in the partial substratum, the overall response rate is expected to be 74.2% (3,525 respondents) under the stratified sampling compared to 73.7% (3,500) under the non-stratified sampling. There is also a potential for increasing the weighted response rate due to higher concentration of effort to a smaller sample selected from the hard to reach households in the partial substratum. However, due to the increase in variation of weights

Table 2. Sample size, response rate and number of contacts in MEPS Panels 17 and 18 combined for substrata under Scheme 1.

NHIS Interview status (Scheme 1)	Number of households		Response rate (%)	Average contacts per household	Average contacts per complete
	Sampled	Responded			
Complete	6,599	5,042	76.4	7.93	10.38
Partial	1,183	692	58.5	10.03	17.15
Total	7,782	5,734	73.7	8.25	11.22

Table 3. Sample allocation and effectiveness analysis for Scheme 1 that is, NHIS complete/partial interview status.

	Default scheme: no stratification/ no variation in sampling rate				Stratification with optimum allocation/ variation in sampling rate			
NHIS Interview status (Scheme 1)	Sampled	Sampling rate (%)	Responded	Cost (contacts)	Sampled	Sampling rate (%)	Responded	Cost (contacts)
Complete	4,028	61.0	3,078	31,957	4,169	63.2	3,185	33,072
Partial	722	61.0	422	7,245	581	49.2	340	5,834
Total	4,750	61.0	3,500	39,201	4,750	61.0	3,525	38,906
Response rate			73.7%				74.2%	
CV of weights		0				9.0		
Effective sample size*			3,500				3,497	
Cost difference relative to default								− 295
Savings (%)							**+0.50%**	**−0.75%**
Difference in effective sample sizes							− 3	

*Effective sample size = (total number who responded)/(1 + CV^2) (e.g., 3,525/(1 + (0.09^2))

(9% CV) under the stratified sampling, the effective responding sample size will come down slightly from 3,500 to 3,497. On the other hand, since there is no additional variation in weights under the non-stratified sampling, the effective sample size will remain the same at 3,500. Under both designs the effective responding sample size is almost the same (3,500 and 3,497). No further adjustment is made to the overall sample size under the stratified sampling. Therefore, while the effective sample size remains almost the same under both schemes, the total number of contacts under the stratified sampling comes down, and the response rate goes up slightly. The number of contacts is used as a proxy for cost, decreasing the number of contacts means there will be a decrease in the costs.

4.2. *Stratification Using Complete/Partial Status and Likelihood of Response Status*

In this section we examine how stratification Scheme 1 can be made even more beneficial by utilizing an additional NHIS paradata variable, the likelihood of response in a subsequent linked survey as assessed by the interviewer in the NHIS, which we refer to as Scheme 2.

Table 4 shows response rates and average number of contacts for the two paradata variables and their cross classification. The last column of the table shows how the groups were collapsed to form three substrata to be used in Scheme 2. The groups were combined based on similarity of cost (number of contacts) and response rate.

Table 5 shows the response rates and cost (number of contacts) for Scheme 2 after the groups are collapsed into three substrata. The response rate ranges from 48.3% in Substratum 3 to 77.2% in Substratum 1 and the average number of contacts per complete ranges from 10.18 in Substratum 1 to 21.85 in Substratum 3. In comparison, the response rates for Scheme 1 are 58.5% for the Partials and 76.4% for the Completes.

Similar to Table 3, Table 6 presents a comparison of the default scheme and the optimal allocation under Scheme 2. The same sample size of 4,750 households was allocated under both scenarios. The sampling rates under the stratified design with optimal allocation are 63.7% in Substratum 1, 52.0% in Substratum 2 and 43.5% in Substratum 3. The mean number of contacts is negatively associated with the sampling rate; therefore, the cost

Table 4. Response rate and cost (number of contacts) by complete/partial interview status and likelihood of response in MEPS Panels 17 and 18 combined.

NHIS Interview status	Likelihood of response	Response rate (%)	Average contacts per household	Average contacts per complete	Scheme 2 substrata
Complete	Definitely agree	79.8	7.66	9.60	1
	Probably agree	73.9	8.13	11.01	1
	Probably refuse	61.7	9.14	14.81	2
	Definitely refuse	55.0	9.40	17.09	3
Partial	Definitely agree	70.3	8.23	11.71	1
	Probably agree	64.3	10.03	15.60	2
	Probably refuse	48.7	10.88	22.32	3
	Definitely refuse	42.9	9.53	22.24	3
Total		73.68			

Table 5. Sample size, response rate, and number of contacts in MEPS Panels 17 and 18 combined for substrata under Scheme 2.

Scheme 2 substrata	Number of households		Response rate (%)	Average contacts per household	Average contacts per complete
	Sampled	Responded			
1	6,358	4,910	77.2	7.86	10.18
2	909	575	63.3	9.67	15.29
3	515	249	48.3	10.57	21.85
Total	7,782	5,734	73.7	8.25	11.20

(number of contacts) is expected to decrease by 1% under the stratified design. Similarly, the overall response rate is expected to be 74.5% (3,537 respondents) under the stratified design. On the one hand, due to the 11% CV of weights arising from the optimal allocation, the effective responding sample size will only come down slightly to 3,495. On the other hand, since there is no additional variation in weights – as the sampling rate is the same in both sampling substrata under the default scheme – the effective sample size remains the same at 3,500. Since the effective responding sample size under the stratified schemes is very close to that of the default scheme, no further adjustment to the overall sample size is made under the stratified scheme.

Table 7 summarizes the findings for Schemes 1 and 2 compared to the default scheme with no stratification or variation in sampling rate. While the impacts of both stratification methods are similar in terms of a negligible decrease in effective sample sizes, the blended stratification appears slightly better in terms of response rate (higher) and overall costs (lower) for obtaining a response.

5. Discussion

In this article, we discuss an approach to improve sampling strategies in MEPS. The approach is based on optimally allocating the sample to substrata formed using paradata from the linked NHIS. We present a method for optimal allocation of the sample to different substrata to minimize the data collection costs for a fixed variance and present an evaluation approach to select the best alternative stratification.

During the last few years, the NHIS interview status (complete/partial) has been used to form the sampling substrata (Scheme 1) to both reduce costs and potentially increase the response rate. Here we explore using an additional paradata variable, likelihood of response in a subsequent survey as assessed by the NHIS interviewers and blend it with the complete/partial interview status variable to form more effective substrata (Scheme 2). The sampling substrata formed by grouping the households with similar response propensities and number of contacts helps to develop a sampling strategy that reduces the data collection effort for a harder to reach population. An evaluation comparing Scheme 1 and Scheme 2 shows that there are slight savings under both methods in terms of reducing costs (number of contacts) and increasing the response rate but Scheme 2 performs slightly better than Scheme 1.

The results illustrate how a tailored sampling scheme with optimum allocation in substrata formed using paradata can help reduce the cost and potentially increase the

Table 6. Sample allocation and effectiveness analysis for Scheme 2, that is, blending of NHIS complete/partial interview status and likelihood of response.

	Default scheme: no stratification/ no variation in sampling rate				Stratification with optimum allocation/ variation in sampling rate			
Scheme 2 Substrata	Sampled	Sampling rate (%)	Responded	Cost (contacts)	Sampled	Sampling rate (%)	Responded	Cost (contacts)
1	3,881	61.0	2,997	30,513	4,053	63.7	3,130	31,868
2	555	61.0	351	5,367	473	52.0	299	4,574
3	314	61.0	152	3,321	224	43.5	108	2,368
Total	4,750	61.0	3,500	39,201	4,750	61.0	3,537	38,809
Response rate			73.7%				74.5%	
CV of weights		0				11.0		
Effective sample size*			3,500				3,495	
Cost difference relative to default								− 392
Savings							**+0.79%**	**−1.00%**
Difference in effective sample sizes							− 5	

*Effective sample size = (total number who responded)/(1 + CV^2) (e.g., 3,537/(1 + (0.11^2))

Table 7. Comparison of two alternative optimum allocation sampling procedures compared to the default scheme.

Sampling substrata	Increase in response rate (%)	Reduction in number of contacts' (%)	Difference in effective sample size
Scheme 1: Complete/Partial	**0.50**	**−0.75**	**−3**
Scheme 2: Blended complete/partial and likelihood of response	**0.79**	**−1.00**	**−5**

response rate without reducing the efficiency of the estimates. Our results indicate that, in a typical MEPS panel, the existing and new stratifications (Scheme 1 and Scheme 2) with optimum allocation of sample can reduce the cost by about 0.75% and 1%, respectively. Similarly, the response rate can increase by about 0.50% under the current scheme and 0.79% under the proposed new cross classification scheme. Although the cost savings by using the existing or the new sampling scheme are marginal, they are achieved with virtually no loss in terms of effective sample size. The method could continue to be improved by applying differential sampling rates using additional available paradata or by collecting and utilizing more relevant paradata to offer further gains in cost reduction. The framework presented for allocating the sample and evaluating cost effectiveness of alternative stratification will be useful in other similar surveys.

It should be noted that with the optimal sampling scheme, it is possible that there could be additional variability in the nonresponse adjusted weight because of variation in substrata sampling rates within a stratum. Our analysis did not address this additional variability, in part, because previous research with MEPS data has shown that the differential sampling within a stratum or a domain has little effect on the overall variability of the weights because the nonresponse adjustment to the weights are made by combining cases from all domains with wide variation in base weights (Chowdhury and Baskin 2014). However, if necessary, any likely increase in the variability of the nonresponse adjusted weight can be mitigated by reinvesting some of the cost savings either by increasing the stratum sample size as discussed in the Methodology section or by making additional contact attempts during data collection to increase the response rate in a harder to reach substrata with lower sampling rates.

Although Scheme 2 has not actually been implemented in the field, MEPS plans to utilize additional paradata variables to tailor the sampling in future surveys. The integration between NHIS and MEPS offers a unique opportunity to continually improve sampling strategies in MEPS.

6. References

Agency for Healthcare Research and Quality. 2016. "MEPS HC-171 2014 Full Year Consolidated Data File." Available at: https://meps.ahrq.gov/data_stats/download_data/pufs/h171/h171doc.pdf. (accessed March 2017).

Barron, M., M. Davern, R. Montgomery, X. Tao, K.M. Wolter, W. Zeng, C. Dorell, and C. Black. 2015. "Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations." *Journal of Official Statistics* 31: 545–557. Doi: http://dx.doi.org/10.1515/JOS-2015-0034.

Bureau of Labor Statistics. (2016, June). *ATUS User's Guide (PDF)*. Available at: http://www.bls.gov/tus/atususersguide.pdf. (accessed March 2017).

Centers for Disease Control and Prevention. (2016, May 23). *Federal Register: The Daily Journal of the United States Government*. Available at: https://www.federalregister.gov/articles/2016/05/23/2016-12008/proposed-data-collection-submitted-for-public-comment-and-recommendations. (accessed March 2017).

Chowdhury, S.R. and R.M. Baskin. 2014. "PPS Subsampling from NHIS to MEPS – Effect on Precision of MEPS Estimates." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 2014, 2339–2351. Alexandria, VA: American Statistical Association (CD-ROM).

Durrant, G.B., O. Maslovskaya, and P.W.F. Smith. 2014. "Sequence Analysis as a Tool for Investigating Call Record Data." Working paper, University of Southampton. Available at: https://eprints.soton.ac.uk/369102/ (accessed March 2017).

Durrant, G.B., O. Maslovskaya, and P.W. Smith. 2015. "Modelling Final Outcome and Length of Call to Improve Efficiency in Call Scheduling." *Journal of Survey Statistics and Methodology* 3: 397–424. Doi: https://doi.org/10.1093/jssam/smv008.

Eckman, S., J. Sinibaldi, and A. Montmann-Hertz. 2013. "Can Interviewers Effectively Rate the Likelihood of Cases to Cooperate?" *Public Opinion Quarterly* 77: 561–573. Doi: http://dx.doi.org/10.1093/poq/nft012.

Ezzati-Rice, T.M., F. Rohde, and J. Greenblatt. 2008. *Sample Design of the Medical Expenditure Panel Survey Household Component, 1998–2007*, Methodology Report No. 22. March 2008. Rockville, MD: Agency for Healthcare Research and Quality. Available at: https://meps.ahrq.gov/data_files/publications/mr22/mr22.pdf (accessed March 2017).

Fricker, S., T. Yan, and S. Tsai. 2014. "Response Burden: What Predicts it and Who is Burdened Out?" In Proceedings of AAPOR Section: American Statistical Association, August 2014. 4568–4577. Alexandria, VA: American Statistical Association. (CD-ROM).

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society, Series A* 169: 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Groves, R.M., M.D. Mosher, J. Lepkowski, and N.G. Kirgis. 2009. *Planning and Development of the Continuous National Survey of Family Growth. National Center for Health Statistics*. Vital Health Stat, 1(48). Available at: https://www.cdc.gov/nchs/data/series/sr_01/sr01_048.PDF (accessed March 2007).

Hansen, M.H. and W.N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association* 41: 517–529. Doi: http://dx.doi.org/10.1080/01621459.1946.10501894.

Kreuter, F. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by Frauke Kreuter. Hoboken, NJ: John Wiley & Sons, Inc.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Lohr, S.L. 2009. *Sampling: Design and Analysis*. Boston: Richard Stratton.

Luiten, A. and B. Schouten. 2013. "Tailored Fieldwork Design to Increase Representative Household Survey Response: an Experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society A* 176: 169–189. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01080.x.

National Center for Health Statistics, *National Health Interview Survey, 2014. Public-use data file and documentation*. Available at: http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm (accessed March 2017).

National Research Council, N. 2008. *Using the American Community Survey for the National Science Foundation's Science and Enginering Workforce Statistics Programs*. Washington, DC: National Academies Press. Doi: https://doi.org/10.17226/12244.

National Science Foundation. 2016. *National Survey of College Graduates*. Available at: http://www.nsf.gov/statistics/srvygrads/#tabs-1 (accessed March 2017).

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97: 558–606. Doi: http://dx.doi.org/10.2307/2342192.

Wagner, J. 2013. "Using Paradata-Driven Models to Improve Contact Rates in Telephone and Face-to-Face Surveys." In *Improving Surveys with Paradata: Analytic Use of Process Information*, edited by F. Kreuter, 145–170. New Jersey: John Wiley and Sons.

West, B.T. 2013. "An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth." *Journal of the Royal Statistical Society A* 176: 211–225. Doi: http://dx.doi.org/10.2307/23355184.

West, B.T. and F. Kreuter. 2013. "Factors Affecting the Accuracy of Interviewer Observations Evidence from the National Survey of Family Growth." *Public Opinion Quarterly* 77: 522–548. Doi: https://doi.org/10.1093/poq/nft016.

Received January 2016
Revised March 2017
Accepted April 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 385–408, http://dx.doi.org/10.1515/JOS-2017-0019

Web-Face-to-Face Mixed-Mode Design in a Longitudinal Survey: Effects on Participation Rates, Sample Composition, and Costs

Annamaria Bianchi[1], *Silvia Biffignandi*[1], *and Peter Lynn*[2]

Sequential mixed-mode designs are increasingly considered as an alternative to interviewer-administered data collection, allowing researchers to take advantage of the benefits of each mode. We assess the effects of the introduction of a sequential web-face-to-face mixed-mode design over three waves of a longitudinal survey in which members were previously interviewed face-to-face. Findings are reported from a large-scale randomised experiment carried out on the UK Household Longitudinal Study. No differences are found between the mixed-mode design and face-to-face design in terms of cumulative response rates and only minimal differences in terms of sample composition. On the other hand, potential cost savings are evident.

Key words: Attrition; total survey error; nonresponse bias; randomised experiment.

1. Introduction

Combining different modes within a survey has long been thought to provide opportunities to benefit from the strength of each mode (de Leeuw 2005). Biemer and Lyberg (2003) assert that in United States and Western Europe mixing modes is the norm for surveys at present. Since the development of web surveys, mixed-mode data collection methods with a web component are increasingly considered as an efficient possibility by many organisations. Indeed, the inclusion of web into a mixed-mode design has potentials to reduce costs, increase timeliness, and improve quality/sample composition (Groves and Lyberg 2010; Couper 2011; Kreuter 2013).

The opportunities for mixed-mode data collection with web are particularly appealing for longitudinal surveys. Indeed, some of the constraints on implementing mixed-mode surveys are reduced in the longitudinal setting, thanks to the diversity of information that can be collected from sample members at the recruitment/first wave. First, collection of

[1] Department of Management, Economics and Quantitative Methods, University of Bergamo, via dei Caniana 2, 24127 Bergamo, Italy. Emails: annamaria.bianchi@unibg.it, silvia.biffignandi@unibg.it
[2] Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex C04 3SQ, UK. Email: plynn@essex.ac.uk

Acknowledgments: The first two authors would like to acknowledge support by the COST Action IS1004 and by the ex 60% University of Bergamo, Biffignandi grant. The contribution of the third author forms part of the methodological research programme of Understanding Society: the UK Household Longitudinal Study. Understanding Society is funded by the (UK) Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. The authors thank Dr. Annette Jäckle for providing Stata codes.

contact information for sample members permit gains and cost savings to be made by approaching panel members in the most cost-efficient mode. For example, email addresses can be collected at the first wave to facilitate subsequent invitations to complete web surveys. Second, knowledge about which sample members are more or less likely to respond in which mode allows targeting of particular mode strategies at specific subgroups, in the framework of adaptive survey design (Lynn 2014; Calinescu and Schouten 2015; Bianchi and Biffignandi 2014). Finally, the study of the effects of different mode strategies can take advantage of the wide range of information available for each sample member from previous waves, thus providing a rather unique opportunity to identify detailed characteristics of respondents in different modes.

Some other considerations in the introduction of mixed-mode designs are also specific to the longitudinal context. First, high response rates are essential to allow longitudinal analyses (Lynn forthcoming). This is because nonresponding sample members cannot be replaced by new sample members. Thus, response rates and cumulative response rates are more important in the longitudinal framework than in cross-sectional surveys. Second, in an ongoing panel that has previously been interviewer-administered, sample members have prior experience of the interview in another mode and prior knowledge of the survey content. These prior experiences might increase the chances of response in web mode, even in the absence of an interviewer (Jäckle et al. 2015), as the task of introducing the survey and the respondent task is greatly reduced.

The aim of this article is to study the effect of a mixed-mode design including web on several aspects related to data quality in a longitudinal survey. By 'mixed-mode' we refer specifically to a sequential mixed-mode design, where web is offered first, followed by face-to-face follow-up of nonrespondents to the web phase. We compare this mixed-mode design to a simple face-to-face design. In both designs we allow the possible use of different modes in a final 'mop-up' step to boost response (e.g., Computer Assisted Telephone Interviewing (CATI)) as we believe this represents good practice and does not fundamentally affect the nature of the designs. Details of the specific designs upon which our analyses are based are presented below. To the best of our knowledge, this is the first study of the effects of introducing a mixed-mode design including web over multiple waves of a longitudinal survey.

Several issues may arise when using web and mixed-modes for data collection. Participation rates are usually low for web surveys (Fan and Yan 2010). Cooperation may be harder to maintain in the absence of personal interviewer contact. This may particularly be the case when the mixed-mode design uses a lower response rate mode first in a sequential design (Lynn 2013). However, the effect on response rates of including web in a mixed-mode design is not completely clear. Several studies have found a lower response rate with a sequential mixed-mode design including web than with the equivalent design without web (Griffin et al. 2001; Janssen 2006; Lagerstrøm 2008; Leesti 2010; Martin and Lynn 2011; Souren 2012), while others have found that adding web to an otherwise single-mode design does not affect response rate (Fong and Williams 2011; Klausch et al. 2015a).

Jäckle et al. (2015) report on the effects at one wave only with reference to the same experiment we analyse. They found that individual response rates were lower with the mixed-mode design and no subgroup could be identified where the reverse was true. They also found that the mixed-mode design resulted in a lower proportion of households in

which all individuals responded. Gaia (2014) found no significant difference in attrition rates after three waves between the two designs.

The possibility of differential measurement error is a very important concern when considering converting a single-mode interviewer-administered survey to a mixed-mode survey including web. Several studies have identified systematic differences in measurement between modes (Bowling 2005) and in some contexts this has been shown to result in measurement differences between face-to-face single-mode and web-face-to-face mixed-mode data collection (Jäckle 2016; Klausch et al. 2015b). However, effects on measurement are not the focus of this article.

In the longitudinal context, response behavior may be affected by the time sample members have been in the panel and by previous wave outcome. It is well known that wave-on-wave attrition rates in longitudinal studies are highest at the second wave and then decline over time (Lugtig 2014; Schoeni et al. 2013; Uhrig 2008). There is also evidence that the correlates of nonresponse may change over waves of a survey (Farrant and O'Muircheartaigh 1991). Further, a study based on four waves of the UK Household Longitudinal Study found that changes in correlates of nonresponse at each subsequent wave are lower compared to the previous one (Bianchi and Biffignandi 2017). Also, those who have been longer in the panel have more experience of the interview in another mode and prior knowledge of the survey content than those who have entered the panel more recently. These aspects might increase the chance of a successful transition to web interviewing.

It is thus expected that more recent panel members will show higher levels of attrition/nonresponse. Jäckle et al. (2015) found that for longer panel members (original sample) the proportion of interviews of any form was lower with mixed-mode, while there was no difference by mode treatment for more recent panel entrants (refreshment sample).

Previous wave nonrespondents are known to have lower response propensities in subsequent waves (Watson and Wooden 2014; Jäckle et al. 2015). We thus expect higher attrition rates among previous wave nonrespondents, which could result in greater sensitivity to mode treatment amongst this group. Furthermore, an invitation to complete the interview by web offers the opportunity to at least make contact with some sample members who are very hard to contact face-to-face (due to being rarely at home at the times when interviewers visit). Jäckle et al. (2015) found that amongst previous wave respondents the mixed-mode design resulted in a higher proportion of refusals than face-to-face design and amongst previous wave nonrespondents it resulted in a smaller proportion of proxy interviews. Moreover, Jäckle et al. (2015) found several groups to be less likely to give an interview in the mixed-mode treatment than face-to-face: men, white, in rural location, web users, those for whom an email address was available, age 21–30, in a household with children, and individuals who said they would definitely not do the survey by web. If these patterns persist over waves, then they are expected to lead to biases in the estimates of correlated variables. Persistent patterns could guide the implementation of targeted mode assignment.

Thus, our first research question is:

RQ1: Does the mixed-mode design affect participation rates (cumulatively or at each wave separately), either overall or amongst important subgroups, compared to the primarily face-to-face design?

Furthermore, it is possible that subgroup differences in response propensity could differ between modes (Groves and Peytcheva 2008; Voogt and Saris 2005). Heterogeneity across modes in response propensities could result in smaller compositional biases with mixed-mode designs than with single-mode designs. Empirical knowledge on these aspects is rather limited, especially in the context of longitudinal surveys. Voorpostel and Ryser (2011) in the implementation of a web-face-to-face concurrent mixed-mode design for refusal conversion in an otherwise CATI panel survey (the Swiss Household Panel) found that the group that completed the web questionnaire tended to have characteristics that were slightly different from the CATI group. They argue that, if larger numbers had been reached, this would have diminished the bias in demographic characteristics. No significant differences in sample composition between a sequential mixed-mode design and single-mode face-to-face were found by Lynn (2013), with respect to a CATI-face-to-face design in the UK, or by Klausch et al. (2015a), with respect to CATI-face-to-face, web-face-to-face or mail-face-to-face. The relevance of sample composition measures depend on the substantive analytical objectives of data users. In case of multi-purpose surveys with many users and many equally-important estimates, it is essential that the response set presents no compositional biases with respect to many variables. Our second research question is therefore:

RQ2: Does the mixed-mode design affect sample composition, compared to the primarily face-to-face design? Does any such effect change over waves as attrition cumulates?

Since one of the main reasons for the implementation of mixed-mode designs with a web component is related to cost reduction, we investigate some aspects related to survey costs. First, in the context of household panels where all household members need to be interviewed, a significant cost-saving may be obtained only when all household members respond by web, as this avoids the need for an interviewer to visit the household in the face-to-face follow-up phase. In this respect and with reference to one wave only, Jäckle et al. (2015) found that one in five households fully responded online, suggesting the potential for useful cost savings. We extend the results in Jäckle et al. (2015) by investigating the extent to which households fully respond online over three waves in order to ascertain whether cost savings may increase over time following the introduction of a mixed-mode design. Further, we explicitly evaluate the relative mean field cost per issued household for the mixed-mode design and the primarily face-to-face design and for each wave. In this respect our analysis goes beyond that in Jäckle et al. (2015). So our third research question is:

RQ3: To what extent does the mixed-mode design reduce field work costs over waves, compared to the primarily face-to-face design?

We analyse data from the Understanding Society Innovation Panel. The Innovation Panel is a longitudinal panel designed explicitly to enable methodological research. The size of the panel is large, which provides good statistical power. The survey aims to interview each adult member of the household. At Wave 5, a randomised experiment was carried out, to inform decisions on whether and how the main Understanding Society Survey (Buck and McFall 2012) might move from a single-mode face-to-face survey to a

mixed-mode survey that includes web interviewing. Two-thirds of sample units were allocated at random to the mixed-mode treatment (sequential mixed-mode in which web was followed by face-to-face), with the other one-third receiving the face-to-face treatment. At the time of the experiment, the panel consisted of 1,573 households and 3,040 adults eligible for interview. The experiment continued at Waves 6 and 7, so that respondents received the same treatment they were assigned to at Wave 5. This structure of the experiment enables investigation of long term effects of mode treatments on panel attrition. Minor changes to the design were applied at Waves 6 and 7, with reference to incentive levels and follow-up procedures. Particularly, at the end of the fieldwork a final 'mop-up' phase was included, which introduced CATI and web options in the face-to-face treatment and CATI in the mixed-mode treatment. Thus, in Waves 6 and 7 the face-to-face treatment was not strictly single-mode. However, as modes used in the 'mop-up' stage played a very small part in overall response (see Subsection 2.1 below), we will use the term 'primarily face-to-face' for the face-to-face treatment in Waves 6 and 7.

Positive effects of incentives on response rates have been found for web surveys (Göritz 2006, 2010, 2015). Incentives have found to be effective also in longitudinal surveys (Laurie and Lynn 2009; Jäckle and Lynn 2008). Thus, respondent incentives were provided in both treatment groups, though the level and nature of the incentives differed between the groups, reflecting the reality that sample members might require additional motivation in the absence of an interviewer. Each of the two mode treatments therefore represents a realistic overall design, though it must be taken into account that the unit cost of incentives is slightly higher in the mixed-mode treatment. Details of the incentive strategies are set out in Section 2 below and a cost comparison is presented in Section 6.

In a Total Survey Error (TSE) perspective (Biemer 2010; Groves and Lyberg 2010; Lynn and Lugtig 2017), this article represents a step towards the optimisation of surveys by maximising certain aspects of survey quality within a budgetary constraint. For example, if cost savings are found by the introduction of mixed-mode with a web component, a larger sample could be afforded for the same budget, which in turns leads to lower variance of the estimates.

In the next section of the article, we describe the data and the experimental study. Next, we present results on participation (Section 3), sample composition (Section 4), and costs (Section 5). Sections 6 and 7 conclude.

2. Data

We use data from the Understanding Society Innovation Panel (Uhrig 2011). More precisely, we consider data from a randomised experiment carried out at Wave 5 and continued at Waves 6 and 7. Subsection 2.1 describes the main characteristics of the panel, Subsection 2.2 provides details on the experimental design.

2.1. The Understanding Society Innovation Panel

The Understanding Society Innovation Panel is an ongoing longitudinal survey which has collected data in annual waves since 2008 (Lynn and Jäckle, forthcoming). The target population for the Innovation Panel is all individuals aged 16 or over and living in

England, Scotland, or Wales. The sample had two components: those who were invited to take part at each wave since Wave 1 and those who entered the survey at Wave 4. We refer to these two sample components as the original sample and the refreshment sample, respectively. Another refreshment sample was added in Wave 7, but is excluded from our analyses.

Both samples are stratified, clustered, probability samples of persons. Primary sampling units are postal sectors, secondary sampling units are residential addresses selected from the Postcode Address File (Lynn and Lievesley 1991) and sample elements are persons. The sample of persons is therefore initially clustered within households (though that clustering reduces over waves of the panel). Further details on the Innovation Panel sample design can be found in Lynn (2009).

The Understanding Society Innovation Panel involves interviews at twelve-month intervals with the initial sample members and all members of the current household of each sample person. Household response at any wave can thus be complete if all household members answer the survey or partial, if only some of the household members participate. Only sample members who were in participating households at the first wave for that sample were re-approached for interview at subsequent waves. Sample members were followed to their new location if they moved anywhere within Great Britain. From Wave 2 onwards, nonresponse at one wave did not preclude an interview attempt at the next wave. Households in which no person responded at two successive waves are no longer issued to the field. Thus, in the sample issued to the field at Wave 5 – which forms the base for most of our analyses – the original sample included all individuals who were in households that had responded at either Wave 3 or Wave 4 and the refreshment sample only included individuals in households that had responded at Wave 4. Thus, at Wave 5 it is only the original sample that includes previous wave nonrespondents.

Interviews cover a wide range of topics, such as household dynamics, economic activity, income, health, housing, and political attitudes. The survey is a multi-purpose survey intended as a major research resource, with thousands of users from different disciplines and a very diverse range of analytical objectives (Buck and McFall 2012).

Proxy interviews are allowed on behalf of individuals who cannot be interviewed in person, but only after considerable efforts have been made to obtain a personal interview. The decision to allow a proxy interview is made subjectively on a case-by-case basis by field staff. At Waves 5, 6, and 7 – the field outcomes which are the subject of our analyses – the proportion of interviews completed by proxy was 6.9%, 5.9%, and 3.2%, respectively. As for modes used in data collection, at Waves 1, 3, and 4, all interviews were carried out face-to-face. Experimentation with a mixture of face-to-face and CATI was carried out at Wave 2 in 2009 (Lynn et al. 2010). The main conclusion from that experiment was that a CATI-face-to-face sequential mixed-modes design, if implemented in a way that would save costs, was likely to result in lower response rates (Lynn 2013). For that reason, CATI was not included as an initial mode at Waves 5 to 7.

2.2. *Experimental Design*

At Wave 5, all sample members were randomly allocated to one of two treatment groups. The allocation was at the household level, so all individuals in the same household

received the same treatment. Interviewers are assigned to households based on geographic location, a factor that had no influence on the allocation to treatment, so each interviewer assignment included households in both treatment groups. One third of the sample was allocated to the primarily face-to-face treatment and two thirds were allocated to the web-face-to-face sequential mixed-mode design. The experiment was continued (with the same treatment allocation) at Waves 6 and 7. The distribution of the issued sample of households across samples and mode treatments is summarised in Table 1.

In Wave 5, the face-to-face treatment involved standard Understanding Society procedures. Each adult sample member (aged 16 or over) was sent an advance letter with a prepaid unconditional incentive, after which interviewers visited to attempt face-to-face interviews. In each household, one person was asked to complete the household enumeration grid and the household questionnaire. All household members aged 16 or over were asked for an individual interview, including a self-completion component administered by computer-assisted self-interviewing (CASI).

In the mixed-mode treatment group, sample members aged 16 or over were sent a letter with a prepaid unconditional incentive, inviting them to take part by web. The letter included the URL and a unique user ID, which was to be entered on the welcome screen. A version of the letter was additionally sent by email to all sample members for whom an email address was available (around half of the sample: of the emails sent, 10% bounced, 30% were opened by the recipient and 60% were left unopened). For people who had indicated at previous waves that they do not use the internet regularly for personal use, the letter mentioned that they would also have the opportunity to do the survey with an interviewer. Up to two email reminders were sent at three-day intervals. Sample members who had not completed the web interview after two weeks were sent a reminder by post and interviewers then started visiting them to carry out face-to-face interviews. The interviewer visits began in the same week that the reminder letter would have been received. Face-to-face interviewers thereby had their full allocation at the start of their fieldwork, rather than having nonresponding web individuals being passed to them during the fieldwork period. The web survey remained open throughout the fieldwork period.

The first household member to log on to do the web survey was asked to complete the household grid, which collects information on who is currently living in the household. The web grid included an additional question to identify who is responsible for paying bills. The household questionnaire could be completed by either this person or their spouse/partner. For these sample members the household questionnaire was displayed first, then leading on to the individual questionnaire. Once one partner had completed the household questionnaire, it would not appear for the other partner. The web questionnaire was based on the face-to-face one, with some adaptations, for example incorporating interviewer instructions into question wording, removing references to showcards, and making 'help' screens more respondent-appropriate. There were no differences in questionnaire content, question order or routing. The web survey was not suitable for completion using a small mobile device. If a mobile device was used to access the log-on page, the respondent was automatically directed to a page requesting that they log on from a computer.

The same procedures were carried out in Waves 6 and 7, with a few small differences. First, respondents accessing the survey from a mobile device were no longer blocked from completing it, though they were still presented with a warning message suggesting that it

Table 1. Allocation of households to experimental groups in Wave 5 and distributions in Waves 6 and 7.

		Wave 5		Wave 6		Wave 7	
Sample component	Previous wave outcome	F2F	MM	F2F	MM	F2F	MM
Original sample	Responded	321	615	292	544	277	544
	Did not respond	43	111	41	89	21	37
Refreshment sample	Responded	168	315	148	263	141	250
	Did not respond	-	-	17	29	12	15
Total	-	532	1,041	498	925	451	846

Notes: Households ineligible at the respective wave (e.g., deceased, moved abroad) have been excluded; F2F = face-to-face; MM = mixed-modes; Responded means that at least the household grid and either the household interview or an individual interview was completed at the previous wave.

would be easier to complete the survey on a PC or laptop. In the mixed-mode treatment group, the proportion of individual web interviews completed on a mobile device was 7% at Wave 6 and 18% at Wave 7. Second, the proportion of sample members in the mixed-mode treatment who had supplied a valid email address and could therefore be sent a survey invitation by email increased at each wave, being around 60% at Wave 6 and 65% at Wave 7. Third, "nonresponse mop-up" procedures to obtain participation of individuals who had not participated by the end of the fieldwork period were extended to include additional modes. This included nonresponding individuals in partially responding households. Nonrespondents in the face-to-face group were sent a letter offering the opportunity to participate by web. The letter included the URL of the web instrument and a unique log-on code. For those whose email addresses were available, this invitation was also sent by email. A few days later, an interviewer attempted contact by telephone with all those for whom a phone number was known in order to remind them of the web questionnaire, and to administer a CATI interview if possible. Telephone contacts were also attempted with all remaining nonrespondents in the mixed-mode group. The telephone interviewer reminded the sample member that they could participate on the web, but was also able to administer the interview by CATI. Cases for which a telephone number was not known were not contacted again at this stage. CATI was included in this final stage at Waves 6 and 7 on the grounds that an additional contact mode might increase the chances of contact being made with some of the most difficult to contact sample members. At Wave 6, just five individual interviews (0.7% of all interviews) in the face-to-face treatment group were completed by CATI and fifteen (2.2%) by web. In the mixed-mode group, fourteen interviews (1.0%) were completed by CATI. At Wave 7, just one individual interview (0.1% of all interviews) in the face-to-face treatment group was completed by CATI and 25 (3.2%) by web. In the mixed-mode group, three interviews (0.2%) were completed by CATI. It is clear that these additional modes had only a minor impact on response outcomes.

At each wave all sample members received an unconditional incentive, enclosed with the advance letter. The value of the incentive was manipulated as part of a separate experiment. Allocation was at the household level, so all individuals in the same household received the same incentive. At Wave 5, in both mode treatment groups original sample members received either GBP 5 or GBP 10, while refreshment sample members received GBP 10, GBP 20, or GBP 30. Additionally, a conditional incentive experiment was carried out within the mixed-mode group (fully crossed with the unconditional incentive experiment) to test ways of increasing web participation. Half of the households were offered an additional incentive of GBP 5 per person conditional on all eligible household members completing the web survey within two weeks. This was mentioned in the advance letters to all household members in this treatment group. Detailed analyses of the impact of incentives at Wave 5 are presented in Bianchi and Biffignandi (forthcoming).

At Wave 6, the incentive experiment was restricted to the mixed-mode part of the sample. Individuals were allocated in equal proportions to three treatments: GBP 10 unconditional incentive, GBP 30 unconditional, or GBP 10 unconditional incentive with an additional GBP 20 per individual conditional on all adult household members taking part online within the two-week web-only period. For the primarily face-to-face part of the sample, all sample members were provided a GBP 10 incentive.

At Wave 7, all continuing sample (original and Wave 4 refreshment) members were again administered the same incentive as at Wave 6.

The analyses carried out in Sections 3 and 4 are on households and individuals aged 16 or over. For households, analyses are restricted to households issued to the field at the respective wave, excluding ineligible households at that wave. For Wave 7, households from the Wave 7 refreshment sample are also excluded. Sample sizes are 1,573 for Wave 5, 1,423 for Wave 6, and 1,297 for Wave 7. As for individuals, we restricted to individuals issued to the field at Wave 5 and eligible at Waves 5, 6, and 7 – counting individuals not issued to later waves as (eligible) nonrespondents (any household that did not respond at either wave $w - 1$ or w would not be issued at $w + 1$). For those individuals issued to Wave 5 and not issued to later waves, nonresponse is classified using last wave available nonresponse classification. The sample size is 2,756. For individuals, we use variables from the most recent available interview as covariates. The cost analysis in Section 5 is based on all households issued to field.

As mentioned above, Jäckle et al. (2015) perform similar analyses as ours, but using only Wave 5 data. With respect to the samples used in Jäckle et al. (2015), we consider the same sample for households at Wave 5. The sample of individuals is not the same as we consider individuals eligible over Waves 5, 6, and 7 (not 5 only). As a consequence, results for households at Wave 5 (first three columns of Table 5) are consistent with those in Jäckle et al. (2015), while results for individuals at Wave 5 are not exactly the same.

3. Participation

The first aspect that we consider is the impact of mixed-mode data collection on participation (RQ1). Notice that all our analyses are conditional on being issued to the field at Wave 5, which means that all Wave 1 nonresponding households and some who adamantly refused or were persistent nonrespondents at Waves 2 to 4, have been dropped from the sample. Our focus is on the effect of mode treatment on attrition at Waves 5, 6, and 7, the waves at which the randomised experiment was carried out. In Subsection 3.1, we consider individual participation, while in Subsection 3.2 we investigate household participation, as interest lies also in how any differences in individual participation cluster within households.

3.1. Individual Participation

A particularly important outcome in the context of longitudinal studies is the cumulative response rate over waves, as this is related to the possibility of performing longitudinal analyses. For analyses of change, observations need to be available from each wave of interest and different patterns of missingness across waves may lead to a large number of cases being dropped from the analyses.

Table 2 compares mixed-mode data collection with primarily face-to-face data collection in terms of the number of waves (out of three) at which the sample member provides a full interview, as well as full interview response rate in each wave separately. No significant differences are found between treatments for the cumulative response rate over three waves ($P = 0.45$). Looking at response in each wave separately, the effect of

Table 2. Individual response rates (in %) – F2F = face-to-face; MM = mixed-modes; P-values from Pearson χ^2 tests, corrected for the survey design (strata and clusters).

Response	F2F	MM	*P*
Waves 5–7 response			
3 full interviews	47.3	49.1	0.45
2 or 1 full interviews	32.9	31.3	0.57
0 full interviews	19.8	19.6	0.92
Wave 5 full interview	71.0	68.4	0.30
Wave 6 full interview	69.3	70.7	0.52
Wave 7 full interview	56.1	59.1	0.21
N	940	1,816	

mixed-modes on the proportion of full interviews went from −2.6 percentage points at Wave 5 to +3.0 at Wave 7, though none of these differences are statistically significant.

Turning to individual response by subgroups of interest (Table 3), no difference between the mixed-mode design and primarily face-to-face design was observed with respect to the cumulative response rate, in the original sample ($P = 0.86$), the refreshment sample ($P = 0.30$), the original sample Wave 4 respondents ($P = 0.81$), or the original sample Wave 4 nonrespondents ($P = 0.11$). Amongst Wave 4 nonrespondents in the original sample, the mixed-mode design resulted in a lower proportion of no interview over three waves than face-to-face (54.9% vs 66.5%, $P = 0.09$). Separate analyses for each wave show that the proportion of full interviews did not differ significantly between treatments for either the original sample ($P = 0.16$) or the refreshment sample ($P = 0.67$) in Wave 5. In Waves 6 and 7, amongst Wave 4 nonrespondents in the original sample, the mixed-mode design resulted in a higher proportion of full interviews than face-to-face design (32.9% vs 20.0%, $P = 0.06$ in Wave 6 and 28.0% vs. 18.7%, $P = 0.08$ in Wave 7). In Wave 7, the proportion of full interviews is higher for the mixed-mode group for both the original and the refreshment samples, even though the differences did not reach statistical significance.

To investigate whether the mixed-mode design had different effects on attrition for different subgroup characteristics, we fitted a logit model predicting full response over three waves (versus proxy or nonresponse in any one of the three waves) using individual characteristics and interactions of those characteristics with treatment as predictors. Individual characteristics were measured in Wave 4 (or last available interview before Wave 5). Results for the original responding sample are summarised in Table 4, which shows the estimated coefficients from the model, together with *p*-values of t-tests for significance (adjusted for sample design). At the five percent level, the only significant interaction is between mode and web preference, with respondents who said at Wave 4 that they would definitely/maybe respond to a web survey having higher probabilities to respond in the mixed-mode group. The effect is stronger for those who declared they would definitely respond to a web survey.

Table 3. Individual response rates (in %) by subsample – F2F = face-to-face; MM = mixed-modes; P-values from Pearson χ^2 tests, corrected for the survey design (strata and cluster).

	Original Sample									Refreshment Sample		
	Total			Wave 4 responding			Wave 4 nonresponding			Total		
	F2F	MM	P	F2F	MM	P	F2F	MM	*P*	F2F	MM	*P*
Waves 5–7 response												
3 full interviews	46.5	47.0	0.86	61.0	61.9	0.81	8.4	13.4	0.11	49.0	54.0	0.30
2 or 1 full interviews	31.4	30.8	0.84	32.2	29.4	0.41	25.2	31.7	0.25	35.8	32.5	0.46
0 full interviews	22.1	22.2	0.97	6.8	8.7	0.39	66.5	54.9	**0.09**	15.2	13.5	0.61
Wave 5 full interview	68.6	64.4	0.16	85.2	80.3	0.12	20.6	26.3	0.21	75.8	77.6	0.67
Wave 6 full interview	67.3	68.6	0.65	84.1	83.5	0.79	20.0	32.9	**0.06**	73.2	75.5	0.56
Wave 7 full interview	55.2	57.9	0.37	68.7	71.3	0.46	18.7	28.0	**0.08**	57.7	61.9	0.39
N	630	1,268		454	858		155	350		310	548	

Table 4. Logistic regression results for giving full interview in Waves 5, 6, and 7 – Original sample (Wave 4 respondents only), N = 1,296 – based on a logit model including the allocated mode, characteristics of the sample members, and interactions between the mode and characteristics as predictors.

Variable	Category	Coefficient	Std. Error	*P*-value
Intercept	-	−1.01	0.65	0.13
Mixed-Mode (MM)	Mixed-mode group	0.33	0.82	0.69
(Ref. Face-to-face goup)				
Gender	Male	−0.05	0.18	0.78
(Ref. Female)				
Race	White	0.76	0.48	0.12
(Ref. Nonwhite)				
Working Status	In work	0.03	0.28	0.93
(Ref. Not in work)				
Urbanicity	Urban	0.41	0.29	0.16
(Ref. Rural)				
Webuser	Yes	0.08	0.33	0.80
(Ref. No)				
Email given	Yes	0.52	0.28	**0.06**
(Ref. No)				
Age	16–20	−0.79	0.53	0.14
(Ref. 41–50)	21–30	−0.63	0.49	0.21
	31–40	0.31	0.34	0.36
	51–60	0.38	0.36	0.29
	61–70	1.52	0.45	**0.00**
	71+	0.20	0.49	0.68
Household type	Single	0.19	0.34	0.58
(Ref. Couple)	Single, children	−0.23	0.57	0.69
	Couple, children	0.22	0.42	0.60
	2+ unrelated adults	−0.08	0.44	0.86
	2+ unrelated adults, children	0.14	0.42	0.75
Web preference	Maybe	−0.42	0.33	0.20
(Ref. No)	Yes	−0.57	0.38	0.14
MM#Gender	MM#Male	−0.09	0.23	0.68
MM#Race	MM#White	−0.27	0.55	0.62
MM#Working Condition	MM#In work	0.20	0.35	0.56
MM#Urbanicity	MM#Urban	−0.04	0.35	0.91
MM#Webuser	MM#Yes	−0.22	0.42	0.61
MM#Email given	MM#Yes	−0.08	0.33	0.82
MM#Age	MM#16-20	0.65	0.62	0.30
	MM#21-30	0.22	0.63	0.73
	MM#31-40	−0.53	0.46	0.26
	MM#51-60	0.14	0.47	0.77
	MM#61-70	−0.71	0.53	0.18
	MM#71+	0.30	0.52	0.56
MM#Household type	MM#Single	−0.20	0.44	0.66
	MM#Single, children	−0.44	0.67	0.51
	MM#Couple, children	−0.71	0.52	0.18

Table 4. Continued.

Variable	Category	Coefficient	Std. Error	*P*-value
	MM#2+ unrelated adults	−0.62	0.53	0.25
	MM#2+ unrelated adults, children	−0.86	0.54	0.12
MM#Web preference	MM#Maybe	0.84	0.42	**0.05**
	MM#Yes	1.09	0.45	**0.02**

The Nagelkerke R^2 is 0.119.

To answer the first research question on participation rates (RQ1), the mixed-mode design does not affect individual participation either overall or amongst those who have been in the panel for longer or shorter periods. The mixed-mode design appears to have a positive effect for those who had not responded at Wave 4, though statistical significance is borderline. As for other subgroups, which had been identified to be less likely to give an interview at Wave 5 in Jäckle et al. (2015), only expressed preference to respond by web showed to have a positive effect on participation in the mixed-mode group with respect to the primarily face-to-face group. No other difference between mode treatments was found.

3.2. *Household Participation*

For households, we analyse outcomes for each wave separately, since a concept of longitudinal household does not make sense as household composition and location may change over time.

The proportion of households participating in the original sample (Table 5) did not differ significantly between treatments in Wave 5 ($P = 0.22$) or Wave 6 ($P = 0.79$), while the mixed-mode design resulted in a 6.5 percentage point higher participation rate than face-to-face in Wave 7 ($P = 0.03$). As for the proportion of complete households, in the original sample it is 7.1 percentage points lower ($P = 0.03$) with the mixed-mode design than with face-to-face only in Wave 5, and by Wave 7 it becomes 10.5 points higher ($P = 0.00$). Non-contacts and refusals in the mixed-mode group are higher than in the face-to-face group in Wave 5 ($P = 0.08$), not significantly different in Wave 6 ($P = 0.33$ and $P = 0.89$. respectively), and lower in Wave 7 ($P = 0.07$ and $P = 0.06$, respectively).

These effects differ between previous wave respondents and nonrespondents. Amongst previous wave responding households in the original sample, the proportion of refusals with the mixed-mode treatment compared with face-to-face was higher at Wave 5 (12.4% vs. 6.9%, $P = 0.03$), not different at Wave 6 ($P = 0.60$), and lower at Wave 7 (7.2% vs. 11.2%, $P = 0.03$). No statistically significant differences are observed for previous wave nonrespondents in the original sample.

For those who have entered the panel more recently (refreshment sample), no statistically significant difference between the mode treatment groups was observed in any wave with respect to household participation and complete household participation (results not shown).

Table 5. Household response rates for the original sample.

	Wave 5			Wave 6			Wave 7		
	F2F	MM	*P*	F2F	MM	*P*	F2F	MM	*P*
HH response rate (complete+partial)	78.0	74.2	0.22	84.4	83.7	0.79	74.4	80.9	**0.03**
Complete HHs	58.2	51.1	**0.03**	63.4	65.2	0.59	51.3	61.8	**0.00**
Partial HHs	19.8	23.1	0.20	21.0	18.5	0.32	23.1	19.1	0.10
Non-contact	8.0	8.4	**0.08**	5.7	6.5	0.33	11.1	7.9	**0.07**
Refusal	11.3	15.8	**0.08**	8.1	8.4	0.89	13.1	9.1	**0.06**
Other nonresponse	2.7	1.5	0.24	1.8	1.4	0.63	1.3	2.1	0.49
N	364	726		333	633		298	581	

Notes: F2F = face-to-face; MM = mixed-mode; HH = household; P-values from Pearson χ^2 tests, corrected for the survey design (strata and clusters); Partial means that at least one individual failed to complete the individual interview.

4. Sample Composition

In this section, we explore whether the two different mode treatments had different effects on sample composition (RQ2). More precisely, we investigate whether there is a mode difference in whether sample composition at each wave and, especially, in the sample that responded at all three waves, differs from the composition at the start of the experiment. We test this assumption by comparing the distribution of covariates collected at Wave 4 (or last wave interview before Wave 5) for different subgroups of respondents. The statistical test for differences in sample composition with respect to a variable is performed by fitting a logistic regression model predicting response in which predictors are mode treatment, the variable under consideration, and the interaction between mode and the variable. The Wald test on the interaction coefficients is a test of whether the association between the outcome and the variable differs by mode. We consider different groups of respondents: individuals responding at Wave 5, individuals responding at Waves 5 and 6, and individuals responding at Waves 5, 6, and 7.

The variables that we considered are those where we expect the greatest chance of a mode difference, on the basis of results in previous studies. More precisely, we consider variables found to be related to response behaviour at Wave 5 in Jäckle et al. (2015, Tables 8 and 9). All these variables are related to at least some substantive variable of interest. For example, ethnicity is an important predictor in studies on social inequalities (Wallace et al. 2016; Chng et al. 2016), while urbanicity figures prominently in research on commuting effects (Evandrou et al. 2016). Therefore any effect of nonresponse on sample composition in respect of these variables has the potential to introduce bias in substantive estimates of interest to researchers.

Results are shown in Table 6. For respondents at Wave 5, only household type shows a significant difference between the mixed-mode and face-to-face group ($P = 0.04$). As for respondents at Waves 5 and 6 and respondents at all three waves, the only variable showing a mode difference in how sample composition differs from the composition at the start of the experiment is expressed web preference ($P = 0.06$ and $P = 0.08$, respectively). The proportion of respondents who said at Wave 4 that they definitely would not respond to a web survey decreases by 2.6 percentage points in the mixed-mode group for respondents at all three waves with respect to the initial sample, while it increases by 1.6 percentage points in the face-to-face group.

Overall, and to answer the second research question (RQ2), we conclude that differences between the two treatment groups in sample composition are minimal.

5. Costs

As stated earlier, one of the main reasons for the use of web in a mixed-mode design is to save costs. In this section we provide an indication of the scale of the data collection cost differential between the two mode treatments (RQ3). The estimate can only be indicative as the realised cost saving from a mixed-mode design such as this one in any particular situation will depend on several aspects of the survey context, survey design, and contractual and remuneration arrangements. Furthermore, focusing solely on data collection costs ignores the possibility that a survey agency might incur initial setup costs

Table 6. Marginal distributions of each treatment group for the sample issued at Wave 5, respondents at Wave 5, respondents at Waves 5 and 6, and respondents at Waves 5, 6, and 7.

		Wave 5 issued (1)		Wave 5 responding (2)			Waves 5 and 6 responding (3)			3 waves responding (4)			Difference (4)-(1)	
Variable	Categories	F2F	MM	F2F	MM	P	F2F	MM	P	F2F	MM	P	F2F	MM
Gender	Male	46.7	46.8	44.9	45.4	0.73	44.1	44.7	0.61	45.0	44.3	0.56	−1.7	−2.5
Ethnic group	White	93.6	91.9	93.0	91.8	0.32	94.5	93.2	0.98	96.0	94.3	0.65	+2.4	+2.4
Working condition	In work	57.4	54.3	59.5	55.3	0.48	57.7	55.1	0.79	55.8	55.8	0.12	−1.6	+1.5
Urbanicity	Urban	76.4	75.3	76.3	75.1	0.92	76.2	75.0	0.96	78.7	74.2	0.14	+2.3	−1.1
Web user	Web user	67.2	69.0	69.2	70.9	0.93	69.0	70.0	0.44	68.6	71.3	0.61	+1.4	+2.3
HH internet	Living in HH internet	84.6	83.2	86.3	84.9	0.97	86.5	83.9	0.36	86.3	84.5	0.83	+1.7	+1.3
Age	16–20	7.6	8.4	5.9	7.1	0.51	5.2	6.7	0.27	4.6	5.1	0.47	−3.0	−3.3
	21–30	10.3	12.4	8.7	10.2		8.0	7.9		6.1	7.0		−4.2	−5.4
	31–40	14.7	13.3	15.4	13.5		15.6	13.0		16.1	12.8		+1.4	−0.5
	41–50	19.1	19.5	19.8	20.1		19.4	19.8		17.5	20.6		−1.6	+1.1
	51–60	19.1	19.5	20.3	18.3		19.6	18.9		19.8	19.7		+0.7	+2.9
	61–70	16.7	14.3	19.6	15.7		20.8	17.5		24.1	18.8		+7.4	+4.5
	71+	12.5	15.3	10.3	15.2		11.3	16.2		11.8	16.1		−0.7	+0.8
HH type	Single	10.3	13.9	13.4	15.2	**0.04**	13.5	16.9	0.35	13.0	17.7	0.85	+2.7	+3.8
	Single, children	5.5	4.6	4.8	4.4		5.3	4.8		4.9	4.0		−0.6	−0.6
	Couple	31.1	29.1	30.9	30.8		32.0	33.4		35.5	34.5		+4.4	+5.4
	Couple, children	22.8	21.7	24.0	21.0		22.9	19.7		23.4	20.3		+0.6	−1.4
	2+ unrelated	18.6	18.8	16.0	17.7		15.6	16.6		14.8	16.6		−3.8	−2.2
	2+ unrelated, children	11.7	11.9	11.1	10.9		10.8	8.7		8.3	6.8		−3.4	−5.1
Web preference	No	32.0	31.1	32.5	29.6	0.24	33.8	29.7	**0.06**	33.6	28.5	**0.08**	+1.6	−2.6
	Maybe	37.1	40.0	37.4	40.8		36.3	41.1		37.5	41.2		+0.4	+1.2
	Yes	30.9	28.9	30.1	29.5		29.9	29.2		29.0	30.3		−1.1	+0.8

Notes: – F2F = face-to-face; MM = mixed-mode; HH = household; *P*-values from Wald tests, corrected for the survey design (strata and clusters).

in introducing a mixed-mode system, and that the cost of some office-based tasks may be greater for a mixed-mode survey. Despite these limitations, the analysis presented here may give a useful impression of the scale of cost-savings with a mixed-mode design.

The main driver of the difference in data collection costs between the two mode treatments is the fact that some sample households do not require an interviewer visit in the mixed-mode treatment. The proportion of households fully responding by web can therefore be used as an initial indicator of potential cost savings, as a full response by web negates the need to send an interviewer to visit the household. The proportion of fully responding households who fully responded by web increased over time, from 42.7% in Wave 5 to 57.5% in Wave 7 (Table 7). This increase over time is apparent for both the original sample (previous wave respondents) and the refreshment sample (results not shown), though at every wave the proportion of households fully responding by web is higher in the refreshment sample than in the original sample. For example, at Wave 7 the proportion of fully-responding households who fully responded by web was 56% in the original sample, compared to 72% in the refreshment sample. It is noteworthy that in Waves 6 and 7 more than one-third of all households fully responded by web (37.1% and 35.1%, respectively).

If field costs per issued sample household – excluding the cost of incentives – were assumed to be approximately GBP 110 with the primarily face-to-face treatment, and GBP 5 per household for the web phase of the mixed-mode treatment, this would imply that costs in the mixed-mode design would be around GBP 5 for each household that fully responds by web and GBP 115 for each other household. Applying these unit costs to the response outcomes in Table 7 would imply that the mixed-mode design could bring about reductions in the cost per household issued to the field of around 19% at Wave 5, 33% at Wave 6 and 31% at Wave 7 (Table 8, rows 2 and 5). However, these figures do not include the costs of incentives which, for Waves 6 and 7, were higher in the mixed-mode treatment group. Rows 1 and 4 of Table 8 show the mean cost of incentives per issued household in each mode for each wave, taking into account the proportion of households in the mixed-mode sample that qualified for the conditional incentives, as well as all unconditional incentives. Incorporating these into the overall data collection costs (rows 3 and 6), the cost differential between mode treatments reduces, with the result that the mixed-mode design is now estimated to bring cost savings of around 15% at Wave 5, 8% at Wave 6 and

Table 7. Proportion of households fully responding by web and proportion of households fully responding at waves 5, 6, and 7.

	Wave 5	Wave 6	Wave 7
Mixed-mode sample			
% fully responding by web (A)	23.8	37.1	35.1
% fully responding (B)	55.7	66.7	61.0
(A)/(B)	42.7	55.6	57.5
N	*1,041*	*925*	*846*
Face-to-face sample			
% fully responding	58.8	62.8	52.8
N	*532*	*498*	*451*

Table 8. Mean field cost per issued household for each treatment group and for each wave.

	Wave 5 (GBP)	Wave 6 (GBP)	Wave 7 (GBP)
Mixed-mode treatment			
Incentives	29.74	49.35	41.91
Other	88.82	74.19	76.39
Total	*118.56*	*123.54*	*118.30*
Face-to-face treatment			
Incentives	29.70	24.58	22.77
Other	110.00	110.00	110.00
Total	*139.70*	*134.58*	*132.77*

11% at Wave 7. It should be noted, however, that these estimated cost savings may have limited generalisability as realised savings will depend on factors such as the cluster sample size, the geographical dispersion of sample addresses within the cluster, and whether interviewers are remunerated equally for interviewing a web-nonrespondent household as they would have been for interviewing a household in a simple face-to-face survey.

6. Conclusions

Regarding possible effects of the mixed-mode design on response rates, either overall or amongst important subgroups (RQ1), for individual participation no difference between mode treatments was detected overall (both cumulative response rate and response rate in each wave). Also, no differences were found in either the original sample or the refreshment sample as a whole, while the mixed-mode design performed slightly better amongst previous wave nonrespondents in the original sample. As for covariates, only expressed mode preference has been found to be related to participation in the mixed-mode group rather than in the primarily face-to-face group. These are very useful results with respect to the implementation of a mixed-mode design with web in a longitudinal survey. They suggest that such a design should not damage participation rates over several waves and may even improve participation amongst sample members who are otherwise less likely to participate. The finding regarding expressed mode preference suggests that answers to a question such as this could usefully be taken into account as part of a strategy for targeted allocation of sample members to mode treatments (Lynn 2014).

As for household participation, no differences could be found in Wave 5 and Wave 6 overall, but the mixed-mode design showed a better performance than face-to-face in Wave 7: higher household participation, higher complete household interviews, and lower non-contact rates. For those who had entered the panel more recently (refreshment sample), no difference in household participation could be detected in any of the three waves. For those who had been longer in the panel (original sample), the mixed-mode design resulted in smaller proportion of households fully responding and higher proportion of non-contacts and refusals in Wave 5; in Wave 7, the situation was completely reversed.

With respect to possible effects of the mixed-mode design on sample composition (RQ2), differences between the two treatment groups in sample composition are minimal. The data provide little evidence of mode treatment affecting sample composition.

With regard to possible cost savings related to the use of the web in the mixed-mode design (RQ3), the mixed-mode design appears to have potential to deliver substantial cost savings. At both Waves 6 and 7, more than one-third of households issued to the field fully responded by web. Our estimates suggest possible field cost savings per issued household in the region of ten percent, compared to face-to-face. The extent to which this saving would be realised in practice depends on, amongst other things, whether the amount of field effort required per household amongst the two-thirds of mixed-mode households that need to be issued to a face-to-face interviewer differs from that amongst the face-to-face sample. Analysis of call record data (results not shown) suggests that in this study the mean number of interviewer visits to a sample household was actually lower in the mixed-mode group (amongst households issued to a face-to-face interviewer) than in the primarily face-to-face group. This suggests that the indicated cost savings could well be fully realised.

7. Discussion

The introduction of web-face-to-face sequential mixed-mode data collection as a cost-saving alternative to single-mode face-to-face has been considered by many surveys but has generally been treated with caution due to concerns about possible negative impacts on nonresponse and measurement. This article has not considered measurement issues, but with regard to nonresponse we suggest that the concerns seem largely unwarranted, at least in the context of an ongoing panel survey. We have found no differences between the mixed-mode and primarily face-to-face designs in cumulative response rates over three waves of the panel, nor were significant differences found in the composition of the responding sample. Meanwhile, the potential for worthwhile field cost savings is evidenced by the sizeable proportion of sample households in which all adult members completed the questionnaire by web. This study therefore paints a rather positive picture of the potential for mixed-mode data collection in panel surveys.

However, some unresolved issues remain. Not least amongst them is the question of whether, and in what circumstances, measurement can be considered to be equivalent between the modes. The considerable literature on mode effects suggests that certain question characteristics tend to be associated with measurement differences between modes, particularly between self-completion and interviewer-administered modes (Couper 2011; de Leeuw 2005; Krosnick and Alwin 1987; Schwarz et al. 1991). For any particular survey considering the introduction of a mixed-mode design, the questionnaire content could be reviewed in the context of this literature, while effects on nonresponse error could be considered in the context of the findings of the current study, thus contributing to an overall evaluation of total survey error.

That said, it would be reasonable to question whether our findings would apply in different survey contexts (different topics of questioning, different study populations, different levels of prior survey engagement, etc.). Sensitivity to context is of course possible. However, we can draw some strength from that fact that our findings were broadly similar for the two different samples involved and for several demographic subgroups. The former suggests that our broad conclusions apply equally to sample members with only one previous wave and to those with four previous interviewer-administered waves, and

therefore that the degree of prior survey engagement does not have a strong influence on the outcomes studied. The latter suggests that the results might equally apply to study populations with rather different demographic profiles. Taken together, these findings provide some indication that our conclusions are at least somewhat robust.

8. References

Bianchi, A. and S. Biffignandi. 2014. "Responsive Design for Economic Data in Mixed-Mode Panels." In *Contribution to Sampling Statistics*, edited by F. Mecatti, P.L. Conti, and M.G. Ranalli, 85–102. Springer International Publishing.

Bianchi, A. and S. Biffignandi. 2017. "Representativeness in Panel Surveys." To appear in *Mathematical Population Studies*.

Bianchi, A. and S. Biffignandi. Forthcoming. "Survey Experiments on Interactions: a Case Study of Incentives and Modes." In *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*, edited by P.J. Lavrakas, E. de Leeuw, A. Holbrook, C. Kennedy, M.W. Traugott, and B.T. West. Hoboken NJ: John Wiley & Sons.

Biemer, P.P. 2010. "Total Survey Error: Design, Implementation and Evaluation." *Public Opinion Quarterly* 74(5): 817–848. Doi: http://dx.doi.org/10.1093/poq/nfq058.

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. New York: John Wiley.

Bowling, A. 2005. "Mode of Questionnaire Administration Can Have Serious Effects on Data Quality." *Journal of Public Health* 27: 281–291.

Buck, N. and S. McFall. 2012. "Understanding Society: Design Overview." *Longitudinal and Life Course Studies* 3: 5–17.

Calinescu, M. and B. Schouten. 2015. "Adaptive Survey Designs to Minimize Survey Mode Effects – a Case Study on the Dutch Labor Force Survey." *Survey Methodology* 41(2) : 403–425.

Chng, S., M. White, C. Abraham, and S. Skippon. 2016. "Commuting and Wellbeing in London: the Roles of Commute Mode and Local Public Transport Connectivity." *Preventive Medicine* 88: 182–188. Doi: http://dx.doi.org/10.1016/j.ypmed.2016.04.014.

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

Evandrou, M., J. Falkingham, Z. Feng, and A. Vlachantoni. 2016. "Ethnic Inequalities in Limiting Health and Self-Reported Health in Later Life Revisited." *Journal of Epidemiology & Community Health* 70: 653–662. Doi: http://dx.doi.org/10.1136/jech-2015-206074.

Fan, W. and Z. Yan. 2010. "Factors Affecting Response Rates of the Web Surveys: a Systematic Review." *Computers in Human Behavior* 26: 132–139.

Farrant, G. and C. O'Muircheartaigh. 1991. "Components of Nonresponse Bias in the British Election Surveys." In *Understanding Political Change*, edited by A. Heath, J. Curtice, R. Jowell, S. Evans, J. Field, and S. Witherspoon, 235–249. London: Pergamon Press.

Fong, B. and J. Williams. 2011. "British Crime Survey: Feasibility of Boosting Police Force Area (PFA) Sample Sizes Using Supplementary Recontact Surveys." Report for the Home Office, TNS-BMRB, London.

Gaia, A. 2014. "Does a Mixed-Mode Design Increase Panel Attrition? Evidence from the UKHLS Innovation Panel." Paper presented at the Internet Survey Methodology Workshop, Bolzano, December 1–3.

Göritz, A. 2006. "Incentives in Web Studies: Methodological Issues and a Review." *International Journal of Internet Science* 1: 58–70.

Göritz, A. 2010. "Using Lotteries, Loyalty Points, and Other Incentives to Increase Participant Response and Completion." In *Advanced methods for conducting online behavioural research*, edited by S. Gosling and J. Johnson, 219–233. Washington DC: American Psychological Association. Doi: http://dx.doi.org/10.1037/12076-014.

Göritz, A. 2015. "Incentive Effects." In *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis, 339–350. London: Routledge.

Griffin, D., D. Fischer, and M. Morgan. 2001. "Testing an Internet Response Option for the American Community Survey." Paper presented at the annual conference of the American Association for Public Opinion Research Montreal." Quebec, Canada, May 17–20.

Groves, R. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5): 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065.

Groves, R.M. and F. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72: 167–189. Doi: https://doi.org/10.1093/poq/nfn011.

Jäckle, A. 2016. "Mode Effects on Measurement in *Understanding Society*." Paper presented at the International Panel Survey Methods Workshop, Berlin, June 20–21. Available at: http://www.diw.de/en/diw_01.c.534396.en/program_psmw2016.html (accessed March 2017).

Jäckle, A. and P. Lynn. 2008. "Respondent Incentives in a Multi-Mode Panel Survey: Cumulative Effects on Nonresponse and Bias." *Survey Methodology* 34: 105–117.

Jäckle, A., P. Lynn, and J. Burton. 2015. "Going Online with a Face-to-Face Household Panel: Effects of a Mixed Mode Design on Item and Unit Non-Response." *Survey Research Methods* 9(1): 57–70. Doi: http://dx.doi.org/10.18148/srm/2015.v9i1.5475.

Janssen, B. 2006. "Web Data Collection in a Mixed Mode Approach: An Experiment." Paper presented at the European Conference on Quality in Official Statistics (Q2006), Cardiff, April 24–26. Available at: webarchive.nationalarchives.gov.uk/20140721132900/http://ons.gov.uk/about/newsroom/events/q2006—european-conference-on-quality-in-survey-statistics-24-26-april-2006/agenda/session-19-wednesday.pdf (accessed March 2017).

Klausch, T., 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(4): 945–961. Doi: http://dx.doi.org/10.1111/rssa.12102.

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

Kreuter, F. 2013. "Facing the Nonresponse Challenge." *The Annals of the American Academy of Political and Social Science* 645: 23–35. Doi: https://doi.org/10.1177/0002716212456815.

Krosnick, J.A. and D.F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement." *Public Opinion Quarterly* 51: 201–219. Doi: https://doi.org/10.1086/269029.

Lagerstrøm, B. 2008. "Cost Efficiency in a Mixed-Mode Survey – Evidence from the Norwegian Rent Market Survey." Paper presented at the 19th International Workshop on Household Survey Nonresponse, Ljubljana, September 15. Available at: http://www.nonresponse.org/db/3/558/Biblliography/Cost%20efficiency%20in%20a%20mixed-mode%20survey%20Evidence%20from%20the%20Norwegian%20Rent%20Marked%20Survey/?&p1=308&p2=74&p3=551 (accessed March 2017).

Laurie, H. and P. Lynn. 2009. "The Use of Respondent Incentives on Longitudinal Surveys." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 205–233. Chichester: Wiley.

Leesti, T. 2010. "Canadian Labour Force Survey Internet Data Collection Pilot Test." Paper presented at the Fifth Workshop on Labour Force Survey Methodology, Paris, April 15–16.

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

Lugtig, P. 2014. "Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers." *Sociological Methods and Research* 43(4): 699–723. Doi: http://dx.doi.org/10.1177/0049124113520305.

Lynn, P. 2009. "Sample Design for Understanding Society." Understanding Society Working Paper 2009-01, ISER, University of Essex, Colchester. Available at: www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2009-01 (accessed 16 March 2017).

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.

Lynn, P. 2014. "Targeted Response Inducement Strategies on Longitudinal Surveys." In *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. Routledge/Psychology Press.

Lynn, P. Forthcoming. "Tackling Panel Attrition." In *The Palgrave Handbook of Survey Research*, edited by D.L. Vannette and J.A. Krosnick. Palgrave.

Lynn, P. and A. Jäckle. Forthcoming. "Mounting Multiple Experiments on Longitudinal Social Surveys: Design and Implementation Considerations." In *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*, edited by P.J. Lavrakas, E.E. de Leeuw, A. Holbrook, C. Kennedy, M.W. Traugott, and B.T. West. Hoboken NJ: Wiley.

Lynn, P. and P. Lugtig. 2017. "Total Survey Error for Longitudinal Surveys." In *Total Survey Error in Practice*, edited by Paul Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West, 279–298. Hoboken, New Jersey: Wiley.

Lynn, P. and D. Lievesley. 1991. "Drawing General Population Samples in Great Britain." London: SCPR.

Lynn, P. S.C.N. Uhrig, and J. Burton. 2010. "Lessons from a Randomized Experiment with Mixed-Mode Designs for a Household Panel Survey." Understanding Society, Working Paper Series, 2010-03.

Martin, P. and P. Lynn. 2011. "The Effects of Mixed Mode Survey Designs on Simple and Complex Analyses." ISER Working Paper Series, 2011-28. Colchester: Institute for Social and Economic Research, University of Essex. Available at: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2011-28 (accessed February 5, 2013).

Schoeni, R., F. Stafford, K. McGonagle, and P. Andreski. 2013. "Response Rates in National Panel Surveys." *The Annals of the American Academy of Political and Social Science* 645: 60–87. Doi: https://doi.org/10.1177/0002716212456363.

Schwarz, N., F. Strack, H.-J. Hippler, and G. Bishop. 1991. "The Impact of Administration Mode on Response Effects in Survey Measurement." *Applied Cognitive Psychology* 5: 193–212.

Souren, M. 2012. "Multi-Mode Surveys at Statistics Netherlands: Implications, Experiences and Open Issues." Paper presented at Opening Conference of the European Statistical System Network (ESSNet) on Data Collection for Social Surveys using Multiple Modes, Wiesbaden, October 11–12, 2012.

Uhrig, S.C.N. 2008. "The Nature and Causes of Attrition in the British Household Panel Study." *Institute for Social and Economic Research Working Paper* 2008-05. Available at: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2008-05 (accessed 16 March 2017).

Uhrig, S.C.N. 2011. "Using Experiments to Guide Decision Making in Understanding Society: Introducing the Innovation Panel." In *Understanding Society: Early Findings from the First Wave of the UK's Household Longitudinal Study*, edited by S.L. McFall and C. Garrington. Colchester: University of Essex. Available at: http://research.understandingsociety.org.uk/findings/early-findings (accessed 16 March 2017).

Voogt, R. and W. Saris. 2005. "Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects." *Journal of Official Statistics* 21(3): 367–387.

Voorpostel, M. and V.A. Ryser. 2011. "Mixed Mode Data Collection as a Strategy to Decrease Panel Attrition in the Swiss Household Panel." FORS Working Paper 2_11. Available at: http://ohs-shp.unil.ch/workingpapers/WP2_11.pdf (accessed 28 March 2017).

Wallace, S., J. Nazroo, and L. Bécares. 2016. "Cumulative Effect of Racial Discrimination on the Mental Health of Ethnic Minorities in the United Kingdom." *American Journal of Public Health* 106(7): 1294–1300. Doi: http://dx.doi.org/10.2105/AJPH.2016.303121.

Watson, N. and M. Wooden. 2014. "Re-Engaging with Survey Non-Respondents: Evidence from Three Household Panels." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 177(2): 499–522. Doi: http://dx.doi.org/10.1111/rssa.12024.

Received February 2016
Revised December 2016
Accepted January 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 409–426, http://dx.doi.org/10.1515/JOS-2017-0020

Interviewer Effects on Non-Differentiation and Straightlining in the European Social Survey

Geert Loosveldt[1] *and Koen Beullens*[1]

In this article we examine the interviewer effects on different aspects of response styles, namely non-differentiation and straightlining, which in general refers to the tendency to provide the same answers to questions in a block of questions. According to research about response styles, the impact of the interviewer on this kind of response behavior is rare. Five blocks of items in the questionnaire in the sixth round of the European Social Survey (2012) are used in the analysis. These data also allow for an evaluation of the differences between countries in terms of non-differentiation and straightlining. Five different measurements of these aspects of response style are used in the analysis. To disentangle the impact of respondents and interviewers on these aspects of response style, a three-level random intercept model is specified. The results clearly show interviewer effects on the respondent's tendency to select a response category that is the same as the response category for the previous item. In some countries the proportion of explained variance due to differences between interviewers is larger than the proportion of variance explained by the differences between respondents.

Key words: Response style; three level random intercept model.

1. Introduction

Many survey questionnaires contain lists of statements about a particular topic or objects with the same response categories (for example, answers on a five-point scale ranging from 'strongly agree' to 'strongly disagree'). Respondents are asked to think about an object or statement and to select the response category that matches their opinion or position. Researchers assume that each respondent's rating is based on a thorough evaluation of all the response categories when answering this type of question. However, researchers are also aware that respondents sometimes insufficiently differentiate between questions and select the same response category for all the items or objects on a list, even though the items are not identical and may express something different. The tendency to provide the same answers to all of the questions in a block of questions about the same topic is called non-differentiation or straightlining. This kind of response style can be considered a source of systematic measurement error. It is clear that when all the items concerning a particular topic are formulated in one direction (positive or negative), straightlining can have a serious impact on the correlation between the items.

[1] Centre for Sociological Research, KU Leuven, Parkstraat 45, 3000 Leuven, Belgium. Emails: Geert.Loosveldt@kuleuven.be and koen.beullens@kuleuven.be

Straightlining can be considered as one of the types of respondent satisficing, a term coined by Simon (1956) and was later introduced in survey methodology research. Satisficing can occur when the cognitive demands of answering questions exceed the respondent's motivation and/or ability to optimize the response process (Krosnick 1991). Research into the stability of different response styles concludes that these are stable individual characteristics (Weijters et al. 2010a). This means that the respondent is mainly responsible for straightlining and that the respondent's characteristics and personality are relevant to explaining their response style. In their literature overview, Van Vaerenbergh and Thomas (2013) conclude that sociodemographic variables affect response styles, but the findings are not always consistent. There is more support for the relationship between personality and response styles, which is particularly true for extreme response styles. However, demographic and personality variables explain only a relatively small proportion of the variance in response styles, whereas culture and country-level characteristics seem to explain a relatively large proportion (Van Vaerenbergh and Thomas 2013). This result indicates that it is relevant and advisable to explain response styles not only based on respondent characteristics, but also on relevant contextual factors. In the current article, we consider the presence of an interviewer in face-to-face interviews as an obvious relevant contextual element. In the next section, we review literature and research concerning the impact of interviewers on response styles.

The importance of this article is twofold. First, we seek to provide evidence as to whether interviewers – as well as respondents – contribute to response styles involving non-differentiation and straightlining. Second, the European Social Survey (ESS) is used for the data analysis. This survey has become an increasingly important source of scientific output in the social sciences. According to Malnar and Müller (2014), using Google Scholar, some 89 publications based on ESS data could be found in 2003, and this number increased to 381 annual publications in 2013. As of June 2016, there are 94,317 registered data users of the ESS (url: http://www.europeansocialsurvey.org/about/user_statistics.html). For this reason, we seek to attract the attention of social science researchers in order to mitigate their scientific claims based on this face-to-face collected data source if interviewer variance is observed regarding response styles related to straightlining or non-differentiation.

2. Interviewers and Response Styles

Although the factor 'interviewer' is not completely absent in studies about response styles in face-to-face interviews, interviewers certainly do not have a dominant place in this type of research. Hox et al. (1991) identified interviewer effects on acquiescence after controlling for some relevant respondent variables. They measured acquiescence by the number of 'agree' responses to all the items on a balanced scale, and this indicates a respondent's systematic tendency to give answers irrespective of the content of the questions or items. In their analysis it was not possible to explain this interviewer effect by using the available interviewer characteristics. More recent research confirms the previously observed presence of interviewer effects on acquiescence, and shows that interviewer experience, after controlling for the length of the interviews, can explain a significant but small part of the variance (1.2–1.3%) in acquiescence across interviewers.

More experienced interviewers obtain higher levels of acquiescence than inexperienced interviewers, which means that interviewer experience is associated more clearly with a variance in acquiescence compared with the respondent's education level (Olson and Bilgen 2011). These results support the idea that, in face-to-face interviews, a response style is not just a matter of respondents' cognitive efforts, but also relates to how interviewers deal with this particular response behaviour. Olson and Bilgen (2011) conclude that models for acquiescence focus on respondent characteristics but tend to ignore the role of interviewers.

We examine whether the scarce results related to interviewer effects on acquiescence extend to the response style that can be termed non-differentiation, or straightlining. Based on the observation that a response style can be influenced by contextual factors, we expect that interviewers will also have an effect on a respondent's tendency to select the same or a nearby response category. We assume that in face-to-face interviews, not all interviewers react in the same way to this kind of response behaviour, and that some interviewers might be more inclined than others to facilitate or inhibit non-differentiation or straightlining. As a consequence, one can expect that differences between interviewers explain a significant proportion of the variability of such response styles. In the analysis of interviewer effects, we control for some respondent characteristics that are suggested to be related to respondent satisficing (motivation and ability). This also means that we take into account the differences between interviewers concerning the group of respondents. Based on the previous observed difference between countries in terms of response style, it also seems appropriate to evaluate the differences between countries.

All this makes clear that it is relevant to evaluate the impact of interviewers on straightlining and non-differentiation while controlling for some relevant respondent characteristics, and that it is advisable to do this for different countries. This general objective is specified in the data and models section. First, we start with a discussion of the different measurements of straightlining and non-differentiation, which will be the dependent variables in our analysis.

3. Measurement of Straightlining and Non-Differentiation

In relevant literature, several approaches are used to measure different types of straightlining. The starting point is the general definition of straightlining as the tendency to give the same answers to questions regardless of their content. Clear examples of straightlining can be observed in online surveys, where multiple items with the same response scale are displayed in a grid. In such a grid, the items are the rows and the response categories are the columns. A pure pattern of straightlining occurs when the selected answers are in a perfect vertical line, which means that the same response category was selected for each item. In their analysis of the association between speeding and straightlining in online surveys, Zhang and Conrad (2013) use eight grid questions with more than two statements and the number of grid questions on which respondents straightline (pure straightlining pattern) is used as a measurement of this response style. This specific operationalization of the measurement of straightlining can be used to discuss several aspects that are relevant to measuring straightlining.

3.1. Different Aspects of the Measurement of Straightlining and Non-Differentiation

3.1.1. Homogenous Versus Heterogeneous Sets of Items

In the example from Zhang and Conrad (2013), the topic of the items is the same within each grid (for example, (non)-working mothers, the role of father and mother in the household, etc.). This means that items in a grid are relatively homogeneous and express different aspects or nuances of the same topic. It is assumed that respondents take these nuances into account during the cognitive process in which they create and formulate their answers, and we therefore expect answers to be similar or consistent, but not identical. The substantially consistent answers are responsible for the correlation between the obtained answers for different items. It should also be noted that a response style such as non-differentiation or straightlining is an underlying factor that contributes to the correlation between items. When there is a pure straightline pattern for all respondents, the correlation between the items will be perfect. It is clear that this is not desirable.

Instead of using a homogeneous set of items about a topic, an alternative is to select a heterogeneous set of items, which can be presumed to be only moderately or poorly related. Greenleaf (1992) considers a set of items with low inter-item correlations as a prerequisite for creating a measurement of an extreme response style. For example, to create a set of heterogeneous items, Weijters et al. (2010b) randomly sampled 21 items from the same number of unrelated marketing scales. This procedure resulted in low inter-item correlations. Unrelated or poorly related items are considered as a necessary condition to ensure that the systematic tendency to select a response category is 'regardless of the content'. We have already noted that straightlining can increase the correlation between items. The question can be posed as to whether it is still possible to observe straightlining or non-differentiation with a set of independent items, because it appears difficult to observe the cause (response style) in a condition where the effect or result (correlation) is supposed to be completely absent.

3.1.2. General Straightlining or Non-Differentiation Versus Specific Response Categories

The second observation on the measurement procedure used in Zhang and Conrad's article (2013) is that straightlining is not specified for a response category. For example, no difference is made between the systematic use of extreme response categories or the middle scale category. Therefore, one can consider this a general measurement of straightlining in comparison with a specific measurement for the systematic tendency to select a specific response category, for example Extreme Response Style (ERS) or Midpoint Response Style (MRS).

3.1.3. Pure Pattern Versus Tendency

In the approach of Zhang and Conrad (2013), only a pure pattern – choosing the same response option for all the items in a grid – is considered to be evidence of straightlining. This is the traditional view on the operationalization of straightlining. In this article, we extend the operationalization of this concept in a more flexible way in order to obtain a

variety of tendencies of straightlining, or the closely adjacent concept of non-differentiation. Particularly, the number of items in a grid is not taken into account when measuring pure straightlining. This means that it is easier to fulfil the condition of straightlining when the grid contains only a small number of items (for example, three instead of eight). When the number of items in a block of questions on a topic increases, one can assume that it will be more difficult to observe a pure pattern. However, the absence of a pure response pattern does not mean that there is no clear tendency to systematically select a particular response category. The most obvious variant of the binary assessment (present or absent) of pure straightlining is counting the number of items with the same score (all rated the same, all but one rated the same, etc.) (Krosnick and Alwin 1988). One can also count the number of times a response category is selected and calculate the log odds ratio. The odds refer to the ratio of the number of times that a response category is selected, to the number of times that this category is not selected (Weijters et al. 2010). Other measurements of straightlining are based on the proportions or percentages of responses in a particular category.

3.1.4. Is the Order of the Responses Taken into Account?

This criterion is perhaps the crucial element that distinguishes straightlining from non-differentiation. The response sequence 7, 6, 6, 7, 6, 7 indicates the same degree of non-differentiation as the response sequence 6, 6, 6, 7, 7, 7. Nevertheless, the second sequence is more likely to provide evidence of straightlining. Non-differentiation is usually measured using a distance metric, such as the standard deviation of the responses, or the average square root of the absolute difference between any two answers from the same respondents to a block of questions (Chang and Krosnick 2009).

3.1.5. Response Scale Format

The eight grid questions used in the article by Zhang and Conrad (2013) do not all have the same response scale: there are six five-point scales ('fully disagree' to 'fully agree'), one different five-point scale ('certainly not' to 'certainly yes') and one three-point scale ('full-time', 'part-time', 'no job at all'). In their article, Weijters et al. (2010b) demonstrate that the labelling of the scale format components and the number of response categories affect different types of straightlining, and that accordingly, empirical results based on different scale formats may not be comparable. Although one might assume that the tendency to select the same response category will decrease when the number of response categories increases, the results do not support this assumption. This also means that respondents do not necessarily differentiate their answers when a scale is used that has more response categories. In fact, this is not what a researcher expects when deciding to use a response scale with more categories. Nevertheless, it seems necessary to evaluate straightlining or non-differentiation for a particular type of response scale.

3.2. Indicators of the Measurement of Straightlining and Non-Differentiation

The discussion of several characteristics that are relevant to qualify the measurement of straightlining or non-differentiation makes it clear that there is no evident simple and univocal measurement. Depending on the survey design characteristics (for example, the

mode and frequency of grid questions with a particular response scale) and the research questions, one measurement may emphasize a different aspect of the response style more than another.

Therefore, it seems more appropriate to opt for more than one measurement. This allows us to assess the sensitivity and robustness of the results for different operationalizations of the concept. Five different, but probably closely related, measurements are used in the analyses here.

3.2.1. Pure Straightlining and the Maximum Sequence of Identical Responses

In line with Zhang and Conrad (2013), straightlining is indicated by a 0-1 binary variable, where the presence of straightlining only applies if all the responses are identical. In the data used in our analyses (ESS), most blocks of questions are relatively long, so that pure straightlining is somewhat exceptional ($<5\%$). In addition, because the analyses use a three-level (residual – respondent – interviewer) data structure, the multilevel models will be very likely to fail to converge. Instead, *the maximum string of identical responses* is determined for each respondent and for each block of questions. Item nonresponse (don't know, refusal, or no answer) breaks a sequence, even if the next response is identical to the previous. This indicator is labelled here as 'MAX'. For example, the maximum sequence in '7, 7, 7, 6, 6' is three; the maximum sequence in '7, 7, DK, 7, 7' is two. Notice that the absolute number is used. This measurement takes the order of response into account, but is not related to a particular response category.

3.2.2. The Percentage of Responses That are Identical to the Response to the Previous Question

In the response sequence 6, 6, 8, 8, 6, two out of four responses (50%) are the same as the previous ones (although there are five responses, the first evidently cannot be compared with a 'previous' one). In fact, one can consider this measurement as an indicator of *response inertia*. Similar to the previous measurement, the order or the sequence of the responses is important in order to assess straightlining. Unanswered questions can never contribute to the numerator determining the fraction, but always add to the denominator. For example, the sequence 5, DK, DK, 5 counts zero out of three potential straightline answers. This indicator is labelled here as '%STR'. The next two indicators do not take the order of the sequences into account.

3.2.3. The Standard Deviation of All the Responses of One Respondent in One Block of Questions

This indicates the degree to which respondents differentiate between questions. Higher scores indicate more differentiation (as opposed to the first two indicators, where higher scores indicate less differentiation or more straightlining). This indicator is labelled here as 'SD'.

3.2.4. Mulligans' Score

Mulligan's score is closely related to the standard deviations measurement for straightlining (Chang and Krosnick 2009). It is a distance metric, measuring the average

square root of the absolute difference between any two answers from the same respondents in a block of questions, or:

$$\binom{n}{2}^{-1} \sum_{q=1}^{n} \sum_{q'>q}^{n} \sqrt{|x_q - x_{q'}|}$$

where n is the number of questions in the grid, and x is the answer of the respondent to question q. Similar to other indicators, unanswered questions do not contribute to the calculation of the distance measurements. For example, the responses 4, 5, NA, 6 will generate the same score as if the sequence of scores was 4, 5, 6. This indicator is labelled here as 'MUL'.

3.2.5. The Average Distance Between Two Subsequent Answers

The last indicator combines the distance approach of indicators SD and MULL, and also takes the order of the responses into account. The average distance is determined between response q and the response to the previous $(q - 1)$ question. For example the sequence 6, 5, 6, 4 will have a score of $(1 + 1 + 2)/3 = 1.33$, whereas the result for the sequence 6, 6, 6, 5 is $(0 + 0 + 1)/3 = 0.33$. This indicator is labelled here as 'DEV.PREV'.

It should be noted that none of these indicators will be capable of watertight detection of the response style that is intended to be measured. Although the response sequence 7, 7, 7, 6 is very likely to generate scores that indicate straightlining or non-differentiation, this sequence of responses can still be a truthful reflection of the respondent's beliefs or attitudes. For the purposes of this article, revealing false positives of this type is possible but not really problematic. The overall level of straightlining or non-differentiation is not of primary interest here, as long as it is equal among interviewers. Nevertheless, we expect to observe interviewer variance regarding these indicators, which in turn should alert researchers who use the ESS data (or other survey data that is prone to such interviewer effects) that the data is not faultless and that, as a consequence, it should be treated cautiously. As already mentioned, non-differentiation or straightlining may artificially increase the correlations between items and it is explicitly not expected that individual interviewers will advance such processes. Evaluation of these interviewer effects is the main objective of this article. Additionally, the presence of such interviewer effects should aid the data producers to invest more in interviewer (and questionnaire) management in order to avoid these unwanted effects.

4. Data and Models

Data from the European Social Survey Round 6 (ESS6) is eminently suitable for the analysis of interviewer effects in general and on straightlining in particular. The ESS6 was organized in 2012 in 29 European countries (see website: http://www.europeansocialsurvey.org/data/download.html?r=6) and the data allows us to evaluate interviewer effects on straightlining within and between countries. Five blocks of items in the questionnaire of the ESS6 (2012) are used in this analysis. A block consists of consecutive items measured on an eleven-point response scale. The topic of the items and the labels of the extreme points of the eleven-point response scale can vary within one

block. Therefore, a block is not necessarily a homogeneous set of items with the same eleven-point scale. Blocks 1, 4, and 5 are more homogeneous (the same topic and the same eleven-point response scale), whereas Block 2 and Block 3 are more heterogeneous (several topics and different eleven-point response scales). The number of the block corresponds to its order in the questionnaire.

- Block 1 (B2–B8). Political trust: seven items about trust in the police and several political institutions; one eleven-point scale (0 = no trust at all; 10 = complete trust).
- Block 2 (B18d–B25). Evaluation of politics and policy: nine items about the importance of and satisfaction with democracy and the state of the education and health services; five different eleven-point scales.
- Block 3 (D28–D35). Wellbeing: eight items about the time respondents have to do things they really want to do and how much of the time they generally are interested in, absorbed in or enthusiastic about what they are doing; four different eleven-point scales.
- Block 4 (E1–E15). Democracy in general: 15 items about democracy in general; one eleven-point response scale (0 = not at all important for democracy in general; 10 = extremely important for democracy in general).
- Block 5 (E17–E30). Democracy in the country: 14 items about democracy in the respondent's country (0 = does not apply at all; 10 = applies completely).

For each respondent, we can calculate each indicator presented in the previous section for each of the five blocks separately. Each of these indicators in each block can be considered to be a repeated measurement within a respondent producing this response style. Therefore, for each respondent, there are five measurements for each indicator (one for each block) nested within the respondent. This results in a three-level hierarchical data structure that can be analyzed using a three-level random coefficient model. The five measurements in each block are the first or lowest level (measurement level), the respondents are the second level (respondent level) and the interviewers are the third or highest level (interviewer level) in this hierarchical data structure. Within this structure, $INDIC_{bij}$ is the measurement of one of the five indicators in a block b for respondent i interviewed by interviewer j (with $b = 1,2,3,4,5$; $i = 1, \ldots I$; $j = 1, \ldots J$), therefore all five indicators are separately used as dependent variables in the specified models. The country level is not considered a level in the data structure, and the analysis is carried out separately for each country. The main reason for this choice is that countries might follow very different strategies to recruit, pay, train, and monitor their interviewers. The comparisons of the results of separate analysis in each country will clearly demonstrate the prevalence of the response style, as well as interviewer variances regarding the response style in each country.

The first model for $INDIC_{bij}$ in the three-level data structure is as follows:

$$INDIC_{bij} = \gamma_{000} + \sum_{b=2}^{5} \gamma_{b00} Block_b + \mu_{0j} + k_{0ij} + \varepsilon_{bij} \qquad \text{(Model 1)}$$

In this model, the only independent variable is the block information for which 5–1 parameters (the first block is the reference category) are accommodated, and μ_{0j}, k_{0ij}, and ε_{bij} are respectively the unique parts of the intercepts at the interviewer level, the

respondent level and the error at the measurement level in the block. The variances of these unique parts of the intercept are, respectively, $\sigma^2_{\mu 0}$, σ^2_{k0}, and σ^2_{ε}. With this model, we can break down the variance across the three levels and calculate the proportion of variance explained by the respondent ($\rho_{respondent}$) and the interviewer ($\rho_{interviewer}$) (Hox 2010; Loosveldt and Beullens 2013). The expressions for the proportions of explained variance are:

$$\rho_{respondent} = \frac{\sigma^2_{k0}}{\sigma^2_{\mu 0} + \sigma^2_{k0} + \sigma^2_{\varepsilon}};$$

$$\rho_{interviewer} = \frac{\sigma^2_{\mu 0}}{\sigma^2_{\mu 0} + \sigma^2_{k0} + \sigma^2_{\varepsilon}};$$

Both expressions are appropriate to evaluate the impact on the indicator of the response style both of interviewers and respondents. It is preferable for the proportion of variance explained by the interviewer to be small, or at least much smaller than the proportion of variance explained by the respondent.

The interpenetration of interviewers and areas was not accommodated during the design stage. This means that area effects may erroneously be taken for interviewer effects (and vice versa). Therefore, we extend the model by adding covariates at the respondent level in order to make the groups assigned to the interviewer more similar. A first extension (Model 2) includes the following variables:

- level of education of the respondent (a seven-point EISCED scale),
- gender of the respondent,
- age of the respondent, and
- rank (logarithm). This count variable indicates the chronological rank of the respondent within each interviewer. It may be expected that a response style is more likely to be observed as the interviewer becomes more familiar with the survey or its questionnaire.

After including these covariates, the model is estimated as if all respondents have a similar age, gender, level of education, and rank within the interviewer. In Model 3, two area variables are included:

- Density: a five-point scale indication of population density, self-reported by the respondent.
- Region: a geographical area that usually coincides with provinces, counties or any other NUTS2 or NUTS3 subnational entity.

Adding these variables may have a strong filtering effect, distinguishing area and interviewer effects. However, since interviewers are recruited, trained, monitored, and coached along the same geographical lines, real interviewer effects may erroneously be taken for area effects. In this way, a larger than ideal amount of real interviewer effects may be separated out of the interviewer variance. A final extension (Model 4) is made by adding the variables that indicate the motivation and cognitive skills of the respondent, as assessed by the interviewers:

- RESBAB: the respondent tried to answer the questions to the best of his/her ability (five-point scale).
- RESUNDQ: the respondent understood the questions (five-point scale).

A potential problem of adding these covariates to the model is that interviewers rate their own respondents. Because it is very likely that these variables are prone to containing interviewer effects themselves (on average countries show intra-interviewer correlations of 0.34 for RESBAB and 0.20 for RESUNDQ), there is a risk that these variables artificially explain too much interviewer variance regarding the response style of non-differentiation or straightlining.

We run the models per country. This is particularly relevant, because the main responsibilities for interviewer management are located at the country level. Each of the participating countries in the ESS needs to recruit, train, remunerate, and monitor its own interviewers. From the ESS6 documentation report (which can also be accessed using the link shown earlier), it becomes immediately clear that countries do not follow a uniform approach regarding the length of interviewer training, the materials they use for the training, interviewer payment, and so forth. Therefore, providing results per country seems to be most appropriate.

5. Results

We start with descriptive results for one of the indicators of non-differentiation and straightlining in the ESS6 (%STR). The first column in Table 1 (Frequency) shows the frequencies with which the categories on the eleven-point scale are chosen. In total, 2,897,669 answers are considered, originating from 54,673 respondents in 29 countries, with each respondent giving 53 answers on an eleven-point scale ($53 \times 54{,}673 = 2{,}897{,}669$). Each of the 53 items belongs to one of the five blocks.

Table 1. Frequencies and percentages of straightlining on an eleven-point scale for 53 items and 54,673 respondents (ESS6).

Scale point	Frequency %	%STR %
0	5.07	48.91
1	2.67	29.84
2	4.08	23.97
3	5.37	22.28
4	5.60	20.45
5	11.26	28.53
6	8.41	21.70
7	11.47	26.08
8	13.69	31.58
9	8.96	33.49
10	19.42	65.53
Refusal (77)	0.16	61.11
Don't know (88)	3.78	43.35
No answer (99)	0.08	51.06

It is apparent that scale-point 10 is chosen most frequently, followed by 8, 7, and 5. Answer 1 and the nonsubstantive answers (don't know, refusal, and no answer) are given the least often. The second column in Table 1 (% STR) shows which scale points are more prone to having the same answer as that given to the preceding question (the second indicator as presented in Section 3: %STR). For example, 48.91% of all the 0 answers were also 0 answers to the previous item. From Table 1, it seems that straightlining or non-differentiation is more likely to occur in the extreme categories, as well as for nonsubstantive categories. However, we do not consider nonsubstantive answers to count as non-differentiation or straightlining in our analysis. It should be noted that this second column is not based on 53 items, but on 48 items (53−5), because providing a similar answer as that to the previous question cannot be assessed for the first item of each block.

Figure 1 shows the average percentage of straightline answers (%STR) in each country for the five blocks and for all the blocks together. The means in the subtitle of the figure are the mean percentages of straightline answers (%STR) in the block for all respondents. For example, for the first block on political trust containing seven items, on average 34.95% of

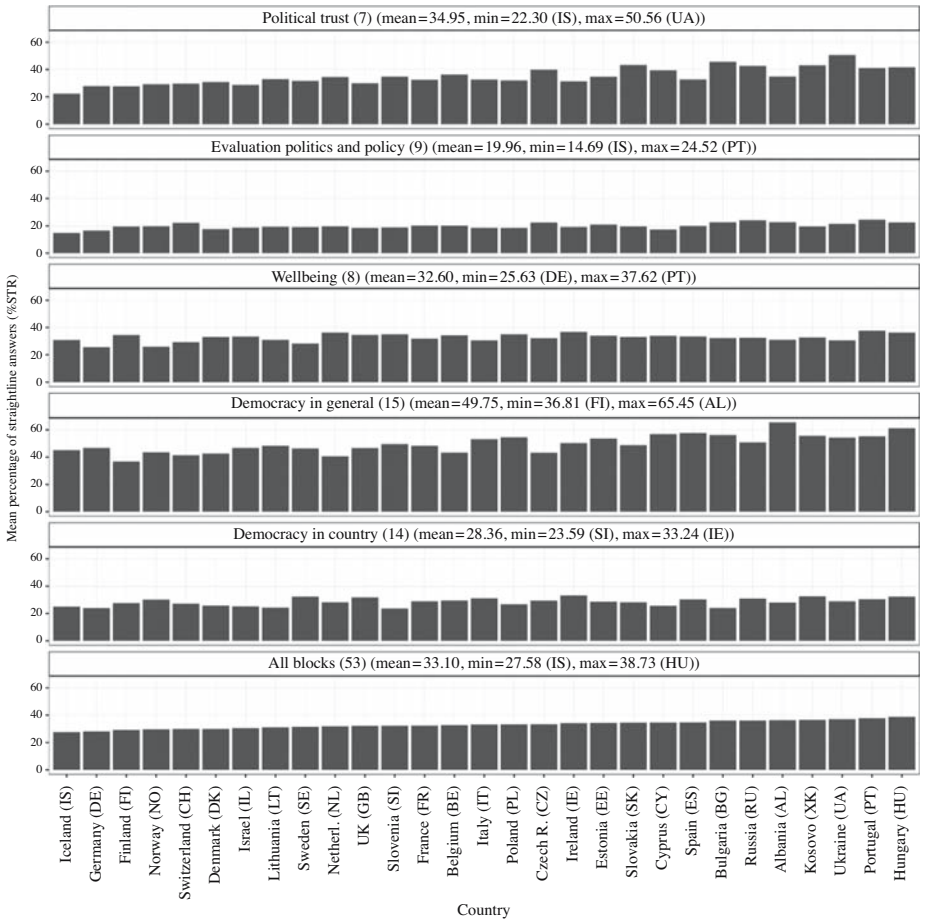

Fig. 1. Mean percentage of straightline answers (%STR) in separate blocks and all blocks together in each country (ESS6).

the answers are identical to the answer to the previous question, with Iceland (IS) showing the minimum (22.30%) and Ukraine (UA) the maximum at 50.56%. The overall view makes clear that there are differences in straightlining between blocks and within blocks between countries. The average percentages of straightlining are especially high for the questions about democracy in general (Block 4). In this block, the mean percentages of answers that are the same as the answers to the previous items are about 60 per cent in some countries. The average percentages are much smaller for the similar questions about democracy in the respondents' own country (Block 5). The more heterogeneous Block 2, with questions about the evaluation of politics and policy, seems to be less sensitive to the tendency to select a response category that is the same as that for the previous item. This makes clear that the measurement used is indeed, as expected, also affected by the homogeneity of the items used. On the other hand, the other block, which was characterized as more heterogeneous with questions about wellbeing, shows higher percentages of straightlining, and is comparable with the other more homogeneous blocks.

So, Figure 1 presents the information concerning one of the five indicators of straightlining and non-differentiation (%STR). Similar figures for the four remaining indicators ('MAX', 'SD', 'MUL', 'DEV.PREV') are available on online supplemental file of Journal of Official Statistics website (available at: http://dx.doi.org/10.1515/jos-2017-0020). In our analysis, we also used the four other indicators, namely 'Max', 'SD', 'MUL', and 'DEV.LAST.

For each respondent and each of the five blocks, each of the five indicators can be determined based on the sequence of the answers given. Table 2 shows the average correlations between these five indicators, only considering measurements within the same block. This means that, for example, 0.89 is the average of five correlations: the correlation between '%STR' and 'MAX' in Blocks 1, 2, 3, 4, and 5. These correlations may be quite high because the indicators are all measured based on the same sequences of answers. Therefore, Table 3 shows the same average correlations, but only correlations between different blocks are allowed to contribute. For example, 0.18 is the average of the correlation between %STR and MAX in ten combinations of Blocks (Block 1 and Block 2, Block 1 and Block 3, . . . Block 4, and Block 5). These average correlations tend to be much lower. It is apparent that two groups of indicators can be distinguished: '%STR' and 'MAX' (similarity measures) tend to be similar, and 'SD', 'MUL' and 'DEV.PREV' (distance measures) also tend to cluster.

The key research question in this article is how much of the observed variability in these indicators of straightlining or non-differentiation can be attributed to the respondent level

Table 2. Average correlations between the five indicators measured within the same blocks. ESS6, 54,673 respondents.

	%STR	MAX	SD	MUL	DEV.PREV
%STR	1.00				
MAX	0.89	1.00			
SD	−0.32	−0.30	1.00		
MUL	−0.60	−0.59	0.87	1.00	
DEV.PREV	−0.55	−0.48	0.83	0.80	1.00

Table 3. Average correlations between the five indicators measured within different blocks. ESS6, 54,673 respondents.

	%STR	MAX	SD	MUL	DEV.PREV
%STR	0.19				
MAX	0.18	0.17			
SD	−0.01	0.01	0.21		
MUL	−0.08	−0.07	0.13	0.12	
DEV.PREV	−0.07	−0.05	0.17	0.14	0.17

and in particular, how much to the interviewer level. We assume that a response style is not only a matter of respondent behaviour, but in line with the results of Olson and Bilgen (2011), that it may be affected by the impact of the interviewer. As a result, we expect to observe that some part of the variance of any of the five indicators for non-differentiation or straightlining as a response style is explained by the interviewers. Because the five indicators tend to be correlated, but potentially measure different aspects of the response style related to straightlining or non-differentiation, it seems appropriate to present the results for all five indicators. In this way, we can assess to what extent our analysis of interviewer and respondent variance is sensitive to the choice of indicator. First, we will discuss the results regarding one indicator ('%STR'), after which a summary of the four remaining indicators will be provided.

Figure 2 shows per country how much variance of '%STR' can be attributed to the interviewer level (black lines) and to the respondent level (grey lines). These lines result from applying Model 1 and its extension when respondent background characteristics are added (Model 2), subsequently area variables are added (Model 3), and variables are added that indicate the motivation and cognitive skills of the respondents (Model 4). The variance components under Model 1 are shown on the left sides of each subgraph and connected with the variance components of the three other models. In some countries, the estimated interviewer effect depends on the applied model. This can be observed in Bulgaria, Hungary, Italy, Kosovo, Lithuania, Portugal, and Slovenia. Particularly between Model 2 and 3 (where area variables are added), the shift is strongest. Nevertheless, as already mentioned, this covariate information may not only contain regional effects between respondents (which is the major reason for including these variables), but may also provide unintended information explaining differences between interviewers, as their interviewing styles may also be trained locally. In the other countries, there are no considerable changes in the estimated shares of variances of interviewers and respondents according to the applied model. Only the estimated impact of the respondents may in some countries differ between Model 1 and 2 (after the inclusion of the respondent variables age, gender, and level of education, and the variable indicating which chronological rank the respondent has within the interviewer). Such clear effects can be observed in Denmark and Belgium.

There are many countries in which the interviewers tend to have a rather small share in the variance of the indicator '%STR'. These countries are Albania, Belgium, Denmark, Finland, France, Germany, Iceland, Israel, Italy, the Netherlands, Norway, Spain, Switzerland, and the United Kingdom. In these countries, it can be observed that the black

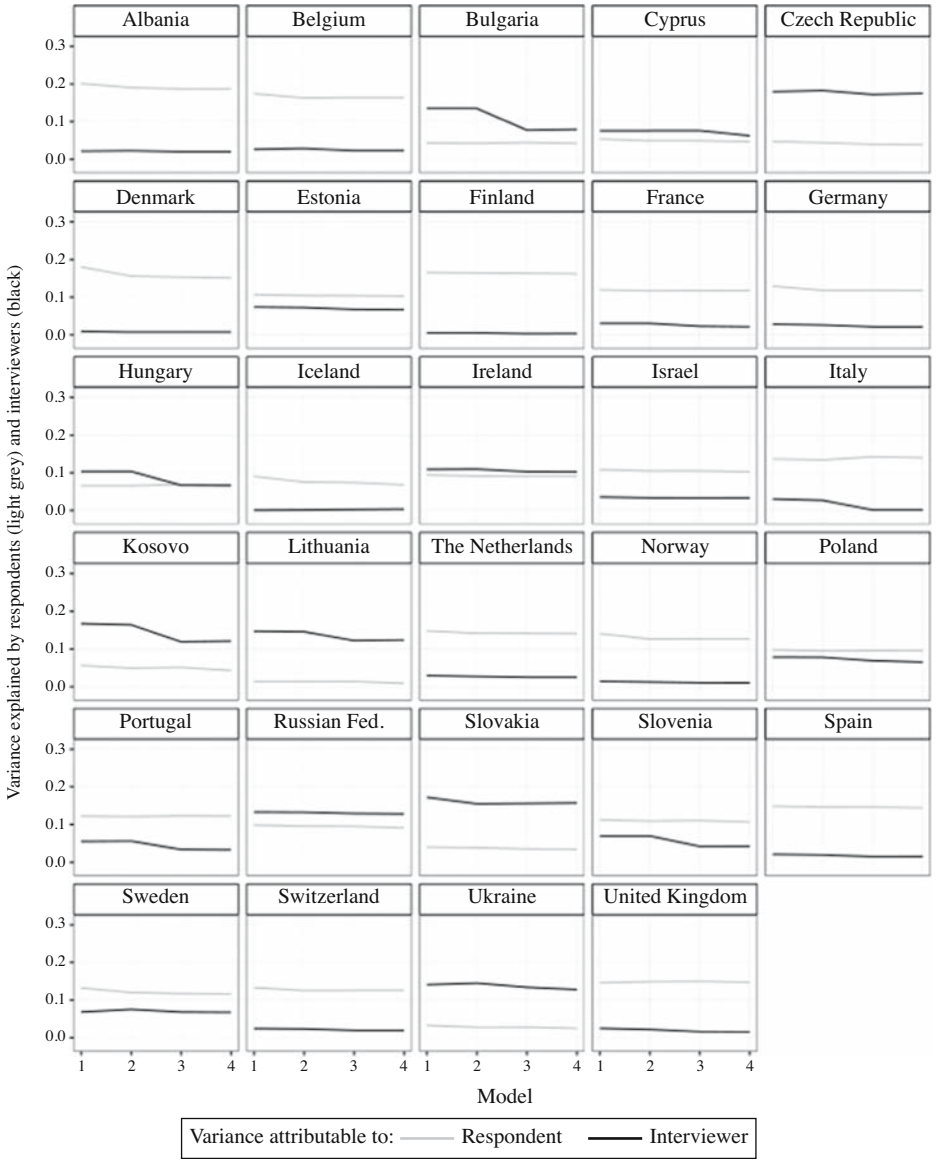

Fig. 2. Intra-respondent (light grey) and intra-interviewer (black) correlations for the '%STR' indicators for four model specifications in 29 countries of the ESS6.

line is very close to the origin (although in most cases it is still statistically significant from zero). In some other countries, interviewer effects on '%STR' are more substantial, although the respondent effects on this indicator are also considerable. This can be observed in Cyprus, Estonia, Hungary, Ireland, Kosovo, Poland, Portugal, Russian Federation, Slovenia, and Sweden. Finally, in a few countries, the interviewer effects clearly surpass the impact of the respondents. This applies to Bulgaria, the Czech Republic, Lithuania, Slovakia, and Ukraine. It is clear that in these countries, intra-interviewer correlations of 0.10 and more are no exception. Therefore, in these countries,

interviewers explain a considerable amount of the variation in the response style '%STR' and it becomes difficult to claim that straightlining or non-differentiation is a trait that is only attributable to respondents.

The graphs of the four remaining indicators ('MAX', 'SD', 'MUL', 'DEV.PREV') are available on online supplemental file of Journal of Official Statistics available at: http://dx.doi.org/10.1515/jos-2017-0020. Most of the results as portrayed in Figure 2 also apply to the other indicators of straightlining and non-differentiation. Nevertheless, some noteworthy differences can be summarized as follows. The indicators 'MAX' and 'MUL' tend to show lower levels of variance components for both interviewer and respondents. 'SD' tends to show higher respondent effects as compared with the four other indicators. Generally, the estimates of interviewer variance are more stable across indicators, whereas the estimates of respondent variance are more varied across indicators. For some countries, some noteworthy remarks need to be made. For Bulgaria, in Figure 2 the respondent variance is about 0.05 for '%STR'. For the other four indicators, this variance reduces to <0.01. This also applies to Kosovo. In Israel, the interviewer estimate for '%STR' is about 0.04 and is therefore much smaller than the respondent impact (about 0.1). However, in the case of 'DEV.PREV' and 'SD', both interviewer and respondent variance are estimated at about 0.1. In Portugal, interviewer variance is estimated to be smaller than the respondent variance regarding '%STR'. However, regarding the four other indicators, respondent and interviewer variance have about the same magnitude.

The parameter estimates for all the added covariates of Model 4 can be found in the online appendix. However, there do not seem to be clear patterns as to which variables are (strongly) predictive of the response style. Gender does not have a clear effect on non-differentiation or straightlining, although in some countries, men tend to have higher scores for 'SD', 'MUL', and 'DEV.PREV'. Generally, age is positively related to '%STR' and 'MAX'. Strangely, age also positively relates to larger differences between the answers (as indicated by 'SD', 'MUL', and 'DEV.PREV'). In some countries, however, the age effect is in the opposite direction. Level of education seems to be related to the response style in some countries, but the direction of the relationship is not very clear, as opposing effects are observed. Population density does not seem to be closely related to the response style, and in the few countries in which effects are observed, they are inconsistent. Because the list of the different provinces or counties used in Model 3 and Model 4 is very long, the related parameter estimates are not provided in the appendix. Nevertheless, regions can differ considerably from one another. This is particularly so in Bulgaria, Hungary, Italy, Lithuania, Portugal, Slovenia, Slovakia, and Kosovo. In these countries, there is a noticeable decrease of the intra-interviewer correlation between Model 2 and Model 3, where the latter includes regional covariates and the former does not.

The effect of the chronological rank (logarithm) of the respondent within the interviewers is not very strong, but it is consistent. Respondents who are interviewed later by the same interviewer tend to show higher levels for '%STR' and 'MAX', indicative of straightlining, and lower levels for 'SD', 'MUL', and 'DEV.PREV', indicative of non-differentiation. In only a few countries, more motivation on behalf of the respondent – as observed by the interviewers (RESBAB) – tends to be related to more differentiation. Whether the respondent understood the questions (as assessed by the interviewer,

RESUNDQ) tends to relate somewhat more strongly to the five indicators, although the directions of these relationships are rather ambiguous. Respondents showing that they understood the questions tend to straightline more ('%STR' and 'MAX'), but also tend to differentiate more ('SD', 'MUL', and 'DEV.PREV').

6. Conclusion and Discussion

In an article about the past, present, and future of total survey error, Groves and Lyberg (2010) conclude that the study of the interplay of various different error sources must be part of the agenda for future survey methodological research. In line with this conclusion, in the current article we try to combine research into response styles with the assessment of interviewer effects, a combination that is rare in survey methodological research. Research concerning response behaviour makes clear that the process can sometimes be characterized by a response style showing a lack of effort to obtain adequate and correct answers. In the total survey error framework, response styles are a source of measurement error for which the respondent is responsible. However, research into interviewer effects makes clear that interviewers can have an impact on the registered responses, and that interviewers can also be considered as another source of measurement error. In the current article, we evaluate the impact of interviewers on the respondent's response style: particularly the tendency to provide the same answer as the one to the previous question (straightlining) and non-differentiation. It should be emphasized that we do not seek to provide evidence that respondents are straightlining or failing to adequately differentiate between questions, but instead we want to provide evidence of the extent to which interviewers mediate these processes. Using different indicators and different data sets from various countries, our results clearly illustrate that in most countries, interviewers have a significant impact on these response tendencies. This makes it clear that analyzing the interplay between the respondent and the interviewer as sources of measurement error was fruitful and that it allows for an interpretation of response styles from a different perspective. Response style is not only a matter of the respondent's cognitive processes, motivation or other characteristics. This interpretation is too limited. Response styles are also influenced by situational factors, and in face-to-face interviews the interviewer is not a negligible factor. The assessment of interviewer effects is a diagnostic analysis. Our results clearly indicate that interviewers can have an impact on response patterns, but we do not know how and why. It is possible that interviewers are suggestive, or that they reinforce a respondent's tendency to select similar response categories. A straightforward way to find out how response styles operate during an interview would be to record and analyze the interviewer-respondent interaction. The results of this type of analysis can also be used during training to remedy interviewers' shortcomings. The observation of substantial interviewer effects highlights the importance of training, and the results of this study reinforce that. During interviewer training, it is advisable to pay adequate attention to response styles. It is necessary to ensure both that interviewers do not induce a response style, and that they know how to handle different kinds of response behaviour during the interview.

Currently, there is an increased attention by survey researchers and practitioners to collect paradata, also during data collection. One of the aims is to monitor and potentially

improve the quality of the data while the data collection process is ongoing. In that sense, regular monitoring of the data of completed interviews can be done in order to assess the degree to which certain interviewers tend to show signs of unfavourable patterns in the obtained answers from their respondents. Such close quality control might be the basis for continual interviewer coaching during the fieldwork.

The observed differences between countries are remarkable and the interpretation of this in terms of 'cultural differences' seems too general and inconclusive. Although it is possible that in some countries straightline answers or non-differentiation are in agreement with the 'true attitude' of the respondent (for example, a very negative evaluation of all aspects of the democratic system), this cannot be an explanation for the differences between interviewers. Differences in 'survey culture and practices' and in fieldwork capacity are probably responsible for differences between countries. The fact that in some countries the differences between the interviewers explain more variability in response tendency than the differences between respondents must, at least, be considered an urgent call to closely monitor the way fieldwork procedures (interviewer training and briefings, follow up of the interviewers during the fieldwork, feedback for interviewers, etc.) are implemented in these countries. One must be aware that these differences in interviewer effects on response style can influence substantive comparison across countries, for example because straightlining or non-differentiation may artificially inflate correlations between survey items. It is necessary to ensure that differences between countries in 'survey culture and practices' are not interpreted as real cultural differences.

7. References

Chang, L. and J. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing versus the Internet. Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 74: 641–678. Doi: http://dx.doi.org/10.1093/poq/nfp075.

Greenleaf, E. 1992. "Measuring Extreme Response Styles." *The Public Opinion Quarterly* 56(3): 328–351. Doi: https://doi.org/10.1086/269326.

Groves, R. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *The Public Opinion Quarterly* 74(5): 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065.

Hox, J. 2010. *Multilevel Analysis: Techniques and Applications*. 2nd ed. New York: Routledge.

Hox, J., E. de Leeuw, and I. Kreft. 1991. "The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: a Multilevel Model." In *Measurement Errors in Surveys*, edited by P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. New York: Wiley.

Krosnick, J. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3): 213–236. Doi: http://dx.doi.org/10.1002/acp.2350050305.

Krosnick, J. and F. Alwin. 1988. "A Test of the Form-Resistant Correlation Hypothesis: Ratings, Rankings, and the Measurement of Values." *Public Opinion Quarterly* 52(4): 526–538.

Loosveldt, G. and K. Beullens. 2013. "The Impact of Respondents and Interviewers on Interview Speed in Face-to-Face Interviews." *Social Science Research* 42(6): 1422–1430. Doi: http://dx.doi.org/10.1016/j.ssresearch.2013.06.005.

Malnar, B. and K.H. Müller. 2014. *Surveys and Self-Reflexivity: A Second-Order Study of the European Social Survey (ESS)*. Wien: Echoraum.

Olson, K. and I. Bilgen. 2011. "The Role of Interviewer Experience on Acquiescence." *Public Opinion Quarterly* 75(1): 99–114. Doi: https://doi.org/10.1093/poq/nfq067.

Simon, H. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63(2): 129–138. Doi: http://dx.doi.org/10.1037/h0042769.

Van Vaerenberg, Y. and T. Thomas. 2013. "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies." *International Journal of Public Opinion Research* 25(2): 195–217. Doi: https://doi.org/10.1093/ijpor/eds021.

Weijters, B., E. Cabooter, and N. Schillewaert. 2010b. "The Effect of Rating Scale Format on Response Styles. The Number of Response Categories and Response Category Labels." *International Journal of Research in Marketing* 27(3): 236–247. Doi: http://dx.doi.org/10.1016/j.ijresmar.2010.02.004.

Weijters, B., M. Geuens, and N. Schillewaert. 2010a. "The Stability of Individual Response Styles." *Psychological Methods* 15(1): 96–110. Doi: http://dx.doi.org/10.1037/a0018721.

Zhang, C. and F. Conrad. 2013. "Speeding in Web Surveys: The Tendency to Answer Very Fast and its Association With Straightlining." *Survey Research Methods* 8(2): 127–135. Doi: http://dx.doi.org/10.18148/srm/2014.v8i2.5453.

Received January 2016
Revised February 2017
Accepted March 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 427–454, http://dx.doi.org/10.1515/JOS-2017-0021

The Influence of an Up-Front Experiment on Respondents' Recording Behaviour in Payment Diaries: Evidence from Germany

Tobias Schmidt[1] *and Susann Sieber*[1]

In this article, we analyse the effect of an incentive experiment on German consumers' recording behaviour on the basis of a one-week diary of their point-of-sale expenditure. Part of the experiment, which was carried out shortly before the consumers began filling in their payment diaries, involved consumers rolling a die with a chance of winning either EUR 20 or nothing, that is, they were randomly assigned an incentive. We ask whether respondents' recording behaviour differs depending on whether individuals win or lose. We argue that winners attach a more positive feeling to the survey than losers and therefore show a stronger commitment to the diary. As the incentive experiment is part of a larger experiment to elicit respondents' risk preferences, we also provide evidence on the effect of conducting up-front behavioural experiments in representative surveys. Our results indicate that the outcome of the lottery (rolling of the die) has an impact on the quantity of transactions recorded, but does not affect other aspects of respondents' recording behaviour, such as item nonresponse or rounding. It also has a negligible impact on substantive measures, such as the cash share.

Key words: Incentives; risk experiments; data quality.

1. Introduction

Behavioural economists and psychologists often conduct experiments using convenience samples (e.g., college students). Recently, however, there has been growing interest in embedding behavioural experiments in representative surveys. In many behavioural experiments, participants receive a monetary payoff, which largely depends on the respondents' behaviour, but often contains a random component as well. The interest of behavioural economists lies in the observed behaviour of the participant (e.g., whether a guaranteed payment is preferred over participation in a lottery), and not in the individual payoffs. However, when these experiments are embedded in a standard survey, the payoffs from the experiment can be interpreted as an incentive payment. The payoff may thus have an impact on participants' attitude towards the survey and ultimately affect their reporting behaviour in other parts of the study, for example, the classic questionnaire. Our article provides evidence on this issue. Failure to acknowledge the incentive effects of

[1] Deutsche Bundesbank – Research Centre, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main, Germany. Email: Tobias.Schmidt@bundesbank.de and Susann.Sieber@bundesbank.de.
Acknowledgments: The authors thank Stephanie Eckman, Edith de Leeuw, participants of the ESRA 2015 and 2015 Total Survey Error conferences, as well as four anonymous referees for valuable comments and suggestions.
Disclaimer: The article represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

behavioural experiments and the possible biases and measurement error they induce in the collected data might lead to a misinterpretation of the survey results.

To the best of our knowledge, we are the first to link behavioural experiments to the literature on the role of incentives in surveys. Due to the fact that our incentive experiment constitutes part of a behavioural experiment, it differs from the existing literature on survey incentives in two important respects. First, we do not focus on participation incentives, but on incentives randomly assigned to participants between two stages of the interview process, that is between a regular questionnaire and the self-completed diary collecting data on payment transactions. We will therefore be able to assess whether incentives have an effect on respondents' answers, given participation. Second, our participants are aware that other participants receive a different incentive or no incentive at all. Thus, they might be disappointed if they receive nothing or, on the contrary, be very pleased if they obtain the incentive. This positive or negative feeling – in addition to the monetary value of the incentive – should result in more pronounced effects of the incentive on commitment to the survey, measured along various dimensions.

To assess the impact of the incentive on respondents' recording behaviour, we consider different indicators of item nonresponse: our focus will be on transactions that are not reported at all. To be more precise, we study the possible underreporting of cash and low-value transactions. In addition, we look at common measures of data quality: the incidence of incompletely reported transactions and the rounding of transaction amounts. Our analysis fits well in the Total Survey Error literature, as our study sheds light on a nonsampling error (Groves and Lyberg 2010; Groves 2004; Biemer and Lyberg 2003), namely measurement error as a result of a specific design feature.

The basis of our analysis is a behavioural experiment eliciting respondents' risk preferences, which is carried out between a standard questionnaire-type data collection and a self-completed one-week diary of consumers' point-of-sale expenditure. In the first stage of the risk experiment, consumers have the choice between receiving a guaranteed payment of EUR 10 and participating in a lottery with an expected value of EUR 10. Risk-averse consumers will choose the guaranteed payment while more risk-loving consumers will go for the lottery option. Consumers participate in the lottery by rolling a die. Participants either win EUR 20 if they throw a 4, 5, or a 6, or nothing if they throw a 1, 2, or a 3. While economists running these kinds of experiments are mainly interested in whether the respondent opts for the guaranteed payment (risk-averse) over participation in the lottery (risk-loving), we are more interested in the lottery part. We re-use the behavioural experiment to learn something about the effects of incentives on consumers' recording behaviour.

The lottery can be interpreted as an incentive experiment, with the random assignment of an incentive of EUR 20 for the "winners" and nothing for the "losers" resulting from the rolling of the die. In addition to the monetary value, "winning" the game may induce a positive attitude towards the survey. Both the monetary incentive itself and the positive attitude from winning should lead to a higher commitment to the survey and may affect the respondents' recording behaviour in the diary.

In the analysis of incentive effects, we will focus solely on the lottery and restrict the sample to respondents who participated in the lottery. It is necessary to exclude the risk-averse consumers who chose the guaranteed payment in the first stage of the risk

experiment, because the payoff they receive is not assigned randomly, but chosen by the respondents themselves based on their risk preference. Hence, this group of respondents differs systematically from respondents participating in the lottery. It nonetheless makes sense to analyse the behaviour of those risk-loving consumers who chose the lottery option over the guaranteed payment. If the recording behaviour of this selected group differed depending on whether it wins or loses, the survey results could be biased.

We find that the payoff from the lottery seems to have some impact on consumers' commitment to the diary part of the study in that these respondents record more transactions. However, it does not induce a bias on the key qualitative results, such as the share of cash payments. Our results indicate that the monetary incentive, as well as its non-monetary component ("winner" vs. "loser"), given to the participants during the interview, do not lead to biased outcomes in subsequent parts of the survey. This is good news for economists who plan to embed behavioural experiments with a random payoff in representative surveys. The incentives' main effect seems to be the increase in the number of transactions recorded during the first few days of the diary recording period. Survey designers with short diary studies (e.g., one or two days) may thus be able to increase the number of transactions recorded by respondents by paying an additional incentive during the interview, that is, before the diary recording period starts. However, in general, paying an unconditional incentive during an interview or different parts of an interview has very limited impact on the recording behaviour of respondents.

2. Related Literature

Our analysis can be linked to issues related to measurement errors discussed within the Total Survey Error (TSE) framework (see Groves and Lyberg 2010 for an overview). We are mainly concerned with measurement error due to respondents' recording behaviour, including item nonresponse and omissions by respondents (see Weisberg 2005; Biemer and Lyberg 2003). Biemer and Lyberg (2003) suggest, describing the merits of the TSE framework, that "among the set of alternative designs, the design that gives the smallest total survey error (for a given fixed cost) should be chosen." (Biemer and Lyberg 2003, 850). Our results can help survey designers wanting to include behavioural experiments in their studies to gain a better understanding of the consequences of the experiment for the total survey error and to make a more informed choice about where in the interview/study process to use behavioural experiments.

Our study is also closely related to the literature on the effects of incentives on data quality and respondents' answering behaviour. From a theoretical point of view, incentives can have an impact on the recording behaviour due to a stronger commitment of the respondent to the survey. It is a well-established fact that responding to survey questions is a complex process that imposes a burden on the interviewee (see, for example, Jones 2012; Sharp and Frankel 1983; Groves et al. 1992). Various methods to reduce or counter the response burden for interviewees have been proposed (see Hedlin et al. 2005, for business surveys), among them fostering respondents' interest in and commitment to the survey (Bonke and Fallesen 2010; Davern et al. 2003). In this vein, we argue that providing respondents with an incentive that has both a monetary component (EUR 20) and a non-monetary component ("winner" vs. "loser") will increase their commitment to

the survey and subsequently influence their reporting behaviour. There is some evidence in the literature that this is indeed the case. The study by Bonke and Fallesen (2010) on Danish data comes closest to what we are researching in this article. Bonke and Fallesen (2010) study how different incentives, paid out through a lottery game, explain people's participation rates, choice of survey mode (CATI/CAWI), and data quality in a large-scale Danish survey on time-use and consumption. What makes the study particularly interesting for us is that they investigate respondents' behaviour in the survey diary context. However, in contrast to our study, their participants are not aware that there are lotteries with varying prizes. They find a strong effect of incentives on response rates as well as on mode choice, but no effect *per se* on respondents' behaviour. Neither item nonresponse in the regular questionnaire of their study, nor the number of reported activities or consumed goods and services in the diary differ significantly depending on the incentives provided, if both CATI and CAWI respondents are analysed. They do find some positive effects of incentives with respect to reporting behaviour and data quality for CATI interviewees only. The fact that the impact of incentives on respondents' recording behaviour may be rather limited has also been documented by other researchers. Davern et al. (2003) and Shettle and Mooney (1999) investigate the impact of incentives on classic measures of data quality, such as item nonresponse and the number of edited variables/cases. They find that (prepaid) monetary incentives do not have an impact on data quality. Similarly, Tzamourani and Lynn (1999) show that there is no clear effect of incentives on respondents' recording behaviour, concluding that ". . . the incentives did not affect the respondents' answers in any way, that is they did not induce bias in the responses." (Tzamourani and Lynn 1999, 16). Göritz (2005) documents for a web-based survey that if respondents are offered an incentive that is contingent on completing all relevant questions in the questionnaire, their reporting behaviour in terms of the number of omitted questions and other quality indicators does not differ from that of respondents who are not offered an incentive. The same seems to hold true for web-based studies using access panels (Göritz 2004).

Whether incentives have a positive, negative or no effect at all on respondents' reporting behaviour is nonetheless still an open question. Studies by Goldenberg and Ryan (2009), Singer et al. (2000), Willimack et al. (1995), James and Bolstein (1990), and several of those cited in Laurie and Lynn (2009) found – contrary to the studies cited above – that incentives do have an effect on reporting behaviour and data quality. Goldenberg and Ryan (2009) report that in the US Consumer Expenditure Diary Survey, respondents receiving a pre-paid monetary incentive of USD 20 or USD 40 reported more transactions and also performed better on other indicators of data quality. A similar result is reported by Goldenberg et al. (2009) for the same type of incentives used in the Consumer Expenditure Interview Survey. Singer et al. (2000) show that for some households, paying an incentive reduces item nonresponse. However, the effect is very small: "Only 7 percent is explained by both the demographics and the incentives, and incentives alone explain less than 1 percent of the variance in item non-response." (Singer et al. 2000, 180). They also find an impact of incentives on the distribution of responses. Respondents receiving an incentive seem to be in a better mood (see also Schwarz and Clore 1996) and report more optimistic expectations. Willimack et al. (1995) summarise their findings: "In addition, evidence suggests greater response completeness among

responding incentive recipients early in the interview, with no evidence of increased measurement error due to the incentive" (Willimack et al. 1995, 78). James and Bolstein (1990) find that what they call "large" prepaid incentives of USD 2 lead respondents to expend more effort on completing questions in a mail survey. They measure greater effort by the length of the respondents' answers, the number of comments and number of words written. Interestingly, they also find that large incentives increase ". . . comments that were more favourable towards the survey sponsor" (James and Bolstein 1990, 346), which signals a stronger commitment to the survey. James and Bolstein (1990) cite several older studies (e.g., Godwin 1979, and Shuttleworth 1931) which have also found that respondents receiving monetary incentives have a tendency to provide more comments and more complete responses. A similar result has been found by Goetz et al. (1984).

Given the mixed evidence of incentives on respondents' recording behaviour, it is difficult to derive a clear hypothesis. However, in contrast to the studies cited above which find no effect of incentives, the monetary incentive in our study is substantial (EUR 20) and also has a non-monetary component ("winner" vs. "loser"). We thus formulate the hypothesis that the incentive will have an effect on respondents' diary recording behaviour.

Our study also provides evidence on aspects of incentives that have, to the best of our knowledge, not yet been addressed in the literature. The papers cited above mainly deal with conditional and unconditional *participation* incentives, that is, incentives paid to induce the respondents to take part in the surveys in the first place. In our experiment, the incentive is offered to participants on an unconditional basis. We are therefore able to isolate the effect of incentives on recording behaviour without having to worry about confounding effects due to participation choices induced by the incentive.

It is hard to imagine a real-life survey where an incentive similar to ours is paid to respondents during a survey of this kind. Yet, with behavioural experiments incorporating an incentive component becoming more popular in representative surveys, the question of whether these experiments have a (negative) influence on respondents' recording behaviour is certainly relevant.

3. Data and Variables

In this section we describe the Bundesbank's Payment Survey and the behavioural experiment we conducted in more detail. We also provide some information about respondent characteristics. Furthermore, we discuss various measures of respondents' commitment and data quality in a payment diary survey that might be affected by the incentives.

3.1. The Bundesbank's Payment Survey

In 2014, the Deutsche Bundesbank conducted the third wave of its payment behaviour survey entitled "Payment Behaviour in Germany" (see Deutsche Bundesbank 2015). The survey was carried out by the market research institute MARPLAN on behalf of the Bundesbank. The sample for the survey was drawn using a random-route procedure developed by the Association of German Market and Social Researchers, or ADM for short (see Hoffmeyer-Zlotnik 2003, on random route samples). The face-to-face

interviews with the respondents were conducted between May and July of 2014. The net sample comprises 2,036 persons. The survey is representative for the German population aged 18 and above. Care was taken to ensure that consumers from all 16 federal states were included in the gross sample.

The survey consists of two main parts, a CAPI interview and a drop-off paper and pencil diary. The diary was handed out to participants after completion of the interview. Respondents were also given the option of filling in the diary using a smartphone app, but less than two percent of respondents chose this option. The CAPI interviews took about 30 minutes on average and contained questions on topics such as the ownership and usage of payment instruments, cash withdrawal behaviour, perceived risks of payment instrument usage, and respondents' demographics.

The payment diary collected information on actual transactions over a period of seven days and specifically refers to direct payment transactions at the point-of-sale, that is, all transactions apart from recurring transactions, such as rent payments, insurance premiums, telephone and utility bills. The information collected in the paper and pencil diary includes the euro amount for each transaction, the location where the transaction took place (16 different possible locations including "retail purchases for day-to-day needs", "filling stations", "restaurants", "e-commerce", "payments to private individuals", etc.) and the payment medium used to settle the transaction (cash and a list of eleven cashless payment methods, e.g., debit cards, credit cards, e-payment schemes, payment schemes via mobile phone, contactless card payments). In addition, respondents had to provide information on cash withdrawals in the diary. The diary contains space for up to eight transactions for each day and some spare pages in case more than eight transactions were made in any one day. At the top of each page of the diary, the respondents were asked to fill in the date and then list all transactions pertaining to this date. The printed diary also contains a page with an example of how to fill in the diary, and the interviewers explained the transaction recording procedure to the respondents when they handed over the diary.

Interviews were conducted and diary data collected across the entire survey period and on all days of the week (including weekends and public holidays). Participants were instructed to start the diary on the day following the face-to-face interview. In practice, we see that more respondents started filling in their diaries on Wednesdays (19%) compared to Sundays (9%). To a certain extent, the unequal distribution of diary days might be caused by the fact that the probability of reaching a participant for an interview is higher on some days than on others. However, we cannot rule out that some participants did not closely follow the instructions and put off starting the diary until a day on which they had a transaction to report.

Respondents received participation incentives both for answering the survey and for filling in the diary. After completing the CAPI interview, the interviewers gave the respondents a pen and a notepad to help them make notes on their payment transactions during the day and a package of shredded banknotes as a participation incentive. A monetary incentive of EUR 10 was sent to everyone who answered the payment diary and returned it to the market research institute. All these incentives were paid out to respondents, irrespective of the outcome of the risk experiment. Therefore, they do not confound our analysis.

3.2. Behavioural Experiment

A novel feature of the 2014 survey was a behavioural experiment which is supposed to elicit respondents' risk preferences. This type of experiments is typically used in behavioural economics (see, for example, Dohmen et al. 2011; Dohmen et al. 2010; Eckel and Grossmann 2002; Charness et al. 2013; and Awel and Azomahou 2015). The aim of including the experiment in the survey was to have a measure for participants' risk preferences, which could then be linked to information provided in other parts of the survey, for example on the adoption of innovative payment instruments.

The risk experiment was performed directly after completing the face-to-face interview and carried out by the interviewers with all respondents. Respondents were told that they had the possibility to take part in a brief experiment which would take about five minutes and would not cost them anything. They were also told that they had the chance to win a cash prize. Out of the 2,036 persons completing the CAPI, 1,952 respondents (almost 96%) decided to take part in the risk experiment. The behavioural experiment was conducted using a representative sample of the population without considerable bias in the group of participants due to non-participation of respondents. This makes the experiment exceptionally valuable for behavioural economists who are usually constrained to conducting such experiments with a narrow subgroup of the population, such as university students (see, for example, Charness and Gneezy 2010).

The sequence of the behavioural experiment can be seen in Figure 1. Respondents who decided to take part in the experiment were given the choice between receiving a guaranteed payment of EUR 10 or the chance to participate in a lottery draw with an expected value of EUR 10 (50% chance of winning either EUR 20 or nothing at all). The rationale of the experiment is that risk-averse participants will choose the guaranteed payment while risk-loving participants will decide to take part in the lottery. This provides the economist with a rough measure of respondents' risk preference. Out of a total of 1,952 consumers participating in the risk experiment 994 took the guaranteed payment of EUR 10, and 958 consumers chose to participate in the lottery (see bottom line of Table 1).

To test whether respondents who took part in the lottery and those who chose the guaranteed payment differ in terms of sociodemographic characteristics, we run a series of t-tests. As expected, we find significant differences between the groups. Columns (II), (III), and (VI) in Table 1 reveal that lottery players are younger, more likely to be male and have a higher income on average (for literature on the link between sociodemographic characteristics and risk aversion see, for example, Eckel and Grossman 2008; Borghans et al. 2009; and Halek and Eisenhauer 2001). In addition, we use the CAPI interview to check whether the two groups differ in their self-assessed risk preferences, their technological literacy and their attitude towards new payment methods. Tables 1 and 2 in the Supplemental data present detailed results by group (available online at: http://dx.doi.org/10.1515/jos-2017-0021). Participants taking the guaranteed payment are significantly more risk-averse and prudent than those who chose to participate in the lottery, which supports the validity of the behavioural experiment. What is more, they are less technologically literate (i.e., less likely to use the internet and electronic devices) and are more conservative in their payment behaviour (i.e., less open to payment innovations, less likely to own a credit card or use e-payment schemes).

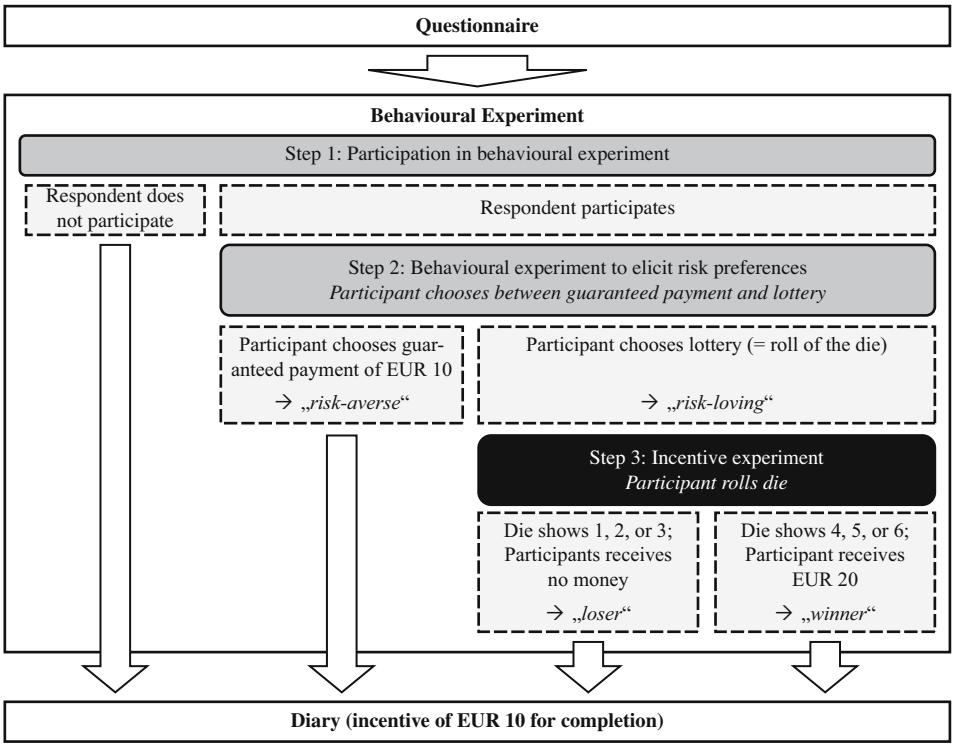

Fig. 1. Sequence of the behavioural experiment.

We also check for interviewer effects in the decision between the guaranteed payment and the lottery (see the lower part of Table 1). Respondents who had female and/or younger interviewers were more likely to choose the lottery option.

3.3. Incentive Experiment

Only the risk-loving participants choosing to participate in the lottery went on to participate in the final step of the risk experiment in which the lottery game was played and the payoffs were determined. The lottery game consisted of rolling a die. Participants either won EUR 20 if they threw a 4, 5, or a 6, or nothing if they threw a 1, 2, or a 3. This last stage of the risk experiment constitutes the incentive experiment, which is the main focus of our research. The payoff amounts resulting from the rolling of the die can be interpreted as an incentive payment. It should be noted, however, that the experiment was not specifically designed to test the effect of incentives on consumers' recording behaviour; instead, we "re-use" the second part of the risk experiment, which allows us to learn something about the effects of incentives on consumers' recording behaviour. In a pure incentive experiment, additional control groups would have been included in the experiment, for example consumers who chose the guaranteed payment would have had to role the die, etc. This limits the scope of our study. Nonetheless we are able to provide some insight into how recording behaviour and data quality is affected by including a standard behavioural risk experiment in a survey. The "assignment" of incentive payments

Table 1. Sociodemographic characteristics of respondents and interviewers by participation in and outcome of the behavioural experiment.

Variable	(I) All participants	(II) Guaranteed payment	(III) Lottery Total	(IV) Lottery Losers	(V) Lottery Winners	(VI) Difference guaranteed payment vs. lottery	(VII) Difference losers vs. winners
Age in years	46.79	47.90	45.64	46.54	45.03	2.25***	1.51
Female	0.55	0.59	0.51	0.50	0.52	0.08***	−0.03
Lives without partner	0.47	0.47	0.48	0.50	0.47	−0.02	0.03
Household size	2.22	2.19	2.25	2.25	2.24	−0.05	0.01
Education (4 groups)	2.05	2.02	2.10	2.08	2.11	−0.08*	−0.03
Household net income (12 groups)	5.04	4.90	5.18	5.30	5.09	−0.28**	0.21
Individual net income (12 groups)	3.45	3.28	3.63	3.72	3.57	−0.34***	0.15
East German	0.19	0.17	0.21	0.18	0.23	−0.04**	−0.05**
Interviewer female	0.52	0.49	0.55	0.53	0.57	−0.07***	−0.04
Age of interviewer	56.74	57.40	56.74	56.15	55.99	1.34**	0.15
Number of observations	1,952	994	958	391	567		

Notes: Number of observations differs for each variable as some values are missing.
*, **, *** mean difference is significant at the 90%, 95%, or 99% level (two-sided test).

Table 2. Results of estimations on the number of transactions and empty diaries.

Variable	(I) Total number of TA (count)	(II) Number of TA on day 1 (count)	(III) Number of TA on day 2 (count)	(IV) Empty diary on day 1 (dummy)	(V) Empty diary on day 2 (dummy)
	Negative Binomial			Probit	
WINNER	0.057* [0.032]	0.095* [0.051]	0.091* [0.055]	−0.229** [0.097]	−0.224** [0.094]
Control variables	Included (see Tables A1 and A5 for results and definitions)				
Observations	949	949	949	949	949
alpha	0.129 [0.011]	0.000 [0.000]	0.000 [0.011]		
Chi2	110.55	113.44	99.21	63.50	72.76
Pseudo-R^2				0.067	0.073

Notes: *, **, *** mean coefficient is significant at the 90%, 95%, or 99% level. Robust standard errors are given in brackets. TA stands for the number of transactions.

to consumers was obviously random if the die was non-biased and interviewers carried out the experiment correctly. However, the share of winners is about ten percentage points higher than expected, at almost 60%. Out of the 958 participants who rolled the die 567 won the EUR 20.

It cannot be ruled out that the interviewers deviated from the instructions and, for example, allowed the respondents to roll the die several times or simply paid out the EUR 20 regardless of the number on the die. If this were the case, this would mean a deviation from an experimental setting with randomly assigned outcomes. It cannot be checked *ex-post* why the realised and expected values do not match. However, in checks carried out by the survey agency after the interviews, respondents confirmed that the interviewers had actually offered the experiment and that they had correctly noted the result of rolling the die. Unfortunately, this does not constitute a thorough check of whether consumers were allowed to roll the die more than once. What is reassuring is that we do not find significant effects of interviewers' gender and age on the outcome of the roll of the die. As an additional check for the existence of interviewer effects, we re-calculated all results presented in Section 4, after eliminating cases from interviewers with very high winning rates. The results did not change.

To make sure that no bias with respect to observable sociodemographics exists between winners and losers, we again run a series of t-tests (for results see columns (IV), (V), and (VII) in Table 1). All but one come up negative, indicating that the composition of the two groups is very similar. We also run probit regressions with the outcome of the roll of the die (win/lose) as the dependent variable and sociodemographic variables as explanatory variables. They broadly support the results of the individual t-tests (see Table 3 in the Supplemental data online at http://dx.doi.org/10.1515/jos-2017-0021). In addition, we do not find any significant differences between winners and losers in their self-assessed risk preferences, their technological literacy and their attitude towards new payment methods as stated in the CAPI interview (see Tables 1 and 2 in the Supplemental data online at: http://dx.doi.org/10.1515/jos-2017-0021).

In summary, the randomisation of incentives worked well despite the fact that there are too many winners. We provide robustness tests below where we will show that our results prevail even after controlling for sociodemographics in a regression.

3.4. Measures of Respondents' Recording Behaviour and Data Quality

A key decision we had to take is along which dimensions we want to assess respondents' recording behaviour in payment diary surveys. An obvious first choice is to examine unit nonresponse and nonresponse bias (see, for example, Bonke and Fallesen 2010, 24). However, we do not consider these measures here as 951 out of the 958 participants who rolled the die returned a completed diary. Instead, we follow the literature cited above in Section 2 (see, for example, Davern et al. 2003; Shettle and Mooney 1999) and look, among other things, at measures related to item nonresponse. Item nonresponse can come in the form of a missing answer for an individual transaction or a missing transaction. We use the "share of incomplete transactions" as our measure of (classic) item nonresponse of the first type. A transaction is incomplete if any of the required information regarding the transaction is missing.

While classic item nonresponse is easy to detect and measure, transactions which are missing completely are harder to examine. Usually, no reference statistics are available which would allow the researcher to detect underreporting of transactions. A comparison of the total number of transactions or activities reported by consumers with incentives and those without has thus been used as an indirect measure (e.g., Fricker and Tourangeau 2010, Axhausen et al. 2002). We follow this literature and look at the total number of transactions reported for each day and their structure with respect to the payment method chosen (cash vs. non-cash payments) and transaction size (payment value below EUR 5). Our approach parallels Fricker and Tourangeau's (2010) classification of activities into different types, like basic daily activities and other activities.

Finally, we look at the share of rounded values (see, for example, Fricker and Tourangeau 2010). The share of rounded values should be an indicator of less commitment to the survey, as providing an exact figure can be bothersome for consumers.

4. Results

The results we present are based on those respondents who took part in the lottery game. The reasons for excluding those participants who opted for the guaranteed payment are discussed in Section 1. Out of the 958 persons who rolled the die, two losers and four winners had to be excluded because they did not return the payment diary. One winner was excluded because this person returned the diary without noting down any transactions. This leaves us with 951 observations on which we base our results. Due to the low number of cases excluded, we are not concerned about any nonresponse bias within the sample of risk-loving participants.

4.1. Number of Transactions

The first indicator we look at is the number of transactions. We find that winners record significantly more transactions, especially on the first two days and the last day of the diary period than respondents who did not win (see Figure 2). Throughout the seven-day diary week, winners noted down 10.40 transactions (1.49 transactions per day), while losers reported only 9.84 transactions (1.41 transactions per day). The number of reported transactions is highest on days 1, 2, and 3 for winners, and on days 1 and 3 for losers. This implies that both winners and losers show a strong initial commitment to the diary, but the level of commitment appears to be even higher for winners than for losers.

In order to rule out that sociodemographic variables or the day of the week confound the descriptive results presented above (see Figure 2), we regress the number of transactions recorded on the variable of interest "WINNER" and control variables. Table 2 provides a summary of those regressions (results for control variables reported in Table A1 in the Appendix). Columns (I), (II), and (III) of Table 2 confirm that winners report more transactions in total and on the first two diary days, even after controlling for consumers characteristics, such as age, gender, income and household size as well as the day of the week (where applicable). The estimated coefficient for "WINNER" of 0.057 in column (I) corresponds to a difference in the number of transactions between winners and losers of almost 0.6 (see also Table A2 in the Appendix). For the first and second day of the diary, the difference between winners and losers also remains significant at conventional levels.

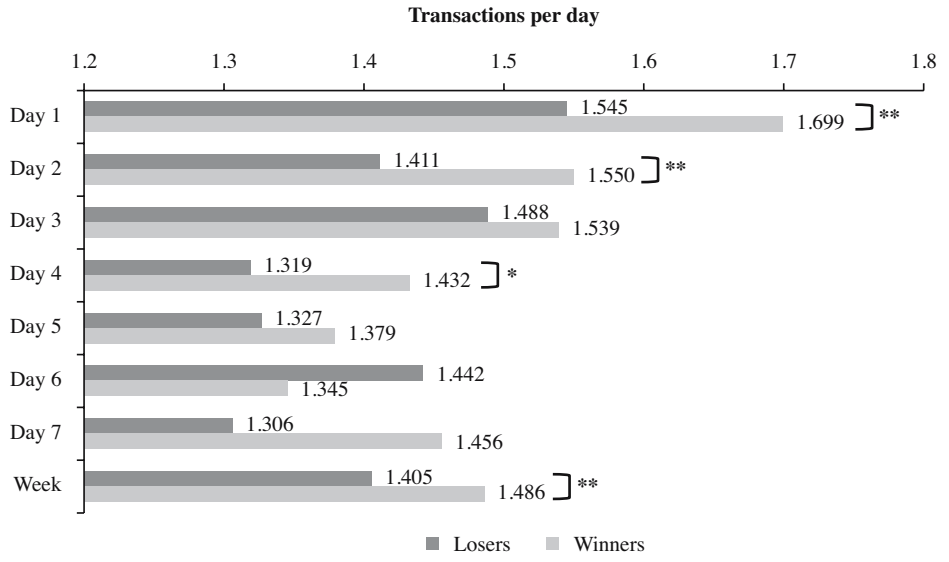

Fig. 2. Average number of transactions per person per diary day.
*Notes: *, **, *** mean difference is significant at the 90%, 95%, or 99% level (one-sided t-test).*

This result is mainly driven by the fact that the share of diaries without any transactions on days one and two is significantly lower for winners than for participants who did not win the EUR 20 (see columns (IV) and (V) in Table 2).

In order to be able to attribute the differences in the number of recorded transactions to an incentive effect, we have to rule out that the higher number of transactions for winners can be traced back to an income effect. In this case, winners would simply feel "richer" due to receiving the EUR 20 and therefore spend more and have more transactions than if they had not received the money. If the income effect existed, it should be larger for consumers with a low income than for those with a high income. Consequently, the difference in the number of transactions reported by winners and losers should be larger for respondents in the lowest income category compared with those in higher income classes. We test this assumption by including an interaction variable WINNER*INCOME in our regressions on the number of transactions reported and the share of empty diaries, which allows us to identify the effect of winning on the quantity of transactions reported in each income category. Figure 3 gives a graphical presentation of the estimated number of transactions in the diary week for winners and losers in different income categories (see also Tables 4 and 5 in the Supplemental data online http://dx.doi.org/10.1515/jos-2017-0021). Winners with a low income report only slightly more transactions than losers in the same income category. The difference is by no means statistically significant. In the middle income category, winners report substantially more transactions than losers (the difference is 1.2 transactions). A Wald test shows that the hypothesis of an equal number of transactions reported by winners and losers in this income group can be rejected at the 95% confidence level. Consequently, the higher number of transactions reported by winners in the estimation without interaction terms can mainly be traced back

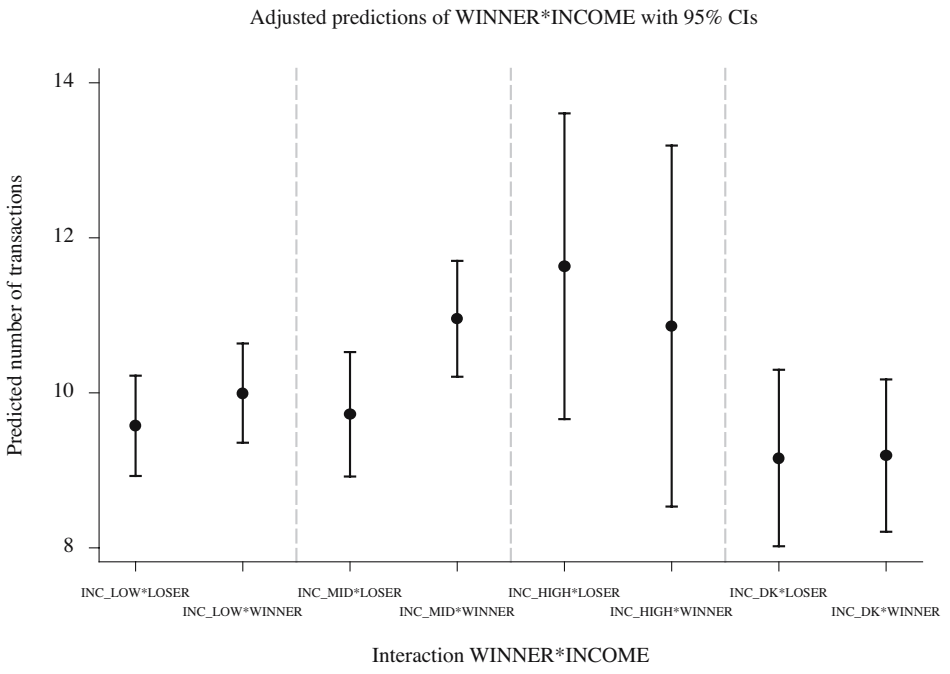

Fig. 3. Predictions of the number of transactions for winners and losers, by income category.

to middle income participants, not to those respondents for whom a potential income effect is expected to be the largest.

As an additional test we look at cash transactions. If winners immediately spent the EUR 20 they receive (in addition to their normal spending), their amount spent (in cash) during the diary reporting week should, all other things being equal, be higher than the amount spent by losers. Using robust regression techniques that down-weight outliers we show that there is no statistically significant difference between winners and losers in total spending and cash spending (see Table A3 in the Appendix). Although the estimated coefficients for winners are positive, they are far below EUR 20. Again, this leads us to conclude that income effects do not drive our results on the quantity of transactions reported in the payment diaries.

To sum up, we find that winners report more transactions than losers, in particular on the first and second diary day. These differences are neither driven by the sociodemographics of the respondents, nor by income effects.

A difference in the number of reported transactions does not affect the substantial outcomes of the survey or response quality unless the additional transactions reported differ systematically from the other transactions. To shed some light on this issue we analyse the structure of payments for each day for the two groups of consumers below.

4.2. Type and Size of Transactions

To investigate whether winners report specific or special payments on days one and two, we first look at cash transactions as a specific type of transaction. Days one and two exhibit

the highest average number of cash transactions for winners (see Figure A1 in the Appendix). However, day one also sticks out for losers. Since cash transactions follow a similar pattern to the total number of transactions described in Subsection 4.1, looking at cash shares is more informative. The cash share is computed by first calculating the share of cash transactions for each individual and diary day and then taking the average across all individuals. Figure 4 shows that the difference with respect to the cash share between those who win and those who lose is insignificant on each of the seven diary days as well as for the whole diary week. Moreover, it is sometimes positive and sometimes negative, and alternates for the first two days.

In addition, we do not find significant differences in the reported cash shares between day one and all other days of the diary week – neither for winners nor for losers. Looking at these factors together, there is no evidence of a disproportional reporting of cash payments on days one and two for winners.

Another classic categorisation of transactions is by their size (see, for example, Bagnall et al. 2016). There is evidence that small-value transactions (below EUR 5) are underreported in diary surveys (Jonker and Kosse 2013). A high number of small-value transactions and, in particular, a high share of this type of transaction among all transactions are signs of good data quality and would suggest a strong commitment to the survey. In our payment diary, the share of transactions below EUR 5 does not differ significantly between winners and losers (see Table 3). The differences observed between winners and losers are insignificant overall and on all individual days. Similarly, the differences in average transaction amounts between winners and losers are insignificant for each diary day and for all days taken together (see Table 3).

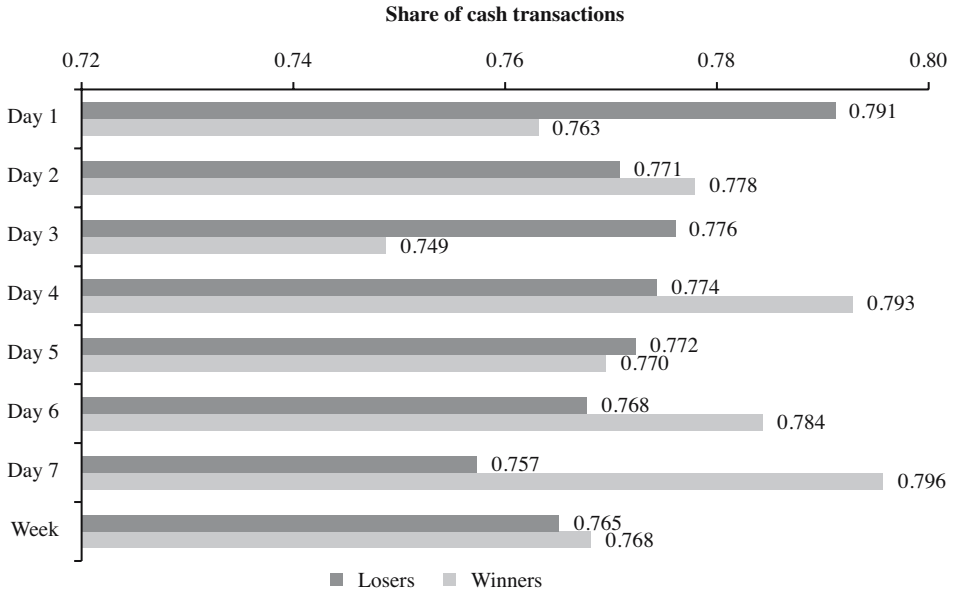

Fig. 4. Average share of cash transactions per person per day.
*Notes: Calculated for individuals with more than zero transactions on a given day. *, **, *** mean difference significant at the 90%, 95%, or 99% level (two-sided t-test).*

Table 3. Average share of small-value transactions below EUR 5 and average transaction amount (in EUR) per person per diary day.

	Share of small-value TA			Average TA amount		
	Losers	Winners	Difference losers vs. winners	Losers	Winners	Difference losers vs. winners
Day 1	0.247	0.226	−0.021	28.48	24.51	−3.97
Day 2	0.217	0.225	0.009	25.29	25.16	−0.13
Day 3	0.195	0.222	0.027	25.53	26.97	1.44
Day 4	0.225	0.210	−0.015	25.90	24.07	−1.83
Day 5	0.195	0.207	0.013	24.22	23.98	−0.24
Day 6	0.194	0.239	0.046	23.99	23.72	−0.27
Day 7	0.216	0.221	0.005	26.42	29.28	2.86
Diary week	0.206	0.222	0.016	26.08	26.98	0.90

Notes: Calculated for individuals with more than zero transactions on a given day. *, **, *** mean difference significant at the 90%, 95%, or 99% level (two-sided t-test). TA is short for transactions.

In order to make sure that differences between winners and losers in the type and size of payments reported are not driven by sociodemographic factors, we run a series of regressions with payment type and payment size as dependent variables, the outcome of the roll of the die (winner or loser) as an independent variable and various sociodemographics as control variables. The regression analysis confirms the descriptive results: being a winner does not have a significant effect on the share of cash payments, the share of small-value transactions and the average transaction amount (see columns (I) to (III) in Table A4 in the Appendix).

All these results indicate that the higher quantity of transactions reported on the first days does not induce a bias in substantive diary results.

4.3. Incomplete Transaction Data and Rounding

In the previous sections we looked at item nonresponse with respect to completely missing transactions. In this section we focus on the classic item nonresponse measures, that is missing information on recorded transactions. In the payment diary, respondents were asked to answer several questions concerning each individual transaction. Besides stating the amount, the payment instrument used and the location where the transaction took place, respondents were also asked about various circumstances that might have influenced their payment choice. Up to eight variables pertain to one transaction. Given that all the requested information requires the respondent to read the diary carefully and to remember the actual transaction precisely, this can be bothersome for some respondents. Similarly, reporting precise amounts in the diary can be a burden for some consumers. They have to remember the exact amount or keep the receipt, for example. Thus, many people round the reported transaction values in payment diaries. In the 2014 Bundesbank payment diary, we see that, on average, about one quarter (29%) of all transactions per person per day are rounded to the next whole euro.

Table 4. Average share of incomplete transactions and average share of rounded transaction values per person per diary day.

	Share of incomplete TA			Share of rounded TA values		
	Losers	Winners	Difference losers vs. winners	Losers	Winners	Difference losers vs. winners
Day 1	0.319	0.282	−0.036	0.237	0.258	0.022
Day 2	0.311	0.322	0.012	0.278	0.267	−0.011
Day 3	0.379	0.329	−0.050	0.294	0.316	0.023
Day 4	0.376	0.342	−0.034	0.271	0.306	0.035
Day 5	0.385	0.345	−0.039	0.356	0.292	−0.064**
Day 6	0.347	0.386	0.039	0.306	0.290	−0.016
Day 7	0.423	0.376	−0.047	0.284	0.302	0.018
Diary week	0.368	0.337	−0.031	0.287	0.293	0.006

Notes: Calculated for individuals with more than zero transactions on a given day.
*, **, *** mean difference is significant at the 90%, 95%, or 99% level (two-sided t-test). TA is short for transactions.

Table 4 shows that, on average, around one in three transactions contains missing information. There are slightly fewer incomplete transactions for winners (33.7%) than for losers (36.8%). Both winners and losers report more incomplete transactions as the diary week progresses. However, the shares of incomplete transactions do not differ significantly overall or for any of the individual days of the diary between winners and losers.

The picture looks similar for the rounding of values. Table 4 shows that the rate of rounded transaction values is only marginally higher for consumers who win the EUR 20. Consumers winning and those losing in the experiment both show significantly increasing rates of rounded values as the diary progresses, under-scoring the value of this measure as a quality indicator. Overall, the difference between the two groups with respect to rounded values is not significant, with the exception of day 5.

Again, regressions with our two quality indicators as dependent variables, the outcome of the lottery (winner or loser) as an independent variable and sociodemographics as control variables confirm that being a winner does not have a significant effect on the share of incompletely reported transactions and rounded transaction values (see columns (IV) and (V) of Table A4 in the Appendix).

5. Conclusions

In this article, we analyse the effect of an incentive experiment on German consumers' recording behaviour in a one-week diary based on their point-of-sale expenditure. As part of the experiment, participants were asked to roll a die and had the chance of winning EUR 20 depending on the number thrown on the die. We interpret this as a random assignment of a monetary incentive, where, in contrast to most other incentive experiments, the persons receiving the incentive (winner) know that they received a higher incentive than some of the other participants. The experiment itself could thus stimulate a positive sentiment in winners and a negative sentiment in losers towards the survey and the

diary in general. Consequently, we expect winners to exhibit a greater commitment to the diary than losers, and thus a differential reporting behaviour.

We measure respondents' recording behaviour using several indicators: the number of transactions recorded, the share of cash transactions, the share of low-value transactions (below EUR 5), the share of rounded transactions values, and the share of transactions with incomplete information.

We find that the overall impact of paying an unconditional incentive during an interview or different parts of an interview on the recording behaviour is rather limited. Our results indicate that the outcome of the game has an impact on the quantity of transactions recorded, but does not affect other aspects of respondents' recording behaviour, such as item nonresponse or rounding. It also has a negligible impact on substantive measures, such as the cash share. The low variation between the two groups with respect to the cash share and the share of small-value transactions indicates that the key findings from the diary, such as the overall cash share of point-of-sale transactions and the share of transactions within certain value ranges, are not biased by the outcome of the experiment.

Our results have implications beyond research on incentive effects. To the best of our knowledge, we are the first to look into the effects of behavioural or psychological experiments with monetary rewards carried out during a representative national survey. The incentive experiment we examine was an integral part of a behavioural experiment eliciting respondents' risk inclination, which was performed on a representative sample of the adult population. Incorporating behavioural experiments in such surveys is a rather new and promising approach considering that these experiments were previously often carried out with a non-representative part of the population (e.g., students). Up to now, little is known about how such experiments influence participants' attitude towards the survey to which the experiment is linked. Our research indicates that including an experiment does not negatively affect the survey's data quality. Survey designers can thus include these types of experiments without having to worry about increasing the total survey error of their study.

We were mainly concerned with the incentive effects of the behavioural experiment, both monetary and non-monetary. Whether other features of the experiment also play a role for respondents' reporting behaviour is still an open issue. It is feasible to assume that the design and administration of the experiment by the interviewers could potentially confound and even counteract the incentive effects. More research on these issues is necessary.

Appendix

Table A1. Results of estimations on the number of transactions and empty diaries.

Variable	(I) Total number of TA (count)	(II) Number of TA on day 1 (count)	(III) Number of TA on day 2 (count)	(IV) Empty diary on day 1 (dummy)	(V) Empty diary on day 2 (dummy)
	Negative Binomial			Probit	
WINNER	0.057* [0.032]	0.095* [0.051]	0.091* [0.055]	−0.229** [0.097]	−0.224** [0.094]
AGE_25_34	0.051 [0.061]	−0.045 [0.111]	0.124 [0.126]	−0.035 [0.221]	0.006 [0.208]
AGE_35_44	0.232*** [0.064]	0.223** [0.110]	0.274** [0.124]	−0.166 [0.228]	−0.159 [0.212]
AGE_45_54	0.205*** [0.062]	0.127 [0.109]	0.211* [0.124]	−0.075 [0.224]	0.051 [0.205]
AGE_55_64	0.193*** [0.069]	0.124 [0.118]	0.246* [0.136]	−0.110 [0.236]	−0.051 [0.224]
AGE_65+	0.208** [0.093]	0.160 [0.159]	0.219 [0.166]	0.126 [0.289]	−0.227 [0.271]
FEMALE	0.022 [0.033]	0.004 [0.053]	−0.002 [0.056]	0.036 [0.103]	−0.001 [0.099]
HH_SIZE_2	0.031 [0.040]	0.135** [0.062]	−0.015 [0.069]	0.007 [0.120]	−0.062 [0.115]
HH_SIZE_3	−0.044 [0.045]	0.019 [0.078]	−0.029 [0.080]	0.089 [0.151]	−0.211 [0.149]
HH_SIZE_4+	0.010 [0.048]	0.127* [0.074]	0.012 [0.082]	0.012 [0.156]	−0.079 [0.146]
EDU_MEDIUM	0.097** [0.041]	0.032 [0.063]	0.105 [0.067]	0.128 [0.122]	−0.155 [0.113]
EDU_HIGH	0.247*** [0.048]	0.155* [0.080]	0.204** [0.082]	0.147 [0.154]	−0.296** [0.147]
EDU_UNI	0.269*** [0.057]	0.213** [0.092]	0.234** [0.094]	0.153 [0.170]	−0.121 [0.169]
INC_MID	0.061 [0.039]	0.086 [0.060]	0.042 [0.071]	−0.169 [0.123]	−0.076 [0.115]
INC_HIGH	0.140* [0.076]	−0.031 [0.113]	0.283** [0.112]	−0.256 [0.225]	−0.468* [0.243]
INC_DK	−0.070 [0.049]	−0.064 [0.081]	−0.009 [0.089]	−0.092 [0.162]	−0.101 [0.160]
REGION_EAST	0.007 [0.039]	−0.037 [0.064]	0.188*** [0.066]	0.069 [0.120]	−0.091 [0.118]

Table A1. Continued

Variable	(I) Total number of TA (count)	(II) Number of TA on day 1 (count)	(III) Number of TA on day 2 (count)	(IV) Empty diary on day 1 (dummy)	(V) Empty diary on day 2 (dummy)
	Negative Binomial			Probit	
OCC_TRAIN	0.046 [0.089]	0.162 [0.150]	−0.003 [0.171]	−0.174 [0.299]	0.175 [0.273]
OCC_WORK	0.047 [0.058]	0.195** [0.090]	0.002 [0.100]	−0.173 [0.171]	−0.110 [0.163]
OCC_OTHER	−0.041 [0.079]	0.069 [0.122]	−0.022 [0.132]	0.265 [0.214]	0.017 [0.209]
SUNDAY	n.a.	−0.683*** [0.123]	−0.644*** [0.124]	0.461** [0.185]	0.987*** [0.199]
TUESDAY	n.a.	−0.030 [0.098]	−0.254** [0.105]	−0.037 [0.178]	0.488** [0.201]
WEDNESDAY	n.a.	0.189** [0.083]	−0.164 [0.103]	−0.530*** [0.174]	0.308 [0.203]
THURSDAY	n.a.	0.036 [0.093]	−0.040 [0.096]	−0.218 [0.175]	0.401** [0.190]
FRIDAY	n.a.	0.089 [0.096]	0.055 [0.091]	−0.449** [0.193]	−0.083 [0.208]
SATURDAY	n.a.	0.110 [0.097]	−0.023 [0.106]	−0.142 [0.184]	0.266 [0.208]
CONSTANT	1.933*** [0.083]	0.009 [0.156]	0.092 [0.173]	−0.496 [0.309]	−0.584* [0.315]
Observations	949	949	949	949	949
alpha	0.129 [0.011]	0.000 [0.000]	0.000 [0.011]		
Chi2	110.55	113.44	99.21	63.50	72.76
Pseudo-R^2				0.067	0.073

Notes: *, **, *** mean coefficient is significant at the 90%, 95% or 99% level. Robust standard errors are given in brackets. TA is short for transactions.

Table A2. Predicted number of events for winners and losers.

	Winner			Loser		
	Estimate	Standard error	95% Confidence interval	Estimate	Standard error	95% Confidence interval
Total number of TA (count)	10.278	0.219	[9.850; 10.707]	9.712	0.227	[9.267; 10.158]
Number of TA on day 1 (count)	1.637	0.052	[1.534; 1.739]	1.489	0.060	[1.371; 1.606]
Number of TA on day 2 (count)	1.493	0.052	[1.391; 1.595]	1.363	0.059	[1.247; 1.479]
Empty diary on day 1 (dummy)	0.159	0.016	[0.128; 0.189]	0.221	0.022	[0.178; 0.264]
Empty diary on day 2 (dummy)	0.194	0.017	[0.161; 0.228]	0.262	0.023	[0.217; 0.307]

Notes: Effects are estimated at the mean of other regressor variables. TA is the number of transactions.

Table A3. Results of estimations on the total amount spent and cash spent during the week.

Variable	(I) Amount spent during the week	(II) Amount spent during the week	(III) Cash spent during the week	(IV) Cash spent during the week
	Robust regression	Median regression	Robust regression	Median regression
WINNER	7.577 [8.545]	6.165 [11.267]	4.885 [5.093]	3.254 [6.525]
AGE_25_34	35.377* [19.717]	40.220* [23.229]	22.607* [11.751]	27.625*** [10.308]
AGE_35_44	44.650** [20.169]	53.603** [23.732]	30.675** [12.021]	39.167*** [10.420]
AGE_45_54	65.811*** [19.602]	75.323*** [24.665]	43.954*** [11.683]	49.040*** [10.760]
AGE_55_64	72.762*** [20.964]	90.105*** [29.363]	49.914*** [12.494]	57.896*** [13.322]
AGE_65+	65.277** [25.480]	78.520** [32.524]	51.374*** [15.186]	63.344*** [20.820]
FEMALE	15.948* [8.931]	5.238 [10.968]	8.052 [5.323]	13.623** [6.754]
HH_SIZE_2	23.566** [10.584]	37.773*** [14.456]	6.358 [6.308]	6.594 [8.274]
HH_SIZE_3	22.915* [13.080]	18.883 [14.365]	0.355 [7.796]	4.181 [8.803]
HH_SIZE_4+	63.653*** [13.212]	65.488*** [18.954]	13.985* [7.875]	15.518 [11.564]
EDU_MEDIUM	17.525* [10.461]	16.515 [11.609]	2.071 [6.235]	3.371 [7.866]
EDU_HIGH	34.942*** [13.200]	42.575** [18.413]	0.869 [7.867]	1.599 [9.562]
EDU_UNI	24.665* [14.879]	44.655* [23.233]	− 11.994 [8.868]	− 16.198 [14.304]
INC_MID	51.947*** [10.545]	65.548*** [14.344]	17.137*** [6.285]	18.461** [8.203]
INC_HIGH	60.137*** [19.076]	103.160*** [35.659]	20.552* [11.370]	29.417* [16.961]
INC_DK	6.001 [14.159]	4.950 [19.316]	− 2.832 [8.439]	− 9.530 [9.956]
REGION_EAST	− 5.813 [10.623]	− 7.188 [12.260]	− 3.208 [6.332]	− 11.125 [6.941]
OCC_TRAIN	− 26.970 [26.068]	− 17.163 [33.842]	− 14.440 [15.536]	− 14.494 [20.049]
OCC_WORK	17.757 [14.953]	21.403 [21.428]	− 4.688 [8.912]	− 1.848 [17.771]
OCC_OTHER	− 41.508** [19.566]	− 22.648 [21.224]	− 19.830* [11.662]	− 22.173 [20.399]
CONSTANT	72.786*** [24.876]	48.143 [30.731]	61.947* [14.826]	49.489** [20.091]
Observations	949	949	949	949
Pseudo-R^2		0.072		0.042

Notes: Median regression with bootstrapped standard errors (1,000 repetitions).

Table A4. Results of estimations on various transaction data features.

Variable	(I) Share of cash TA	(II) Share of small TA	(III) Average TA amount	(IV) Share of incomplete TA	(V) Share of rounded TA
	OLS	OLS	OLS	OLS	OLS
WINNER	0.006 [0.015]	0.012 [0.014]	1.258 [2.189]	−0.022 [0.023]	0.006 [0.014]
AGE_25_34	0.007 [0.034]	−0.063** [0.030]	0.964 [3.639]	−0.002 [0.049]	−0.028 [0.034]
AGE_35_44	0.024 [0.034]	−0.043 [0.031]	1.310 [2.915]	−0.026 [0.050]	−0.046 [0.034]
AGE_45_54	0.002 [0.034]	−0.088*** [0.031]	4.727 [3.250]	0.026 [0.049]	−0.026 [0.033]
AGE_55_64	0.033 [0.036]	−0.105*** [0.033]	6.893 [4.307]	0.078 [0.054]	−0.040 [0.035]
AGE_65+	0.042 [0.044]	−0.096** [0.039]	4.340 [5.255]	0.108* [0.065]	−0.020 [0.041]
FEMALE	0.026* [0.016]	−0.016 [0.015]	−6.283** [3.164]	−0.001 [0.023]	−0.035** [0.015]
HH_SIZE_2	−0.050*** [0.018]	−0.029* [0.017]	8.825** [4.270]	0.024 [0.028]	0.021 [0.017]
HH_SIZE_3	−0.043* [0.023]	0.006 [0.022]	4.171** [1.826]	0.033 [0.034]	0.044** [0.022]
HH_SIZE_4+	−0.078*** [0.023]	−0.062*** [0.020]	8.172*** [2.724]	0.091*** [0.035]	0.079*** [0.022]
EDU_MEDIUM	−0.020 [0.017]	0.032* [0.017]	2.503 [1.716]	−0.080*** [0.029]	−0.001 [0.018]
EDU_HIGH	−0.071*** [0.024]	0.027 [0.023]	3.760 [5.173]	−0.108*** [0.034]	0.007 [0.022]
EDU_UNI	−0.076*** [0.027]	0.044* [0.022]	1.710 [3.331]	−0.118*** [0.039]	−0.030 [0.022]
INC_MID	−0.020 [0.019]	−0.037** [0.017]	2.941 [2.185]	0.015 [0.027]	0.033** [0.017]
INC_HIGH	−0.012 [0.033]	0.002 [0.032]	2.216 [3.473]	0.005 [0.052]	0.003 [0.030]
INC_DK	−0.037 [0.026]	0.021 [0.023]	4.410 [3.461]	0.010 [0.038]	−0.011 [0.023]
REGION_EAST	−0.001 [0.018]	0.027 [0.017]	−2.783 [3.091]	−0.115*** [0.025]	−0.016 [0.017]
OCC_TRAIN	−0.014 [0.045]	0.065 [0.045]	6.136 [13.135]	−0.008 [0.065]	0.050 [0.042]
OCC_WORK	−0.036 [0.026]	−0.001 [0.021]	−1.923 [3.351]	0.046 [0.039]	0.023 [0.022]
OCC_OTHER	0.050 [0.033]	0.045 [0.033]	−7.881** [3.753]	0.116** [0.054]	−0.004 [0.035]
CONSTANT	0.831*** [0.042]	0.280*** [0.038]	20.154*** [4.926]	0.347*** [0.063]	0.284*** [0.041]
Observations	949	949	949	949	949
R^2	0.081	0.077	0.030	0.066	0.049

Notes: *, **, *** mean coefficient is significant at the 90%, 95% or 99% level. Robust standard errors are given in brackets. TA is short for transactions.

Table A5. Construction of regression variables.

Variable name	Type	Description
WINNER	Dummy	Outcome of roll of the die. One, if person wins EUR 20.
AGE_18_24	Dummy	One, if respondent is aged 18 to 24. Reference category.
AGE_25_34	Dummy	One, if respondent is aged 25 to 34.
AGE_35_44	Dummy	One, if respondent is aged 35 to 44.
AGE_45_54	Dummy	One, if respondent is aged 45 to 54.
AGE_55_64	Dummy	One, if respondent is aged 55 to 64.
AGE_65+	Dummy	One, if respondent is aged 65 or above.
FEMALE	Dummy	Gender of respondent. One, if gender is female.
HH_SIZE_1	Dummy	Number of persons living in respondent's household (including children). One, if household size is one. Reference category.
HH_SIZE_2	Dummy	Number of persons living in respondent's household (including children). One, if household size is two.
HH_SIZE_3	Dummy	Number of persons living in respondent's household (including children). One, if household size is three.
HH_SIZE_4+	Dummy	Number of persons living in respondent's household (including children). One, if household size is four or more.
EDU_LOW	Dummy	Educational attainment of respondent. One, if education is low (no educational degree (yet), lower secondary education of less than 10 years) or if education is "other/don't know". Reference category.
EDU_MEDIUM	Dummy	Educational attainment of respondent. One, if respondent has secondary education of at least 10 years.
EDU_HIGH	Dummy	Educational attainment of respondent. One, if respondent has upper secondary education (= qualification for entering university).
EDU_UNI	Dummy	Educational attainment of respondent. One, if respondent has university degree (including university of applied sciences).
INC_LOW	Dummy	Respondent's personal monthly net income. One, if income is less than EUR 1,500. Reference category.
INC_MID	Dummy	Respondent's personal monthly net income. One, if income is between EUR 1,500 and EUR 3,000.
INC_HIGH	Dummy	Respondent's personal monthly net income. One, if income is more than EUR 3,000.
INC_DK	Dummy	Respondent's personal monthly net income. One, if "don't know/no answer".
REGION_EAST	Dummy	Respondent's region of residence. One, if region is East Germany.

Table A5. Continued

Variable name	Type	Description
OCC_HOME	Dummy	Respondent's current occupation. One, if respondent is not working or working at home (pensioner, homemaker). Reference category.
OCC_TRAIN	Dummy	Respondent's current occupation. One, if respondent is in training (student, apprentice, volunteer in federal volunteer service ("Bundesfreiwilligendienst")).
OCC_WORK	Dummy	Respondent's current occupation. One, if respondent is working outside the home (employee, public servant, self-employed person).
OCC_OTHER	Dummy	Respondent's current occupation. One, if unemployed or "other/don't know".
SUNDAY	Dummy	One, if transaction takes place on Sunday.
MONDAY	Dummy	One, if transaction takes place on Monday. Reference category.
TUESDAY	Dummy	One, if transaction takes place on Tuesday.
WEDNESDAY	Dummy	One, if transaction takes place on Wednesday.
THURSDAY	Dummy	One, if transaction takes place on Thursday.
FRIDAY	Dummy	One, if transaction takes place on Friday.
SATURDAY	Dummy	One, if transaction takes place on Saturday.

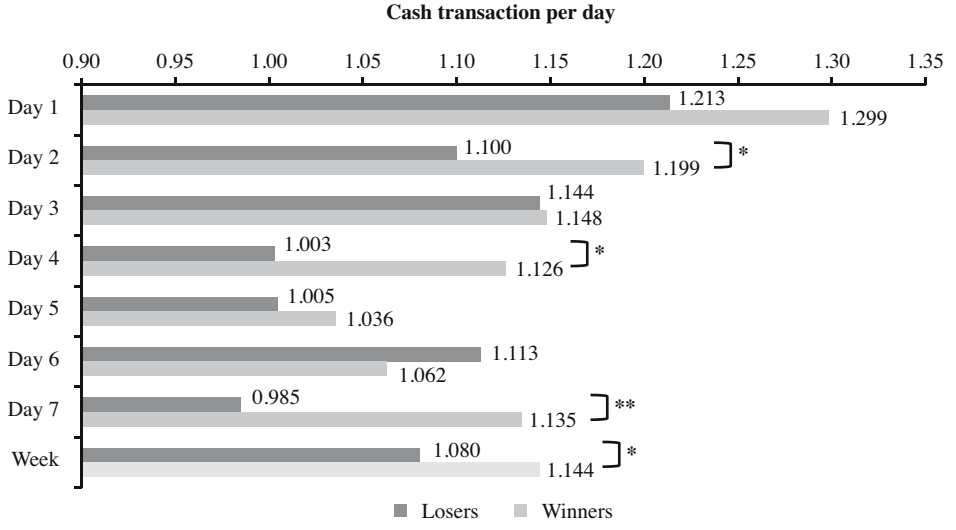

Fig. A1. Number of cash transactions per person per diary day.

6. References

Awel, Y.M. and T.T. Azamahou. 2015. "Risk Preference or Financial Literacy? Behavioural Experiment on Index Insurance Demand." UNU-MERIT Working Papers 2015-005, Maastricht. Available at: http://www.merit.unu.edu/publications/working-papers/abstract/?id=5626 (accessed August 2016).

Axhausen, K.W., A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt. 2002. "Observing the Rhythms of Daily Life: A Six-Week Travel Diary." *Transportation* 29(2): 95–124. Doi: http://dx.doi.org/10.1023/A:1014247822322.

Bagnall, J., D. Bounie, K.P. Huynh, A. Kosse, T. Schmidt, S.D. Schuh, and H. Stix. 2016. "Consumer Cash Usage: A Cross-Country Comparison with Payment Diary Survey Data." *International Journal of Central Banking* 12(4): 1–62. Available at: http://www.ijcb.org/journal/ijcb16q4a1.htm (accessed December 2016).

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. New Jersey: John Wiley & Sons.

Bonke, J. and P. Fallesen. 2010. "The Impact of Incentives and Interview Methods on Response Quantity and Quality in Diary- and Booklet-Based Surveys." *Survey Research Methods* 4(2): 91–101. Doi: http://dx.doi.org/10.18148/srm/2010.v4i2.3614.

Borghans, L., J.J. Heckman, B.H. Golsteyn, and H. Meijers. 2009. "Gender Differences in Risk Aversion and Ambiguity Aversion." *Journal of the European Economic Association* 7(2–3): 649–658. Doi: http://dx.doi.org/10.1007/s11238-016-9565-9.

Charness, G. and U. Gneezy. 2010. "Portfolio Choice and Risk Attitudes: An Experiment." *Economic Inquiry* 48(1): 133–146. Doi: http://dx.doi.org/10.1111/j.1465-7295.2009.00219.x.

Charness, G., U. Gneezy, and A. Imas. 2013. "Experimental Methods: Eliciting Risk Preferences." *Journal of Economic Behavior & Organization* 87: 43–51. Doi: http://dx.doi.org/10.1016/j.jebo.2012.12.023.

Davern, M., T.H. Rockwood, R. Sherrod, and S. Campbell. 2003. "Prepaid Monetary Incentives and Data Quality in Face-to-Face Interviews: Data from the 1996 Survey of Income and Program Participation Incentive Experiment." *Public Opinion Quarterly* 67(1): 139–147. Doi: http://dx.doi.org/10.1086/346012.

Deutsche Bundesbank. 2015. *Payment Behaviour in Germany in 2014. Third Study on the Utilisation of Cash and Cashless Payment Instruments*. Frankfurt am Main: Deutsche Bundesbank. Available at: http://www.bundesbank.de/Redaktion/EN/Downloads/Publications/Studies/payment_behaviour_in_germany_in_2014.pdf?__blob=publicationFile (accessed December 2015).

Dohmen, T., A. Falk, D. Huffman, and U. Sunde. 2010. "Are Risk Aversion and Impatience Related to Cognitive Ability?" *The American Economic Review* 100(3): 1238–1260. Doi: http://dx.doi.org/10.1257/aer.100.3.1238.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G.G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants and Behavioral Consequences." *Journal of the European Economic Association* 9(3): 522–550. Doi: http://dx.doi.org/10.1111/j.1542-4774.2011.01015.x.

Eckel, C.C. and P.J. Grossman. 2002. "Sex Differences and Statistical Stereotyping in Attitudes Toward Financial Risk." *Evolution and Human Behavior* 23(4): 281–295. Doi: http://dx.doi.org/10.1016/S1090-5138(02)00097-1.

Eckel, C.C. and P.J. Grossman. 2008. "Men, Women and Risk Aversion: Experimental Evidence." In *Handbook of Experimental Economics Results* 1(113), edited by C.R. Plott and V.L. Smith. 1061–1073. New York: Elsevier.

Fricker, S. and R. Tourangeau. 2010. "Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys." *Public Opinion Quarterly* 74(5): 934–955. Doi: http://dx.doi.org/10.1093/poq/nfq064.

Godwin, R.K. 1979. "The Consequences of Large Monetary Incentives in Mail Surveys of Elites." *Public Opinion Quarterly* 43(3): 378–387. Doi: http://dx.doi.org/10.1086/268528.

Goetz, E.G., T.R. Tyler, and F.L. Cook. 1984. "Promised Incentives in Media Research: A Look at Data Quality, Sample Representativeness and Response Rates." *Journal of Marketing Research* 21(2): 148–154. Doi: http://dx.doi.org/10.2307/3151697.

Goldenberg, K., D. McGrath, and L. Tan. 2009. "The Effects of Incentives on the Consumer Expenditure Interview Survey." Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, May 14 – 17, 2009, Hollywood, Florida. Available at: https://stats.bls.gov/osmr/pdf/st090100.pdf (accessed December 2015).

Goldenberg, K. and J. Ryan. 2009. "Evolution and Change in the Consumer Expenditure Surveys: Adapting Methodologies to Meet Changing Needs." Paper presented at the NBER Summer Institute 2009 – Conference on Research on Income and Wealth, July 13, 2009, Cambridge, Massachusetts. Available at: https://www.bls.gov/cex/research_papers/pdf/Evolution-and-Change-in-the-Consumer-Expenditure-Surveys-Adapting-Methodologies-to-Meet-Changing-Needs.pdf (accessed December 2015).

Göritz, A.S. 2004. "The Impact of Material Incentives on Response Quantity, Response Quality, Sample Composition, Survey Outcome, and Cost in Online Access Panels." *International Journal of Market Research* 46(3): 327–345. Available at: https://www.mrs.org.uk/ijmr_article/article/79180 (accessed January 2016).

Göritz, A.S. 2005. "Contingent versus Unconditional Incentives in WWW-Studies." *Metodolosky Zvezki* 2(1): 1–14. Available at: http://mrvar.fdv.uni-lj.si/pub/mz/mz2.1/goritz.pdf (accessed January 2016).

Groves, R.M., R.B. Cialdini and M.P. Couper. 1992. "Understanding the Decision to Participate in a Survey." *Public Opinion Quarterly* 56(4): 475–495. Doi: http://dx.doi.org/10.1086/269338.

Groves, R.M. 2004. *Survey Errors and Survey Costs*. New Jersey: John Wiley & Sons.

Groves, R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5): 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065.

Halek, M. and J.G. Eisenhauer. 2001. "Demography of Risk Aversion." *Journal of Risk and Insurance* 68(1): 1–24. Doi: http://dx.doi.org/10.2307/2678130.

Hedlin, D., T. Dale, G. Haraldsen, and J. Jones (eds.). 2005. "Developing Methods for Assessing Perceived Response Burden. Research Report." Stockholm: Statistics Sweden, Oslo: Statistics Norway, and London: Office for National Statistics. Available at: http://ec.europa.eu/eurostat/documents/64157/4374310/10-DEVELOPING-METHODS-FOR-

ASSESSING-PERCEIVED-RESPONSE-BURDEN.pdf/1900efc8-1a07-4482-b3c9-be88ee71df3b (accessed November 2015).

Hoffmeyer-Zlotnik, J.H. 2003. "New Sampling Designs and the Quality of Data." In *Developments in Applied Statistics.* Methodoloski zvezki 19, edited by A. Ferligoj and A. Mrvar. 205–217. Ljubljana: FDV.

James, J.M. and R. Bolstein. 1990. "The Effect of Monetary Incentives and Follow-Up Mailings on the Response Rate and Response Quality in Mail Surveys." *Public Opinion Quarterly* 54(3): 346–361. Doi: http://dx.doi.org/10.1086/269211.

Jones, J. 2012. "Response Burden: Introductory Overview Lecture." Paper presented at the 4th International Establishment Surveys Conference, June 11–14, 2012, Montreal, Canada. Available at: https://ww2.amstat.org/meetings/ices/2012/papers/302289.pdf (accessed July 2016).

Jonker, N. and A. Kosse. 2013. "Estimating Cash Usage: The Impact of Survey Design on Research Outcomes." *De Economist* 161(1): 19–44. Doi: http://dx.doi.org/10.1007/s10645-012-9200-2.

Laurie, H. and P. Lynn. 2009. "The Use of Respondent Incentives on Longitudinal Surveys." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 205–234. Chichester, UK: John Wiley & Sons.

Schwarz, N. and G.L. Clore. 1996. "Feelings and Phenomenal Experiences." In *Social Psychology: Handbook of Basic Principles*, edited by E.T. Higgins and A. Kruglanski, 433–465. New York: Guilford.

Sharp, L.M. and J. Frankel. 1983. "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly* 47(1): 36–53. Doi: http://dx.doi.org/10.1086/268765.

Shettle, C. and G. Mooney. 1999. "Monetary Incentives in US Government Surveys." *Journal of Official Statistics* 15(2): 231–250. Available at: http://www.jos.nu/Articles/abstract.asp?article=152231 (accessed January 2016).

Shuttleworth, F.K. 1931. "A Study of Questionnaire Technique." *Journal of Educational Psychology* 22(9): 652–658. Doi: http://dx.doi.org/10.1037/h0074591.

Singer, E., J. Van Hoewyk, and M. Maher. 2000. "Experiments with Incentives in Telephone Surveys." *Public Opinion Quarterly* 64(2): 171–188. Doi: http://dx.doi.org/10.1086/317761.

Tzamourani, P. and P. Lynn. 1999. "The Effect of Monetary Incentives on Data Quality – Results from the Social Attitudes Survey 1998 experiment." CREST Working Paper 73, Oxford, UK. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.573.5965&rep=rep1&type=pdf (accessed January 2016).

Weisberg, H.F. 2005. *The Total Survey Error Approach. A Guide To The New Science Of Survey Research*. Chicago/London: The University of Chicago Press.

Willimack, D.K., H. Schuman, B.E. Pennell, and J.M. Lepkowski. 1995. "Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to-Face Survey." *Public Opinion Quarterly* 59(1): 78–92. Doi: http://dx.doi.org/10.1086/269459.

Received January 2016
Revised January 2017
Accepted January 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 455–475, http://dx.doi.org/10.1515/JOS-2017-0022

Comparison of 2010 Census Nonresponse Follow-Up Proxy Responses with Administrative Records Using Census Coverage Measurement Results

Mary H. Mulry[1] *and Andrew D. Keller*[1]

The U.S. Census Bureau is currently conducting research on ways to use administrative records to reduce the cost and improve the quality of the 2020 Census Nonresponse Followup (NRFU) at addresses that do not self-respond electronically or by mail. Previously, when a NRFU enumerator was unable to contact residents at an address, he/she found a knowledgeable person, such as a neighbor or apartment manager, who could provide the census information for the residents. This was called a proxy response. The Census Bureau's recent advances in merging federal and third-party databases raise the question: Are proxy responses for NRFU addresses more accurate than the administrative records available for the housing unit? Our study attempts to answer this question by comparing the quality of proxy responses and the administrative records for those housing units in the same timeframe using the results of 2010 Census Coverage Measurement (CCM) Program. The assessment of the quality of the proxy responses and the administrative records in the CCM sample of block clusters takes advantage of the extensive fieldwork, processing, and clerical matching conducted for the CCM.

Key words: 2020 Census; correct enumeration.

1. Introduction

The planning for the 2020 U.S. Census includes a program of research and testing aimed at developing methodology and processes to achieve cost containment and maintain quality. The program includes exploring and creating fundamental changes to the design, implementation, and management of the decennial census. A series of tests investigate proposed changes such as using adaptive strategies for conducting Nonresponse Followup (NRFU) of the housing units that do not self-respond in a census. The examined strategies include using administrative records and a variable number of contact attempts with the goal of reducing costs and improving data quality. One avenue of research focuses on whether administrative records can reduce the 2020 Census Nonresponse Followup (NRFU) fieldwork at addresses where the Census Bureau did not receive a self-response electronically or by mail. In previous censuses, when enumerators were unable to contact

[1] U.S. Census Bureau, Washington, DC 20233, 4600 Silver Hill Rd, Suitland, MD 20746, U.S.A. Emails: mary.h.mulry@census.gov and andrew.d.keller@census.gov

Acknowledgments: The authors thank Tom Mule for his useful advice and consultations. The authors thank Eric Slud and Richard Griffin, three anonymous referees, and the editors for their helpful comments on earlier versions of this manuscript. This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

a household after a specified number of attempts, the instructions were to find a knowledgeable person. This person, perhaps a neighbor or apartment manager, who provided the census information for the residents, was called a proxy respondent. The question is whether a combination of federal and third-party databases provides better census information than the proxy responses.

Our study attempts to answer this question by comparing the quality of the proxy responses in the 2010 Census with administrative records for the same housing units. Previous studies have indicated differences in the quality of reporting of the population count and characteristics of the residents from household member respondents as opposed to proxy respondents. Both U.S. Constitutional and legislated uses of the census data involve the population counts and characteristics, such as age, sex, and race/Hispanic ethnicity so the collection of these data is fundamental to some government functions. Studies of proxy data following the 2000 Census found fewer missing characteristics in responses from household members versus proxies such as neighbors, postal workers or landlords (Chesnut 2005; Wolfgang et al. 2003). Regarding census coverage, Martin (1999) found that proxy reports of 'usual residence' increased undercoverage, particularly for unrelated household members. As part of research associated with the 2010 Census, King et al. (2012) found that self-report respondents provided more complete household membership than proxy respondents did.

The comparison of the quality of proxy responses and administrative records relies on the results of the 2010 Census Coverage Measurement (CCM) Program, which collected and processed the data used in forming estimates of census coverage error (Mule 2012). The goals of our study also include identifying variables that correlate with the quality of proxy responses and administrative records. Such variables, if they exist, would be useful in formulating decision rules for census processing. To provide context, our study also examines the quality of NRFU data from respondents who are household members and the administrative records available for the same addresses.

Ideally, one of the census tests could include a comparison of the proxy response for a housing unit and the administrative records for the same housing unit against a 'gold standard' interview conducted by a highly skilled interviewer with the residents of the housing unit. A determination could then be made as whether the proxy or the administrative records had better information, or whether they were of comparable quality. However, the 2020 Census testing cycle has a tight timeframe, which does not allow for a gold standard interview operation.

This article compares the quality of the 2010 Census NRFU housing units with proxy responses and the administrative records for the same housing units using the results of the 2010 Census Coverage Measurement (CCM) in a sample of block clusters. The approach is similar to a methodology discussed in Mulry and Spencer (2012). The administrative records files in our study come from two sources: (1) the Internal Revenue Service (IRS) 1040 forms filed in all months of 2010, and (2) the Medicare records for all months of 2010. The files from these two sources have the advantage of containing data for households.

This report describes the results of the first phase of our assessment. The second phase continues and includes a comparison of demographic characteristics of NRFU proxy responses and administrative records in corresponding housing units. Another aspect is to develop statistical models to identify the characteristics of NRFU housing units with

corresponding administrative records that have a high probability of being correct. The development of the models will consider characteristics of the households as well as geographic and socioeconomic variables available for census tracts and block groups from the U.S. Census Bureau's Planning Database (U.S. Census Bureau 2015). The Planning Database includes data from the U.S. Census Bureau's American Community Survey and the 2010 Census.

2. Research Approach

2.1. Research Questions

We aim to answer the following questions in order to produce information useful for the strategy design of contacting housing units during the 2020 Census NRFU:

- Are proxy responses for NRFU addresses more or less accurate than the administrative records available for the housing unit?
- What variables correlate with the accuracy of proxy responses for individual records and for records grouped by housing unit?
- What variables correlate with the accuracy of administrative records for individual records and for records grouped by housing unit?

2.2. Population

According to census residency rules, the correct address for a person's enumeration is his/her usual residence around Census Day, which is April 1 of the census year. The population under study is defined as the people whose Census Day residence is a housing unit enumerated in the 2010 Census NRFU by a proxy respondent, and administrative records are available for the housing unit. We consider the quality of two lists of the population using the criteria of whether the person is found at the correct location on Census Day according to census residency rules. One list of this population is the census enumerations, and the other list is the administrative records for the same housing units. For context, we also examine the quality of NRFU enumerations where the respondent is a household member and the administrative records at these addresses.

In this study, the definitions of the populations enumerated by proxy and household member respondents are operational and depend on the conduct of the 2010 Census operations. The housing units enumerated by household member respondents failed to self-respond by mail. The housing units enumerated by proxy failed to self-respond by mail, and none of the household members gave an interview to an NRFU enumerator. In 2010, enumerators had to make six contact attempts prior to taking a proxy interview. Therefore, our analyses, as well as the population definition, are conditional on the type of response observed in the 2010 Census. In addition, the analysis is conditional on the sources of administrative records that we consider.

2.3. Gold Standard

The assessment of the quality of the proxy responses and the records in the selected administrative files takes advantage of the extensive fieldwork, processing, and clerical

matching conducted for the CCM, which is the justification for using the CCM results as a gold standard. The 2010 CCM was designed to measure census coverage error with a post-enumeration survey composed of two samples, the population sample (P-sample) and the enumeration sample (E-sample). The former is a sample of housing units and persons selected independently of the census and designed to support the estimation of people missed in the census. Members of P-sample households are interviewed and then matched to the census on a case-by-case basis to determine whether they were enumerated in the census or missed. The E-sample is a sample of census enumerations (records) in the same areas as the P-sample and designed to support the estimation of erroneous enumerations. The data processing included a computerized search of census records to identify census enumerations for the P-sample and E-sample individuals (Cantwell et al. 2009). In addition, a computer-assisted clerical operation searched for enumerations for the P-sample individuals in the local area as well as duplicates of E-sample enumerations. When there was ambiguity, fieldwork collected additional information to resolve the status. Each P-sample and E-sample record that CCM processed was assigned a residence code indicating one of the following: (1) the person was a resident of the sample block cluster on Census Day, (2) was not a resident on Census Day, or (3) had unresolved Census Day residence. Figure 1 displays an overview of the CCM data collection and processing.

The P-sample interviews occurred in August and September 2010 independently from the 2010 Census. These interviews collected data that enabled constructing the Census Day (April 1) roster for the address by asking when current residents moved to the address

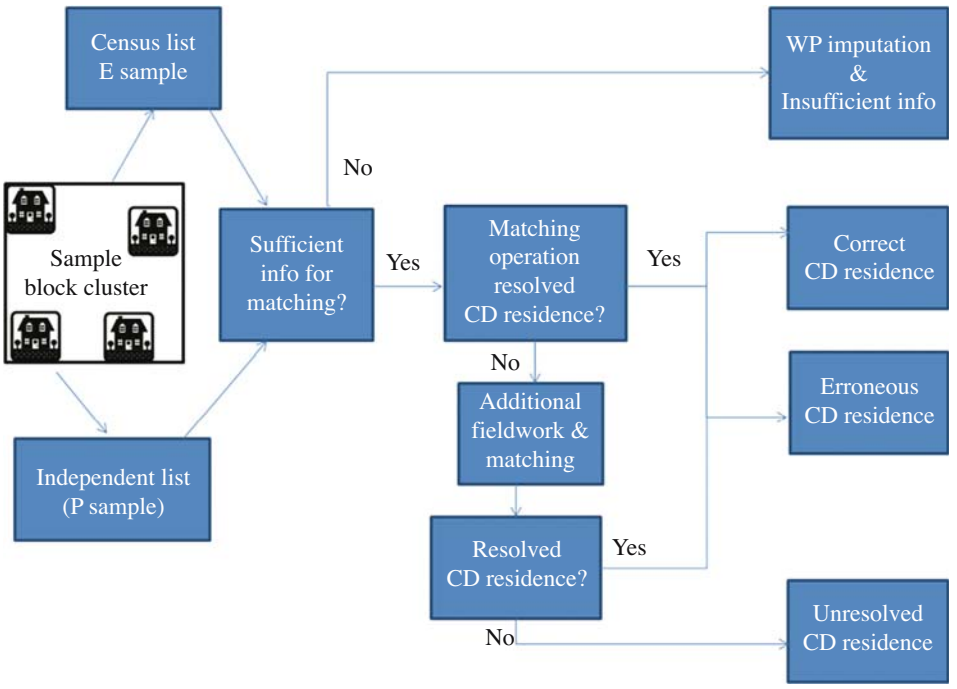

Fig. 1. Overview of CCM data collection and processing that produces codes indicating residence status on Census Day (CD). Note: WP imputation indicates whole person imputation, which is discussed in Subsection 2.4.

and about any Census Day residents who had moved from the address. The Census Bureau used a combination of electronic and clerical operations to match the P-sample people to the 2010 Census enumerations and conducted follow-up interviews in February 2011 to collect additional data when a person's Census Day residence could not be resolved. The CCM operation determined whether the census enumerations and P-sample persons were residents of their sample block cluster or the blocks surrounding the block cluster on Census Day by assigning the statuses of resident, nonresident, and unresolved. The CCM built this tolerance to avoid including minor geocoding error or mail delivery mistakes in the coverage error estimates, which would increase the variability of the estimates.

Since the P-sample is available only for the block clusters in the CCM sample, the comparison has to be restricted to the CCM block clusters. Although the 2010 CCM estimation does not require assuming that the P-sample interview is the 'truth,' the P-sample interviews are believed to be of higher quality because the interviewers have more training and experience since they were chosen from the pool of the best NRFU interviewers. In addition, the CCM interviewers were supported with a Computer Assisted Personal Interviewing (CAPI) instrument and supplied with additional residence probes.

The NRFU enumerations in the E-sample have residence status codes assigned during the CCM processing, but the administrative records in the NRFU housing units do not. We link the administrative records to the E- and P-sample records to retrieve CCM residence status codes. When a person's administrative record links to an enumeration in housing unit enumerated by a proxy response at the same address, the CCM residence code for the proxy response will indicate whether the person's enumeration at the address was correct. For example, if the person was enumerated at two addresses and the address not in the sample block was the correct Census Day residence, the enumeration in the sample block cluster was coded erroneous. This would mean the location of the person's administrative record was also in error. However, when a proxy response for a person and the administrative record file disagree, the CCM results provide information about whether the person should have been enumerated at the address and whether one of the sources is better for the person. Requiring the same address for a person's administrative record and the linking NRFU enumeration to retrieve a CCM residence code lends credibility to the assumption that the person lived at or is associated with the address. An administrative record will be inserted in the census at its address if the Census Bureau decides to use administrative records as enumerations. Requiring the same address from both sources means the correct enumeration rate reflects the accuracy of the use of administrative records at the addresses where they will be inserted in the census.

2.4. Matching Administrative Records to Combined CCM

The comparison of the 2010 Census NRFU housing units with proxy responses and the administrative records data for the housing units in the CCM block clusters requires linking the administrative records to the combined CCM to retrieve residence codes assigned during the CCM processing. The linking between the administrative records data and the combined CCM requires that both sources include Protected Identification Keys (PIKs). These PIKs are essentially encrypted Social Security Numbers or Individual Tax Identification Numbers, which are included when we use the term Social Security

Numbers. Administrative records data comes with Social Security Numbers that the Census Bureau staff converts to PIKs after a validation of their accuracy through matching to Social Security Administration files, a procedure called the Person Identification Validation System (PVS) (Wagner and Layne 2014). When a data file with records for persons does not come with Social Security Numbers, the Census Bureau uses its system to look up Social Security Numbers in Social Security Administration files and encrypt them by assigning PIKs. For this work, census and P-sample person records were assigned a PIK in a cascading search through the four search modules discussed in Wagner and Layne (2014): geographic search, name search, date of birth search, and household composition search. Each module has its own set of user defined blocking passes and parameter score thresholds. Layne et al. (2014) examine the error in PIK assignment by the PVS system associated with each of those search modules. It should be noted that this research assumes all PIKs are assigned with equal accuracy. PIKs have been assigned to the 2010 Census so the NRFU enumerations in the housing units with proxy responses have PIKs. PIKs also have been assigned to all the names collected in the P-sample regardless of the ultimate classification of nonmover, in-mover, out-mover, or never a resident of the sample block. Figure 2 illustrates the process of assigning PIKs and linking the files.

Sometimes the PVS fails to assign a PIK to a record. For example, 90.3% of the 2010 Census enumerations received a PIK from the PVS, but only 97% of the enumerations had enough information for an attempt to assign a PIK (Wagner and Layne 2014). Evaluation studies have shown that missing date of birth in a record is highly correlated with the PVS not assigning a PIK. In addition, an incomplete or fake name in a record is highly correlated with a PIK not being assigned (Wagner and Layne 2014; Mulrow et al. 2011). Nevertheless, it is possible to assign a PIK for someone missing sex or age particularly if other blocking and matching variables exist by which a high quality match can be made. However, a missing matching variable may result in a lower match score. Mulrow et al. (2011) found socioeconomic differences between the records that received PIKs and those that did not in a study using American Community Survey data. For example, the percentage assigned a PIK tended to be higher among those over 35 years of age than those younger. In addition, a higher percentage of those with a college degree received a PIK than those with a high school degree but not a college degree.

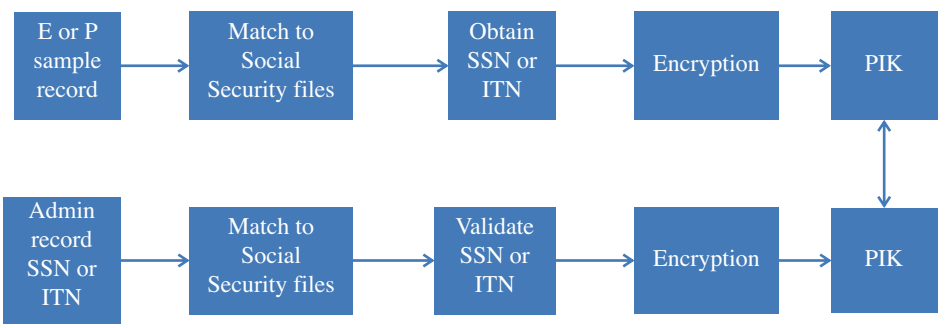

Fig. 2. The PVS assigns a Protected Identification Key (PIK) based on the person's Social Security Number (SSN) or Individual Tax Identification Number (ITN) for matching between the CCM E- and P-sample records and administrative records.

Having the CCM results available to compare proxy responses and administrative records is important because the estimated correct enumeration rate for the 2010 Census was 70.1% for persons enumerated by proxy respondents with 23.1% having all characteristics imputed, 5.6% being duplicates, and 1.1% being erroneous for other reasons. In contrast, 93.4% of the persons enumerated by a household member in NRFU were correct with 1.6% having all characteristics imputed, 4.2% being duplicates, and 0.8% being erroneous for other reasons (Mule 2012, Keller and Fox 2012). Even though enumerations that had all characteristics imputed, called *whole person imputations*, were not processed in the CCM E-sample due to lack of information to identify a person uniquely, the corresponding housing unit was included in the CCM P-sample and usually has information about the residents that can be used to evaluate any administrative records associated with the address. The P-sample also may have residency information for enumerations that are data-defined (i.e., processed in the E-sample) but have insufficient information to be processed in the CCM. The CCM requirement for sufficient information is a name and at least two characteristics because the CCM operations matched the enumerations to the names on the P-sample interview rosters.

When a person is enumerated by a proxy response and is in the administrative records file at the same address, the CCM residence code for the proxy response indicates whether the person's enumeration at the address was correct. If a person appears in the administrative records file but does not link to a combined CCM record at the same address, we can search the PIKs assigned to 2010 Census enumerations to learn if the person was enumerated elsewhere, but are not able to assess the accuracy for enumerations outside the CCM sample block clusters. If the person has an enumeration elsewhere that could not be assigned a PIK, we are not able to detect it using PIK matching.

Other types of electronic matching algorithms that do not rely on the assignment of PIKs, such as the household-based matching used by CCM, were not attempted. Household-based matching may or may not identify additional links between administrative records and the combined CCM. Regardless, our results must be viewed as conditional on the use of PIK matching.

Linking the administrative records to the CCM records enables identifying administrative records that are at the correct Census Day residence and those that are at an erroneous Census Day residence. Then a comparison of the percentages of administrative records and NRFU proxy responses in the CCM sample at the correct Census Day residence provides a measure to answer the research questions in Subsection 2.1.

2.5. *Underlying Assumptions*

This study approach has five major underlying assumptions:

- The results for proxy interviews in NRFU in the 2010 Census are applicable to the proxy interviews that would occur in the 2020 Census. The implementation of self-response and NRFU in the 2020 Census will be different from what occurred in the 2010 Census, and in particular, the procedures for taking proxy interviews in NRFU will differ.
- The 2010 CCM was able to determine whether the people on the rosters in NRFU proxy interviews were enumerated at the correct location, meaning their usual residence.

- The electronic matching algorithm used in this study (described in Subsection 2.4) was able to link a person's administrative record to the same person's record in the CCM P and E-samples.
- The availability of records from the administrative sources used in this study reflects the future availability from these sources.
- When a person has the same address in administrative records and NRFU, the person lives at or is associated with the address.

2.6. *Data*

For this study, we are going to focus on housing units in the CCM sample block clusters that were on the NRFU list in the E-sample and on the independent list of housing units created for the P-sample, and call this group the *combined CCM*. We need both E-sample and P-sample records because some or all the records for an occupied housing unit on the census list may be whole person imputations, but the P-sample interviewers were able to obtain data for the residents. In addition, the P-sample may have information regarding persons in administrative records not listed on the census form. We use the combined CCM to look up residence status codes for the administrative records. We do not form estimates using the combined CCM.

The administrative records file is the merger of the two files unduplicated within housing units: (1) the IRS 1040 forms filed in all months of 2010, (2) the Medicare records for all months of 2010. One reason the files were not unduplicated across housing units is that when duplicate records appear, there is no way to determine which is at the person's usual residence on Census Day. As stated earlier, the files from these two sources contain data for whole households. In addition, the 2014 Census Test operations used only these two sources.

The combined CCM contains 27,724 housing units that were proxy responses in NRFU with 10,416 occupied in NRFU, 15,012 vacant and 2,296 deleted because they did not have living quarters. Table 1 shows that of the 10,416 occupied housing units, 5,310 also have administrative records, the implication being that 5,106 have no records in the administrative records files we are using. Therefore, enumeration of these 5,106 housing units with proxy respondents using the combination of IRS 1040 and Medicare files is not an option unless other administrative sources with records for the housing units are found. However, one must keep in mind that the CCM oversamples hard-to-count areas. For a fit-for-use check, the percentage of the 23.6 million occupied housing units in NRFU that

Table 1. 2010 Census NRFU housing units in the combined CCM by administrative records (AR) status and type of NRFU respondent (unweighted).

	Proxy		HH Member	
AR status of housing units	HUs	%	HUs	%
Person records on AR list	5,310	51.0	16,876	61.3
No person records in AR list	5,106	49.0	10,647	38.7
Total	10,416	100.0	27,523	100.0

Note: Administrative records include IRS 1040 forms and Medicare records for all of 2010.

Table 2. Number of individual records found in administrative records (AR) files and number of individual records found on the combined CCM list in housing units in the combined CCM and occupied in the census by type of NRFU respondent.

Respondent type	AR	NRFU
Proxy	12,880	11,766
Household member	50,876	51,485
Total	63,756	63,251

have records in the combination of IRS 1040 and Medicare files is 56%. Therefore, the combined CCM percentages are reasonably comparable with proxy housing units being a little lower than the overall average at 51% and the housing units with household member respondents being a little higher at 61.3%.

For the NRFU housing units in Table 1 that have administrative records, Table 2 shows the distribution of the number of NRFU person records enumerated by proxy and household member respondents and the corresponding number of administrative records for the same housing units. In each of the two sources, the size of population in the proxy housing units is about 25% of the size of population in the housing units enumerated by household members. The administrative records file has more people in housing units enumerated by proxy than NRFU but fewer people in the housing units enumerated by household members. To see what would happen if all of these NRFU housing units were enumerated using administrative records, we combine the administrative records for NRFU housing units enumerated by both types of respondents and observe that the administrative records file has 505 records more than NRFU, about a 0.8% difference. Late in the analysis, we discovered that 88 of the administrative records persons in the proxy housing units and 237 in the housing units enumerated by a household member had died in 2009. These remain in the analysis but we address this issue for administrative records file construction in the recommendations in Section 4.

The 5,310 housing units with administrative records had 11,766 NRFU enumerations of persons with 9,258 of those having at least two characteristics, which is considered enough information to be an enumeration and is called *data-defined*. One of these characteristics could be a name. The remaining 2,508 were whole person imputations. Therefore, the imputation rate in these housing units is 21.3%, which is lower than the 23.1% for imputations among NRFU proxy enumerations nationally.

For completeness, we note that our analysis does not include 1,048 housing units with proxy respondents in the E-sample that are not also on the P-sample list, making them ineligible for the combined CCM list. The number of these housing units containing administrative records is 231 resulting in 460 administrative records for persons not being evaluated. In addition, the study does not include the 6,154 housing units on the P-sample list that were not on the E-sample list.

2.7. *Evaluation Criteria*

The evaluation of the quality of enumerations from the proxy responses and records in the administrative records file in the same housing units includes the rate of correct enumerations. The assessment also includes comparing the count of persons in each

source. Comparable calculations are made for enumerations and administrative records in housing units with household member responses.

- The total number of people enumerated at the sample addresses in each source.
- The total number of people correctly enumerated at the sample addresses in each source.
- HUs classified by (1) all administrative records are at the correct Census Day residence, (2) at least one administrative record is erroneous (not at the Census Day address) or its Census Day residence is unresolved, and (3) at least one Census Day resident does not have an administrative record at the address.

3. Results

Although the focus of our analyses is the NRFU housing units enumerated by proxy respondents, we are going to present results for NRFU housing units enumerated by household members for comparison. First, Subsection 3.1 considers the quality of the records for persons using the results of the CCM to determine whether the address on the record is in the correct location. Analyzing the quality of individual records provides insight when viewing the quality of the records for complete households, which is the focus of Subsection 3.2. In addition, analyses of individual records provide information about several potential uses of administrative records, such as for enumeration and for use in developing imputation models.

3.1. Quality of Individual Person Records

Even though Table 2 shows the number of records in administrative records and NRFU generally agree, this alone is not enough to evaluate the quality of the individual records in the two systems, which is the topic of our first research question. We need to know whether a person's record is at the correct location of the person's Census Day residence and whether the characteristics of the person and the size and composition of the households are correct.

Two things have to happen to evaluate an administrative record for a person: (1) the person's administrative records PIK has to link to the PIK for a record in the combined CCM and (2) the combined CCM record has to have a resolved residence status.

Table 3 shows the weighted distribution of combined CCM residence status for enumerations and administrative records in NRFU housing units in the combined CCM by NRFU respondent type while Table A1 in the Appendix shows the same results unweighted. The first thing to notice is that the unweighted and weighted distributions of CCM residence status are very similar for each NRFU respondent type. The weighted and unweighted distributions for the administrative records in housing units by NRFU respondent type also are similar. The weights are the CCM E-sample block cluster weights not adjusted for CCM nonresponse. Since the CCM sample design was able to keep the block cluster weights within a tight range, the similarity of the unweighted and weighted distributions is reasonable. We use the weighted results in our discussion.

To compare the distributions of the residence statuses from different types of respondents or different sources, we perform a chi-square test using the Rao-Scott adjustment (Lohr 1999) to account for the sampling design. For the design effect of the

Table 3. Weighted distributions of combined CCM residence status for enumerations and administrative records (AR) in NRFU housing units in the combined CCM by NRFU respondent type (shown in thousands).

	Proxy respondent			
	NRFU		AR	
Census Day residence status	count	%	count	%
Correct residence	5,235.2	56.6	5,017	49.1
Erroneous residence	380.9	4.1	418	4.1
Unresolved residence	1,462.4	15.8	379	3.7
NRFU not processed by CCM				
Insufficient info	258.3	2.8	-	-
Whole person imputation	1,920.6	20.7	-	-
AR PIK not in census at same address			4,397	43.1
Total	9,257.4	100.0	10,212	100.0
	Household member respondent			
	NRFU		AR	
Census Day residence status	count	%	count	%
Correct residence	36,720.2	88.0	29,971	72.5
Erroneous residence	1,058.9	2.5	1,054	2.5
Unresolved residence	2,308.2	5.5	1,283	3.1
NRFU not processed by CCM				
Insufficient info	1,070.9	2.6	-	-
Whole person imputation	583.0	1.4	-	-
AR PIK not in census at same address			9,038	21.9
Total	41,741.2	100.0	41,346	100.0

CCM sample, we examined Table 8 in Olson and Griffin (2012) that contains the means of several ranges of the observed correct enumeration rate, the number of observations in each range, and the standard error of the mean. The design effects varied between 2.5 and 3.5 across the categories. We use a design effect of three for the Rao-Scott adjustment to the chi-square statistics. For the chi-square tests, we use four cells: correct residence, erroneous residence, unresolved residence, and unable to process. For NRFU, we define the unable-to-process cell by collapsing insufficient information for CCM and whole person imputations. For administrative records, we collapse the records found at another census address and those not linked to a combined CCM record.

For the NRFU proxy enumerations, Table 3 shows that CCM found that 56.6% were at the correct residence, and 4.1% were at an erroneous residence. CCM attempted but could not determine Census Day residence for 15.8% of the NRFU proxy enumerations. CCM did not attempt to process the 2.8% that had insufficient information or the 20.7% that were whole person imputations.

For the NRFU enumerations by household members in Table 3, we see that 88.0% are at the correct residence, 2.5% are at an erroneous residence, and 5.5% had an unresolved residence status. However, 2.6% had insufficient information for CCM to process and 1.4% of the proxy enumerations were whole person imputations, which CCM did not process.

Turning to the residence status of the administrative records in NRFU housing units in Table 3, for proxy respondents, links to combined CCM records showed that 49.1% were

at the correct residence, 4.1% were at an erroneous residence, and 3.7% had an unresolved residence. The percentage that did not link at the same address and could not be evaluated is 43.1%. When we examine the administrative records in the housing units with household member respondents, we see that links to the combined CCM found that 72.5% were at the correct residence, 2.5% were at an erroneous residence, and the residence status of 3.1% could not be resolved. The percentage that did not link at the same address and could not be evaluated is 21.9%.

For some insight about the administrative records that did not link, the unweighted data in Table A1 shows that 17.3% of the individual enumerations by a proxy respondent and 10.5% of the individual enumerations by a household member respondent did not link to a combined CCM record at the same address but linked to enumerations elsewhere in the census. In addition, 26.8% of the individual enumerations by a proxy respondent and 12.9% of the enumerations by a household member respondent did not link to a combined CCM record at the same address or elsewhere in the census. For the administrative records found elsewhere in the census, using the administrative records for enumeration would create duplicate enumerations. We do not have information to determine which address was the correct location for their enumeration since the census address was not in the CCM. These people may have moved or may alternate between two residences, such as families with seasonal homes or children in shared custody. In these cases, they may have been enumerated in one location and list the other address as their residence in administrative records. As for the administrative records that did not link anywhere in the census, there are two possible explanations: (1) the person has a census enumeration but it has errors or not enough information for the linking procedure to make the connection; (2) the person was missed by the census.

Next, we compare the distributions of the residence statuses for the NRFU enumerations and the administrative records by respondent. For the housing units with proxy respondents, the chi-square test produced a p-value less than 0.001, which leads us to conclude that the distribution of the residence statuses for the NRFU enumerations with 56.6% correct and the administrative records with 49.1% correct are different. For the housing units with household member respondents, the p-value of the chi-square test is 0.028, which indicates the distributions of the residence codes are different. For both types of respondents, the percentage of NRFU enumerations at the correct residence is higher than observed for administrative records, and the percentage of administrative records that cannot be evaluated is higher than observed for NRFU enumerations.

Both NRFU and administrative records have a substantial percentage of records where this approach is unable to evaluate their residence status. The seemingly high percentage of records that do not link to a combined CCM record at their administrative records address but link to a census address elsewhere causes concern that these administrative records are not at the correct Census Day residence and more importantly, that inserting them as census enumerations would create duplicate enumerations. Since the CCM sample did not include the address where administrative records PIKs were found, the CCM did not evaluate the accuracy of the enumeration of the people at the address. Therefore, the accuracy of administrative records that linked to these enumerations also could not be evaluated.

Interestingly, the percentage of records with a CCM resolved residence status is higher for NRFU enumerations than administrative records in housing units with both

types of respondents. Keep in mind that all the administrative records have PIKs, but the Census Bureau procedure may or may not be able to assign PIKs to the census enumerations.

From another perspective, we compare the distributions of the residence status of the NRFU enumerations for the two types of respondents. A chi-square test comparing produced a p-value less than 0.001; therefore, we conclude that the distributions are different. We see that the percentage of proxy enumerations that are at the correct residence at 56.6% is lower than the percentage of household member enumerations at the correct residence at 88.0%. The most apparent difference is that the percentage of whole person imputations is much higher for the proxy enumerations at 20.7% than for the household member respondents at 1.4%. However, the housing units that are remaining after the attempts to get household member respondents fail get rolled over to the attempts to get proxies. So, almost all the whole person imputations are attributed to the proxies, although both the self-response phase and the NRFU household member response phase also fail to get a response.

Similarly, a chi-square test to compare the distributions of the administrative records for the two respondent types produces a p-value of 0.010, which indicates that the distributions are different. The percentage of administrative records that are at the correct residence is 49.1% in the housing units enumerated by proxy while the percentage correct is higher at 72.5% in the housing units enumerated by a household member. In addition, the percentage that did not link at the same address and could not be evaluated is higher for proxy respondents 43.1% than for household member respondents at 21.9%.

3.2. Characteristics Correlated with Quality

When we consider our second research question, we note that the assignment of PIKs to the combined CCM records proved crucial to evaluating the administrative records in housing units enumerated during NRFU. Therefore, the percentage of NRFU enumerations that received PIKs is an evaluation tool. Table 4 shows the distribution of the residence status of enumerations with PIKs and those without PIKs by NRFU respondent. Of the NRFU enumerations where the PVS attempted to assign PIKs, 73% (SE = 0.9%) of those in housing units enumerated by proxy received PIKs while 92% (SE = 0.2%) of those enumerated by a household member received PIKs. If the whole person imputations are included, the percentage is 58% (SE = 0.8%) for proxy respondents and 91% (SE = 0.2%) for household member respondents. When whole person imputations are included and when they are not, the tests of difference between the percentages of enumerations assigned PIKs for proxy and household member respondents produced p-values less than 0.001, so we conclude there is a difference in the enumerations from the two types of respondents.

In summary, a distinguishing feature that indicates the quality of NRFU enumerations appears to be whether they can be assigned a PIK. Those that receive PIKs tend to be in the correct location at high rate. Table 5 shows the correct enumeration rate for several criteria for the denominator for enumerations with and without PIKs by type of NRFU respondent. We do not conduct statistical testing but use the data in Table 5 to illustrate the effect of the choice of the denominator of the correct enumeration rate.

Table 4. Weighted distributions of combined CCM residence status for enumerations in NRFU housing units by NRFU respondent type and PIK status (shown in thousands).

	Proxy			Household member		
Census Day residence status	with PIK	without PIK	Total	with PIK	without PIK	Total
PIK attempted						
Correct residence	3,625.8	1,609.4	5,235.2	34,322.1	2,398.2	36,720.2
Erroneous residence	266.4	114.5	380.9	844.0	214.9	1,058.9
Unresolved residence	337.5	173.2	510.7	1,713.5	594.7	2,308.2
Insufficient info for CCM	1,124.9	85.1	1,210.0	990.8	80.1	1,070.9
Subtotal	5,354.6	1,982.1	7,336.7	37,870.3	3,287.9	41,158.3
	73%	27%	100%	92%	8%	100%
PIK not attempted						
Whole person imputation		1,920.6	1,920.6		583.0	583.0
Total	5,354.6	3,902.8	9,257.4	37,870.3	3,870.9	41,741.2
	58%	42%	100%	91%	9%	100%

Table 5. Weighted correct enumeration (CE) rate for enumerations in occupied housing units in the combined CCM with several criteria for the enumerations included in the denominator by type of NRFU respondent. (shown in thousands).

Status of enumerations in denominator	Proxy respondent			HH member respondent		
	Total	CE	% CE	Total	CE	% CE
With PIK						
CCM resolved status	3,892	3,626	93	35,166	34,322	98
Data-defined	5,355	3,626	68	37,870	34,322	91
Without PIK						
CCM resolved status	1,724	1,609	93	2,613	2,398	92
Data-defined	1,982	1,609	81	3,288	2,398	73
Data-defined and imputed	3,903	1,609	41	3,871	2,398	62

When the denominator includes only the enumerations where CCM could resolve the residence status, namely those that are correct and erroneous, the percentage correct is not dramatically different from the percentages for the household member respondents without PIKs and both categories for proxy respondents, which range from 92% to 98%. Additionally, Table 3 shows that the percentage of administrative records with a resolved residence status in proxy housing units that are correct is in the same range at 92% (5,017/(5,017+418)).

For the data-defined enumerations with PIKs, 68% from proxy respondents and 91% from household member respondents are in the correct location. However, the correct enumeration rate among enumerations that are data-defined but not assigned a PIK is 81% for proxy respondents and 73% for household member respondents. When the denominator for those without PIKs includes whole person imputations, the correct enumeration rate for proxy respondents is 41%. For household member respondents, rate becomes 62% with the inclusion of the imputations. Keep in mind that whole person imputations are a much smaller percentage of the enumerations by household members than for proxy respondents.

3.3. Quality of Records for Entire Households

As stated in the third research question, our ultimate interest is the quality of administrative records on a household basis because that is most likely the way they will be used for enumeration. Our analysis examines two measures. One is the percentage of housing units where the population counts from NRFU and administrative records are equal. The other is the percentage of NRFU housing units where the combined CCM determines the administrative records roster is perfect. These are descriptive analyses with unweighted data.

Table 6 shows that the percentage housing units where the NRFU and administrative records population counts are the same is 51% for both proxy and household member respondents. However, the administrative records population count being equal to the NRFU population count does not mean that the administrative records roster for the housing unit has the correct Census Day residents. CCM provides a means to determine the accuracy of the administrative records roster.

Table 6. Unweighted comparison of housing unit population counts from NRFU and administrative records (AR) by respondent type.

	Proxy		Household member	
Housing unit population counts	Number of housing units	%	Number of housing units	%
Same AR and census	2,685	51	8,633	51
Different AR and census	2,625	49	8,243	49
Total	5,310	100	16,876	100

Therefore, we examine the accuracy of the administrative records on a household basis for the 5,310 housing units with proxy respondents and 16,876 housing units with household member respondents that have administrative records. Table 7 shows the percentage of housing units in the following categories as determined by the combined CCM:

- Administrative Records Perfect – All administrative records persons in the housing units are Census Day residents at the address and no Census Day residents are omitted from the administrative records roster.
- Administrative Records Erroneous Enumerations and Unresolved Enumerations (E&Us) – At least one administrative record in the housing unit either linked to a combined CCM record coded as not being a Census Day resident at the address or did not link to a combined CCM record with a resolved residence status.
- Administrative Records Omissions – There is at least one person that the combined CCM found to be a Census Day resident at the address, but the person(s) is (are) not on administrative records roster for the address.

When the administrative records in the 5,310 proxy housing units are considered on a household basis instead of a individual basis, 1,722 (32.4%) are perfect in that the combined CCM indicated every record as being at the person's Census Day residence and no persons were omitted. We also find that administrative records for 408 (7.7%) of the housing units omit at least one person that the combined CCM found to be a Census Day resident at the address. The remaining 3,180 (59.9%) have at least one record that the combined CCM found not to be a resident at the address on Census Day, or the person's Census Day residence was not determined because the administrative records did not link to a combined CCM record with a resolved residence status.

Table 7. Status of administrative records (AR) in NRFU housing units in the combined CCM by NRFU respondent type (unweighted).

	Proxy		Household member	
Housing unit status	Number of housing units	%	Number of housing units	%
AR Perfect	1,722	32.4	7,256	43.0
AR E&U	3,180	59.9	6,846	40.6
AR Omissions	408	7.7	2,774	16.4
Total	5,310	100.0	16,876	100.0

Surprisingly, the percentage of housing units with household member respondents who omitted at least one Census Day resident from the administrative records roster was 16.4%. In addition, 43.0% of the administrative records rosters for housing units enumerated by household members are perfect. The percentage of housing units with an administrative record for at least one person who was not a Census Day resident or had an unresolved Census Day residence was 40.6%.

4. Summary

Our investigation discovered that determining whether proxy responses are more or less accurate than administrative records is not as straightforward as it sounds. The percentage of enumerations in housing units with proxy respondents in the correct location units was higher than the percentage for administrative records in the same housing units even though the administrative records sources were all IRS 1040 and Medicare records from 2010. However, the percentage of records that could not be evaluated was higher for the administrative records than for the proxy respondents. The high unresolved rate among administrative records was due to the failure to link the administrative records to a combined CCM record at the same address. The reasons that an administrative record did not link include the individual being enumerated at another address, having a census enumeration or P-sample roster entry that could not be assigned a PIK, or being missed by the census. This research prompted a change from the initial plan that used all administrative records for NRFU enumeration to the search for methods to identify the best administrative records for enumeration. The current methodological approach focuses on the development of predictive models to identify administrative records with a high probability of being accurate (Morris et al. 2016).

In addition, the findings of our study have implications for the census in the areas of administrative records sources used in census enumeration, the risk of duplication, and characteristics of high quality proxy enumerations. We recommend finding additional high-quality administrative records sources to increase the potential for using administrative records to enumerate housing units that cannot be enumerated well by proxy, such as the Supplemental Nutrition Assistance Program files from the states. We found that enumeration with administrative records was not an option for approximately half of the housing units in the CCM E-sample classified as occupied in the census using proxy respondents when the administrative records sources were IRS 1040 and Medicare files for all of 2010. If additional high quality administrative records sources cannot be found, these housing units without administrative records will need to be contacted by NRFU enumerators or imputed. However, the implication of increasing the number of administrative records sources is that it elevates the importance of identifying duplicate records across housing units and developing rules for which address to keep as the person's Census Day address. Algorithms for identifying duplicate records face challenges when sources do not have the same name, age, and/or address for a person. Examples include when one source has a person's nickname while the other source has the given name, and the old versus new address when a person moves. Adding sources of administrative records has the potential to increase the variation in the key variables used in linking the records, thereby increasing the errors in identifying duplicates.

One important finding is that not all proxy responses are bad as demonstrated by the result that over half of proxy enumerations are in the correct location (56.6%). By almost any standard, proxy enumerations that can be assigned PIKs tend to be in the correct location. Therefore, one indicator for a high quality NRFU enumeration appears to be whether it has enough information for the Census Bureau's Personal Validation System algorithm to assign a PIK. The implication is that the design of NRFU operations would profit by including strategies to obtain high-quality proxy responses. Such strategies would include designing the training of interviewers to emphasize the importance of obtaining the name and age of the residents from proxy respondents since these are important for assigning PIKs. Additional advantages may come from developing contact tactics that incorporate the times when knowledgeable proxy respondents are likely to be accessible, namely at home for neighbors or on the premises for multi-unit building managers.

However, the amount of information collected for an individual does not always assure that the PVS will be able to assign a PIK. Some data-defined census enumerations that meet the CCM criteria of sufficient information, which is a name and two characteristics, could not be assigned PIKs but were found by CCM to be enumerated at the correct location. The addresses where these people were enumerated may not have been associated with them in administrative records.

Since administrative records enumeration would occur on a housing unit basis, a comparison of NRFU proxy responses and administrative records for whole households on population count and accuracy of location also is important. The combined CCM found that an unweighted 32% of the administrative records for proxy housing units were enumerated perfectly. That means that all the administrative records persons in the housing unit were Census Day residents and no Census Day residents were omitted from the administrative records roster. The enumerations with unresolved residence status were not considered to be at the correct location. Some likely are, but without enough information to make a determination. When focusing only on population count, the percentage of housing units have an administrative records count that agrees with the census count is an unweighted 51% among housing units with proxy respondents and among housing units with household member respondents.

The results also indicate that duplication may be a problem when using administrative records to enumerate whole HHs. Census operations may need to search census enumerations, particularly self-responses, to be sure that an administrative records enumeration does not create a duplicate. If a search finds another enumeration for a person, the administrative record is not necessarily the one in the wrong location. Self-responses may be in error due to postal delivery errors or misunderstandings about the correct location for enumeration when a person has more than one residence. One approach to identifying which of two enumerations to keep in the census is to consult multiple administrative records sources and make the decision based on the recency and frequency of the appearances of the person at the addresses. The addition of questions regarding other residences to the census questionnaire may aid in avoiding duplicates.

Further research is needed to identify additional characteristics that indicate how the quality of the proxy responses may vary. Additional investigations could examine the demographic, geographic, and socioeconomic characteristics of the housing units where the combined CCM found their individual administrative records to be perfect, that is, the

exact household members were correctly enumerated versus those housing units with administrative records that had errors or could not be evaluated. Additional research could examine relationships between operational characteristics, such as the number of prior contact attempts and correct proxy responses to identify characteristics of housing units with complete correct administrative records among NRFU proxy responses.

The results of our study apply to identifying a person's usual residence and characteristics for census-taking and therefore probably have only limited implications for surveys since the focus of surveys usually is to collect behavior or socioeconomic information. Survey researchers do need to be aware that when linking a survey from a sub-national area to append additional administrative records data to individual records, the respondents may not be in administrative records at the survey address. As for administrative records, our study indicates that while administrative records contain a large amount of information, determining whether that data is truly adequate for the purpose at hand is not always easy.

Appendix

Table A1. Unweighted distributions of combined CCM residence status for enumerations and administrative records (AR) in NRFU housing units in the combined CCM by NRFU respondent type.

	Proxy respondent			
	NRFU		AR	
Census Day residence status	count	%	count	%
Correct residence	6,637	56.4	6,191	48.1
Erroneous residence	481	4.1	519	4.0
Unresolved residence	1,850	15.7	493	3.8
NRFU not processed by CCM				
Insufficient info	290	2.5	-	-
Whole person imputation	2,508	21.3	-	-
AR PIK not in census at same address				
Found at another census address	-	-	2,230	17.3
Not linked to census records	-	-	3,447	26.8
	11,766	100.0	12,880	100.0
	Household member respondent			
	NRFU		AR	
Census Day residence status	count	%	count	%
Correct residence	45,018	87.4	36,084	70.9
Erroneous residence	1,392	2.7	1,258	2.5
Unresolved residence	3,042	5.9	1,645	3.2
NRFU not processed by CCM				
Insufficient info	1,285	2.5	-	-
Whole person imputation	748	1.5	-	-
AR PIK not in census at same address				
Found at another census address	-	-	5,318	10.5
Not linked to census records	-	-	6,564	12.9
	51,485	100.0	50,869	100.0

5. References

Cantwell, P.J., M. Ramos, and D. Kostanich. 2009. "Measuring Coverage in the 2010 U.S. Census." In *JSM Proceedings*, Social Statistics Section, American Statistical Association, Washington, DC, August 1–6, 2009. Alexandria, VA: American Statistical Association. 43–54. Available at: https://ww2.amstat.org/sections/srms/proceedings/y2009/Files/302739.pdf (accessed January 2017).

Chesnut, J. 2005. "Item Nonresponse Error for the 100 Percent Data Items on the Census 2000 Long Form Questionnaire." In *JSM Proceedings*, Section on Survey Research Methods, American Statistical Association, Minneapolis, MN, August 7–11, 2005. Alexandria, VA: American Statistical Association. 2857–2864. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2005/Files/JSM2005-000341.pdf (accessed January 2017).

Keller, A. and T. Fox. 2012. "2010 Census Coverage Measurement Estimation Report: Components of Census Coverage for the Household Population in the United States." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-04. Washington, DC: U.S. Census Bureau. Available at: http://www.census.gov/coverage_measurement/pdfs/g04.pdf (accessed January 2017).

King, T., S. Cook, and J. Hunter Childs. 2012. "Interviewing Proxy Versus Self-Reporting Respondents to Obtain Information Regarding Living Situations." In *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, San Diego, CA, July 28–August 2, 2012. Alexandria, VA: American Statistical Association. 5667–5677. Available at: https://ww2.amstat.org/sections/srms/proceedings/y2012/files/400243_500698.pdf (accessed January 2017).

Layne, M., D. Wagner, and C. Rothhaas. 2014. "Estimating Record Linkage False Match Rate for the Person Identification Validation System." CARRA Working Paper Series. Working Paper #2014-02. Washington, DC: Census Bureau. Available at: https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-02.html (accessed March 2017).

Lohr, S. 1999. *Sampling: Design and Analysis*. Cengage Learning. Boston, MA.

Martin, E. 1999. "Who Knows Who Lives Here? Within-household Disagreements as a Source of Survey Coverage Error." *Public Opinion Quarterly* 63: 220–236. Doi: http://dx.doi.org/10.1086/297712.

Morris, D., A. Keller, and B. Clark. 2016. "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census." *Statistical Journal of the IAOS* 32: 177–188. Doi: http://dx.doi.org/10.3233/SJI-161002.

Mule, T. 2012. "Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-01. Washington, DC: U.S. Census Bureau. Available at: http://www.census.gov/coverage_measurement/pdfs/g01.pdf (accessed January 2017).

Mulrow, E., A. Mushta, S. Pramanik, and A. Fontes. 2011. Assessment of the U.S. Census Bureau's Person Identification Validation System. Report for the U.S. Census Bureau. Chicago, IL: NORC. Available at: http://www.norc.org/PDFs/May%202011%20Personal%20Validation%20and%20Entity%20Resolution%20Conference/PVS%20Assessment%20Report%20FINAL%20JULY%202011.pdf (accessed January 2017).

Mulry, M.H. and B.D. Spencer. 2012. "A Framework for Cost Models Relating Cost and Data Quality." Presentation at the 2012 International Total Survey Error Workshop. Sanpoort, The Netherlands, September 2–4, 2012. Research Triangle Park, NC: National Institute of Statistical Science. Available at: http://www.niss.org/sites/default/files/Mulry_september2012.pdf (accessed January 2017).

Olson, D. and R. Griffin. 2012. "2010 Census Coverage Measurement Estimation Report: Aspects of Modeling." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-10. U.S. Washington, DC: Census Bureau. Available at: http://www.census.gov/coverage_measurement/pdfs/g10.pdf (accessed January 2017).

U.S. Census Bureau. 2015. *Planning Database*. Washington, DC: Census Bureau. Available at: http://www.census.gov/research/data/planning_database/ (accessed January 2017).

Wagner, D. and M. Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications." CARRA Working Paper Series. Working Paper #2014-01. Washington, DC: Census Bureau. Available at: https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-01.html (accessed March 2017).

Wolfgang, G., R. Byrne, and S. Spratt. 2003. *Analysis of Proxy Data in the Accuracy and Coverage Evaluation*, Census 2000 Evaluation O.5. Washington, DC: U.S. Census Bureau. https://www.census.gov/pred/www/rpts/O.5.PDF (accessed March 2017).

Received January 2016
Revised February 2017
Accepted March 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 477–511, http://dx.doi.org/10.1515/JOS-2017-0023

Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ

Giles Reid[1], *Felipa Zabala*[2], *and Anders Holmberg*[3]

Many national statistics offices acknowledge that making better use of existing administrative data can reduce the cost of meeting ongoing statistical needs. Stats NZ has developed a framework to help facilitate this reuse. The framework is an adapted Total Survey Error (TSE) paradigm for understanding how the strengths and limitations of different data sets flow through a statistical design to affect final output quality. Our framework includes three phases: 1) a single source assessment, 2) an integrated data set assessment, and 3) an estimation and output assessment. We developed a process and guidelines for applying this conceptual framework to practical decisions about statistical design, and used these in recent redevelopment projects. We discuss how we used the framework with data sources that have a non-statistical primary purpose, and how it has helped us spread total survey error ideas to non-methodologists.

Key words: Total survey error; multiple data sources; official statistics; survey design.

1. Introduction

Producers of official statistics are facing increasing pressures to save money while maintaining or even increasing the quality and timeliness of outputs. In many countries, response rates for traditional surveys have been falling, but more and more administrative and other non-traditional data have become available. There is an urgent need to find ways to use administrative sources to increase the efficiency and effectiveness of statistical production. Administrative data cannot solve all of our problems, and traditional survey data collection is still needed in many cases. At Statistics New Zealand (Stats NZ), we have been faced with the challenge of redesigning statistical outputs to make better use of administrative data while maintaining the data quality required by our users.

In this article, we propose a quality framework that we have developed to provide a systematic approach to meeting the Stats NZ goal of using administrative data as the first source of data, supplemented by survey data collection only when necessary. This framework is widely applicable to all kinds of input data and statistical outputs and includes considerations of estimation models and continuous improvement alongside its total survey error foundations.

[1] Statistics New Zealand, 2018 Census, PO Box 2922, Wellington 6140, New Zealand. Email: giles.reid@stats.govt.nz.
[2] Statistics New Zealand, Statistical Methods, PO Box 2922, Wellington 6140, New Zealand. Email: felipa.zabala@stats.govt.nz.
[3] Statistics Norway, Division for Methodology, Akersveien 26 Oslo, Norway. Email: Anders.holmberg@ssb.no.

In 2016, Stats NZ released its vision to "unleash the power of data to change lives, which will enable data-led innovation across society, the economy, and the environment". Its aim is "to increase the value of data to decision-makers tenfold in the next 15 years" (Statistics NZ 2016b, 1). Survey methodology provides good ways to answer questions about how to measure and improve data quality, but it requires us to have a high degree of control over the entire process from collection to output. Designing statistical outputs that use administrative data creates many new challenges because we have to give up direct control over many processes, including population definitions, collection methods, classifications, and data editing. Each administrative source has its own particular problems that must be understood both for our own design work and to assure the final users of the data that our outputs are fit for purpose. When we use administrative data instead of a traditional survey, we need new processes, such as data integration and re-coding or adjusting administrative variables, which can introduce new types of errors.

The quality assessment framework presented in this article provides a basis for understanding how these factors fit together. We expand and unify earlier conceptual work from various writers to make it more directly and easily applicable to practical statistical design in official statistics. Based on our experiences from survey redevelopment projects within Stats NZ, we also provide a sequence of practical steps which can be followed during the design process.

Our three-phase framework applies the Total Survey Error (TSE) paradigm (see, for example Groves and Lyberg 2010; Biemer 2010) to the new realm of statistical production, which involves integrating and combining data from various sources. It builds on Li-Chun Zhang's extension of this TSE thinking to administrative and integrated data (Zhang 2012). It makes use of various quality indicators and measures, such as those developed as part of the European Statistical System Network (ESSnet) and BLUE-Enterprise and Trade Statistics (BLUE-ETS) projects, alongside earlier Stats NZ quality work, like metadata templates, output quality reviews, and process reviews. (Burger et al. 2013; Daas et al. 2011; Daas et al. 2012; Statistics NZ 2016a). The framework assists in understanding how well different data sets meet their originally intended purpose (Phase 1) and what their strengths and limitations are. It provides a way of determining what effects these strengths and limitations may have on the quality of a statistical output that makes use of these "found" data sources, statistically designed (possibly sample survey based) data, or a combination of the two (Phase 2). In that sense, our framework suggests an extension of Zhang's work including a third phase with evaluations between design options for a statistical output.

Quality assessments carried out with this framework can help answer statistical design questions on how to use available data to meet user needs in an efficient way. They help to decouple the true statistical needs of our users from design decisions: our goal should be to meet these needs as best we can with the data available. Sometimes reproducing the results of a sample survey using administrative sources may not be feasible, but a new alternative output can be produced which still meets existing needs, or meets emerging needs that the old survey outputs did not.

The framework is also part of Stats NZ's continued efforts to be better equipped for a changing data environment with an increasing array of unconventional data sources. Our new strategy is to increase use and reuse of data already collected, both in the production

of traditional official statistics and for new research projects. More reuse of data also means we need to always consider that all new data we collect may have new uses in the future. To make this reuse possible, documentation is essential; the framework gives a clear guide to what should be recorded and how the documentation should be structured. The framework also helps with managing data from multiple data sources simultaneously, and improving the opportunity to use them in an integrated way.

In this article, we adopt the United Nations Economic Commission for Europe (UNECE) definition of administrative data: "data that is collected by sources external to statistical offices" (UNECE 2011b, 2). and "administrative sources are data holdings containing information which is not primarily collected for statistical purposes," (UNECE 2011b, 4). We use the term "survey" in a classical sense, which does not necessarily mean the data acquired was selected with probabilistic methods.

We used the Organization for Economic Cooperation and Development (OECD) definition of a statistical product to define *statistical outputs*: "an information dissemination product that is published or otherwise made available for public use that describes, estimates, forecasts, or analyses the characteristics of groups, customarily without identifying the persons, organisations, or individual data observations that comprise such groups. This may include general-purpose tabulations, analyses, projections, forecasts, or other statistical reports" (OECD 2007, 745) as well as data sets containing unit record data.

In Section 2 we present the quality assessment framework and discuss how it is used. We also provide a list of quality measures and indicators that can be used to measure different types of error. The versatility of the framework is illustrated by three applications: the redesign of our Quarterly Building Activity Survey in Section 3, the use of tax data to measure personal income in Section 4, and the use of linked administrative data to estimate resident population counts in Section 5. We conclude with a summary and discussion.

2. The Quality Assessment Framework

2.1. *Developing the Framework*

Stats NZ's "administrative data first" goal means that during the design phase of a proposed statistical output, we first have to confirm if an existing data source can be used to provide all or part of the required information and satisfy data needs. To ensure an administrative data approach is comparable to a classical survey design (with a tailored questionnaire, sample selection scheme, controlled data collection, data processing etc.), measures are required to assess the quality of alternative data sources and determine how they fit together to answer statistical needs. Assessments that determine whether the data sources are fit for purpose enable sound decisions on whether using them is a cost-effective alternative to directly collecting new data ourselves.

We first investigate useful quality measures for statistical outputs that use administrative data. The most important motivation is the need to understand in detail, the risks and benefits involved when we are redesigning statistical outputs to make more use of administrative data. We must be able to assure users that our new designs will produce fit-for-purpose data that will meet their needs. Without a thorough understanding

of the sources of error affecting output quality, it is very difficult to evaluate whether the savings and efficiencies from the use of administrative data will be worth the potential loss in output quality.

There are many approaches to the quality assessment of administrative data (see Daas et al. 2010, 2012; Daas et al. 2011; Wallgren and Wallgren 2014; UNECE 2011b). However, our work focuses on how the quality of statistical outputs that use administrative data can be assessed. To do this and to enable an administrative-data first production environment with easy reuse of data, we developed quality measures both for the administrative data we use as input to our statistical outputs ("input quality") and for the statistical outputs produced from administrative data ("output quality").

To assess the input quality of the administrative data entering a national statistics office, our framework includes qualitative as well as quantitative indicators based on the quality concepts given by Daas et al. 2010. These indicators have also been included in Stats NZ's meta-information template for evaluating new data sets (an online version of the template can be found in Statistics NZ 2016a). As for indicators of the output quality, our framework is influenced by the work by Burger et al. (2013). They investigated the use of administrative data to avoid unnecessary reporting burden on businesses and provided quality indicators for statistical outputs that use mixed sources of data. Other agencies have adapted or developed frameworks for measuring quality. Examples include Australian Bureau of Statistics Data Quality Framework (Australian Bureau of Statistics 2009) and Statistics Canada's Quality Guidelines (Statistics Canada 2009). These are of limited practical use in determining what the quality of our outputs will actually be since quality indicators are not explicitly defined. The United Kingdom's Office for National Statistics' Guidelines for Measuring Statistical Quality (Office for National Statistics 2013) provides quality indicators useful in the assessment of output quality.

Li-Chun Zhang's two-phase life-cycle model for integrated statistical microdata (Zhang 2012) helpfully expands the TSE paradigm in a way that makes it applicable to mixed-source statistical outputs. We adopted this model for the first two phases of our framework because its systematic list of the ways in which error arises in statistical outputs, is applicable to designs using traditional survey methods, administrative data, and mixtures of the two. This enables us to compare the various sources of error affecting rival statistical designs aiming to produce the same statistical output with different mixtures of input data. The two phases cover the processes used to create a final unit record data file. In Phase 3 of our framework, the errors that arise from the estimation process are considered, alongside the evaluation and correction of errors. Our framework also gives a useful vocabulary for this error and statistical design comparison, which can be explained to non-methodologists with limited familiarity of administrative data, TSE, or both. It also provides a structure to organise the practical knowledge from processing which analysts have about the sorts of errors that affect their statistical output.

One major attraction of the framework is that it explicitly distinguishes "input quality" and "output quality". Input quality, or how well a single data source meets its original purpose, is particularly relevant to Stats NZ's aim of reusing data and matches well with our existing meta-information template for evaluating new data sets (Statistics NZ 2016a). The sources of error under Phase 1 of Zhang's model are a result of the initial data

collection and processing, and will flow through into any use of the data in the production of a statistical output. Phase 2 errors relate to using these source data sets to produce a *particular* statistical output. They depend on the desired outputs and the design under consideration. When a new statistical output is being designed, previous Phase 1 evaluations of the data sources under consideration can be reused. Some additional Phase 1 evaluation work may still be required but the previous Phase 1 evaluation still saves a lot of information gathering and initial investigations.

To practically apply the concepts in the framework with an overall aim of identifying and understanding all sources of error that affect a statistical output, an organised list of the sources of error, and at least a rough idea of their relative magnitude, is essential. Rigorous measurement is often difficult, but is necessary for design, process monitoring, and reporting to users.

2.2. *Components of the Framework*

The three phases of the quality assessment framework are separated so we can understand the effects of data processing on the quality of the statistical output.

2.2.1. Phase 1

Figure 1 shows a flow chart illustrating Phase 1 of the quality assessment framework from Zhang (2012). The flow chart is similar to those in works such as Groves and Lyberg (2010, Figure 3). The main difference is that Zhang (2012) uses more generic terms that apply to both survey and administrative sources. The most important aspect of this diagram is the flow (shown by arrows) between the rectangular boxes from the initial target concept and target set to the final data stored. At each stage errors can arise (represented by the ovals). Throughout the Phase 1 assessment process, it is the target concept and target set (intended by the organisation that created the data) that we must assess against. Using someone else's data means we cannot control any of their decisions

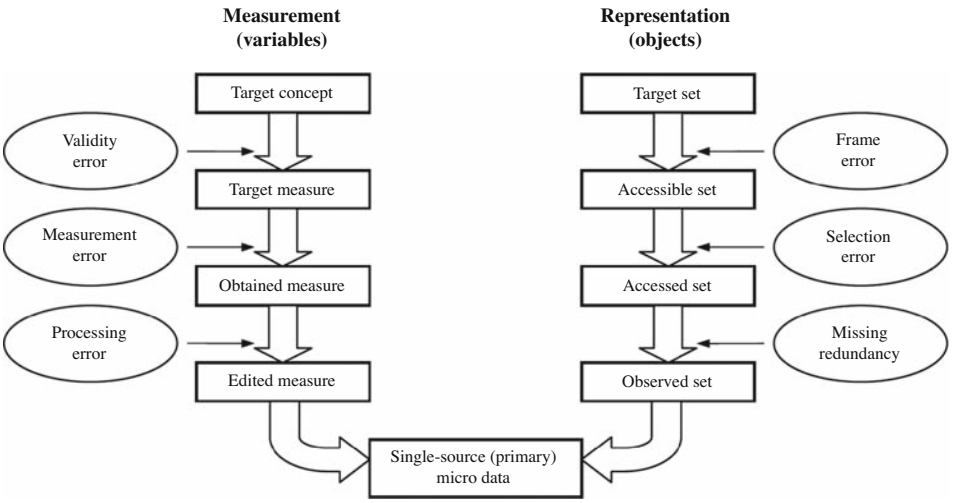

Fig. 1. Phase 1 of the quality assessment framework (Zhang 2012)

on measurements and populations. We need to understand their design decisions so we can determine what to do to turn their data into the information we want. Definitions of terms in Phase 1 of the quality assessment framework are given in Appendix A.

Although the framework applies to both administrative and traditional survey data, different types of errors tend to dominate. Our test cases (Sections 3–5) show that administrative collections, particularly for business data, usually have very good alignment between the target concept and the measure used to capture it. For instance, the value of sales taxes paid by a retail business in a given calendar month is objective and well defined, so validity errors are small compared with conceptually complex individual survey questions about ethnicity or well-being.

The distinction between frame error and selection error can be confusing, especially when the administrative data have been designed with restrictions already in mind. An example of these errors is in the recording of transaction events. Suppose that a retail chain wants to produce statistics on the transactions across all its stores, but the system they use can only record purchases that use electronic cards. Cash transactions could be said to be "inaccessible" since they will never be in the database. On the other hand, if a store manager forgets to run the reporting tool for a week, the transactions missing from the data set due to that mistake will be selection errors: they were accessible, but were not accessed.

Phase 1 of our framework provides some measures for each of the identified error components of a data source. Examples of quality measures for measurement error include the item imputation rate of a variable and the lag time between the reference period and the time of receipt of the data source. Quality measures for frame error include undercoverage and overcoverage. In instances where a metric assessment is not possible, the framework will assist in identifying processes where potential errors may arise so these can be addressed during the design of the output statistic. More complex measures are also possible: Bakker (2012) used a structural equation model to assess bias arising from measurement errors from various data sources, and Scholtus and Bakker (2013) used a simulation study to test the robustness of the model to additional components of measurement error as well as selection errors.

See Appendix A for a list of quality measures and indicators for Phase 1. Note that we focus on administrative data use and the new potential for errors it raises, so our examples are centred on administrative data. Many of the same or similar measures are also relevant to survey data, or can be made so with small modifications.

2.2.2. Phase 2

Phase 2 of the quality assessment framework is illustrated in Figure 2. Phase 2 focuses on errors that arise when data sets from several sources are integrated to produce an output that meets a certain statistical purpose. Phase 2 also includes errors from an output produced mainly from a single administrative data set.

In this phase the reference point is the statistical (target) population we would ideally have access to and the statistical (target) concepts of the units we want to measure in the target population. In practice, it takes some care to precisely define the true targets. In an established survey design, for instance, sometimes there is not a clear distinction between the sampling frame developed for practical purposes and the true target population. Some of the errors that arise during Phase 1 can also propagate through to the final output, and

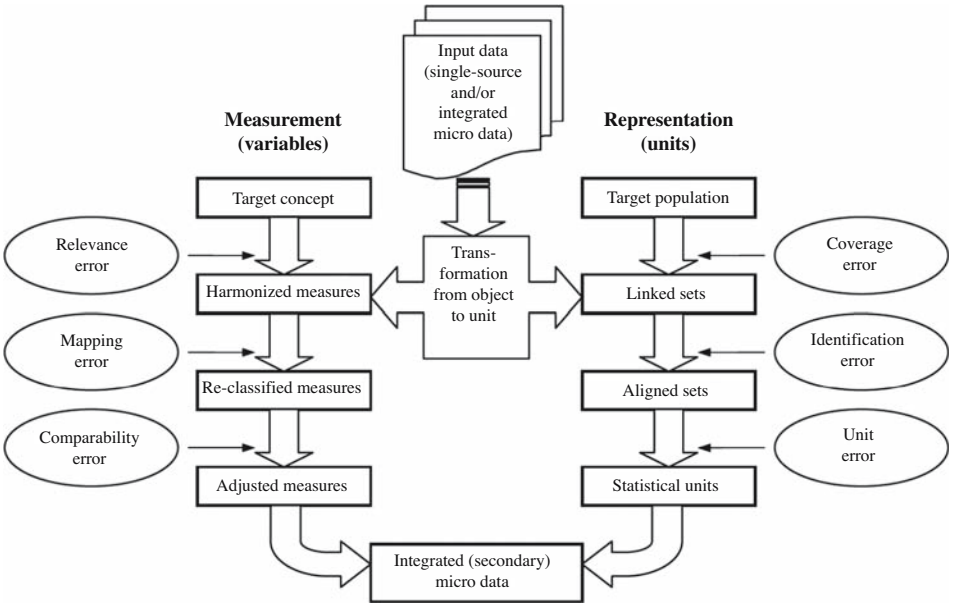

Fig. 2. Phase 2 of the quality assessment framework (Zhang 2012)

the flows in the figures are not necessarily directly related to specific or sequential steps in a statistical production process. See Appendix B for a definition of terms in Phase 2 of the quality assessment framework.

If we again look at sales tax data from our tax agency and consider Phase 2; the sales taxes paid by a business in a given month might not actually correspond to the true sales in that month, which is generally what the statistical output is more concerned with. Depending on the details of the tax system, the sales taxes may be paid in the month after the actual sale of the goods, or there may be sales taxes paid on items at the time they are brought into the store rather than at the time they are actually sold. These mismatches give rise to errors, specifically mapping errors, in the "measurement (variables)" column (see Figure 2).

For the "representation (units)", other difficulties may be encountered. If a particular branch of a retail store franchise changes ownership, New Zealand tax reporting rules often result in an entirely new tax unit being created in the administrative source. From the point of view of the tax agency, as long as the tax owing is paid correctly, this does not result in any error in the units. From the point of view of a business survey, however, such changes result in the old unit being dropped from the survey (because it is marked as ceased in the tax data), while the new unit may not be selected because the magnitude of its tax activity is too small to qualify it for selection. In reality, the store continued in the same way, but our rules for creating and selecting survey units using the administrative data have introduced errors for "representation (units)".

Data integration is also an important source of error in Phase 2. Stats NZ's Integrated Data Infrastructure (IDI), discussed in more detail in Sections 4 and 5, combines information from several government agencies, and so create a central list of individuals who interact with the government. We found many cases where an individual has multiple

records with the same agency. In some cases these duplicates are flagged and linked by the agency, but if they cannot be detected and removed, we will be creating too many individual records. This in turn leads to problems when linking other data sets, because our record-matching process effectively has to choose between two different duplicate records when integrating to another data set, and the result will be rather unpredictable.

Phase 2 of the framework provides measures for each of the identified error sources of an integrated data set. These are listed in detail in Appendix B.

For some data sets there may be no linking, just processing and conversion from a raw input data set into an output. In these cases, measures such as link rates may not be very useful, but concepts such as coverage of the target population, and conversion of administrative objects to statistical units, can still be valuable.

2.2.3. Phase 3

The end point of Phase 2 is a unit record file containing a set of units and a set of variable values for each of these units. Typically, this unit record file is not itself the final output desired: this file is used to derive estimates, such as the unemployment rate, or the population for a range of geographic regions. We included Phase 3 in the framework to account for the processes and errors that can arise in the creation of these final outputs.

In our framework, Phase 3 includes the work done to evaluate or estimate the quality of the final outputs, taking into account all error sources. It also concerns the inaccuracies introduced by estimation methods that attempt to correct for sources of error that arise in the first two phases.

In a traditional survey context, the estimation process can include a variety of techniques, from simple sums and averages to complex model-based methods that use auxiliary data to calibrate or correct for selection or nonresponse biases. Other processes, like seasonal adjustment, may also be carried out to further correct or adjust final estimates. Seasonal adjustment could be thought of as a correction for relevance errors: for example our desired output could be to measure the underlying growth rate of, say, an industry sector, but our raw results only measure the combination of seasonal and true growth. Using seasonal adjustment we can estimate the size of the seasonal movements and remove them, but this process itself is subject to error.

It is difficult to create a generic set of steps for this phase, but the aim is to consider the estimation methods and the corrections that can be applied to deal with various sources of error. It should also include an evaluation of the estimated level of error remaining in the final estimates. Traditionally, the key indicator published by statistical agencies is the sampling error, but as we saw in Phases 1 and 2, there are many more non-sampling errors that, ideally, we would try to estimate. Ultimately, if this error estimation can be done for competing designs that are candidates to estimate the same underlying quantity or concept, then we can use these error estimates as the foundation for a cost/quality trade-off.

In the planning (or design) stage we can use comparative production costs and our best "guesstimates" of the total error in the desired outputs to determine whether an overall statistical design is well-motivated, compared with some other configuration and use of input data sources from Phase 2 (or just Phase 1 if no integration is considered as in many traditional surveys). Ideally, these estimations and evaluations would include optimising a multivariate TSE measure subject to cost restrictions, but this is unrealistic because of

complexities and possible shortcomings in assessing errors in Phases 1 and 2, including the use of different indicators with different scales. Instead, we advocate a practical approach that analyses different options in Phase 3 by appropriately weighting and comparing individual error components, and on that basis reach a decision on design. Although this approach will include a significant amount of judgements, we argue that this is a more thorough, methodological, and systematic way of achieving better-quality outputs than making the statistical-design decision based on a single (first) choice of data set. The approach enforces the practice of setting and thinking of competing objectives and comparing design options. This increases the chances of getting a good outcome that considers the cumulative effect of errors. It is in a phase that compares outputs (estimates) where evaluations that affect final choices on statistical design can be made.

Laitila and Holmberg (2010) give an example of how a Phase 3 comparison can be made. They suggest estimating the total error of an estimator from one data source by deriving lower and upper boundaries for a Total Mean Square Error (TMSE) measure. Let $\tilde{Y}_1(r,m)$ and $\tilde{Y}_2(r,m)$ denote an estimator of a parameter Y under representative (r) and measurement errors (m) from two different Phase 2 data set alternatives. By decomposing TMSE of the estimators with respect to the error sources and comparing them, there is guidance about which one is best. The derivation of $TMSE(\tilde{Y}_i(r,m))$ can be done in different ways (for example Biemer and Lyberg 2003; Biemer 2010; Laitila and Holmberg 2010; Smith 2011). Each of these approaches involves different assumptions, so the best choice depends on the particular case under consideration, the error sources, and the indicators available from Phase 1 and Phase 2. The derivation is an important and non-trivial step. A full consideration of the choice of a total mean square error measure would be too complex to include in this article, but one particular challenge is how to address the cases when randomisation theory does not easily apply to some data sources.

A very important aspect of these comparisons is the recognition that errors can be accounted for and potentially corrected within our estimation process. If we know from an independent survey (for example an audit sample) that a certain administrative data set has systematic undercoverage of our target population, then including this as a correction factor will bring our estimates closer to the true value. Ideally we would repair or eliminate errors at source or during the production of the unit record file, but this may be impossible. For instance, we can quite easily measure the rate of erroneous links in integrated data sets using a clerical sample, but searching through the entire linked file removing all incorrect links is impractical. Instead, we can use the error rate as an input to an estimation model that aims to produce corrected final estimates.

One possibility for estimation in these scenarios has come from work done at Stats NZ to estimate the size of the New Zealand resident population. In the past, we relied on data from the Census of Population and Dwellings, collected in a full-coverage survey of the country, but Bryant and Graham (2015) describe a Bayesian approach for population estimation from administrative data under coverage errors. By expanding that estimation approach to also include other types of errors identified by the framework, and comparing the uncertainties arising from different combinations of data, we have a tool to assist in making a better design choice. The error decomposition and the knowledge from the indicators in Phase 1 and Phase 2 can be used as inputs to the model and contribute to the uncertainty of estimates. Substantial further work is required to develop this idea, more

generally. There are other alternatives, but we believe that these ideas are a promising solution to dealing with errors in administrative sources that cannot necessarily be identified and repaired at a unit record level.

2.3. *Applying the Framework in the Design of a Statistical Output*

Our quality assessment framework is useful for designing a statistical output that considers either the use of a single data source or an integration of several data sources. Because the error categories and concepts in the framework are often quite abstract, it can take some time and effort for analysts to come to grips with them when they are first introduced. When the framework was developed, we carried out practical tests on various outputs together with subject matter analysts who are very experienced in the practicalities of working with their data and processing systems. Based on these tests, we settled on a rough sequence of tasks for applying our framework.

The evaluation process we developed has four steps.

- **Initial metadata collation:** Basic information is collected about each of the source data sets that contributes to the final output. The information relates to the source agency, purpose of the data collection, populations, reporting units, variables, timeliness of the data, and so on.
- **Phase 1 evaluation:** Errors occurring in Phase 1 of the quality framework are determined and categorised for each source data set. This involves detailed consideration of how the methods, purpose, known issues, and other aspects of the original data collection contribute to each of the specific error categories in the Phase 1 flow chart in Figure 1.
- **Phase 2 evaluation:** As in the Phase 1 evaluation, errors arising in Phase 2 of the quality framework are listed and examined in a similar way, taking into account the data set(s) being integrated to produce the final output. These errors are considered with respect to the intended statistical target concepts and population. The effects of Phase 1 errors on the creation of statistical units, or the particular details of the misalignment between concepts on different data sets, must be understood.
- **Phase 3 evaluation:** The previously identified sources of error are evaluated and further investigations are done into how they might be measured, controlled, or reduced. This may include developing and applying tailored quality measures and indicators. It also includes determining which sources of error should be minimised or which data source minimises a specific source of error so that the final statistical output is optimised. The error measurements may eventually feed into an estimation model that attempts to correct known data problems as much as possible.

Once this four step process is completed, the final outputs will include a list of the sources of error that affect both the input sources and the final statistical output, and corresponding measures to be used to assess the size or effect of each of these errors, where possible.

An important principle we agreed on during our tests was that the framework should be used in a flexible way. For some major design projects we might need to examine every detail of every type of error that might arise. In other cases, the goal might be to produce a basic report that explains the data under investigation in general terms and highlights its

main features and potential flaws. The effort spent on an evaluation should depend on the requirements, and the process should never be a routine box-ticking. A more detailed guide to the implementation of our quality assessment framework is available in Statistics NZ (2016a).

2.4. *Case Studies*

The following sections describe three projects that we used to test and develop our framework in practice.

The first, the Building Activity Survey redevelopment, was the first full redesign project carried out at Stats NZ where we tried to apply the process of mapping out the sources of error and systematically measuring or correcting for them. It is a relatively simple survey so was a good test case for administrative data replacement in business surveys, and balancing the cost savings made against any quality risks introduced.

The second case study relates to the measurement of personal income in household surveys, and the potential for using linked personal tax records to replace population census collection of this variable. We have included it because it is a good demonstration of the way our framework can capture, not only the issues with administrative data sets, but also categorise and understand the errors that arise in the traditional collection of this variable.

Our final case study is more of a work in progress, and examines the problem of population estimation from (imperfectly) linked administrative sources. It is important because it shows the value of the Phase 3 thinking that we have introduced in our framework and how new estimation models can take advantage of our systematic approach to the evaluation of error. It is also a good example of how we separate out the causes and effects of the various types of error that arise in a complex way when linking many data sets.

3. Case Study 1: Redesign of the Building Activity Survey

This case study is an example of a redevelopment project in which the aim was to reduce the amount of direct surveying through the use of administrative data. The process we followed is applicable to surveys in which the desired response variable can be approximated by using a statistical model based on a closely related administrative variable (or variables). A more complete statistical discussion of the changes made to the survey is available in Statistics NZ (2015a, 2015b).

3.1. *Introduction to the Building Activity Survey*

In the past, Building Activity Survey estimates were based on a stratified sample survey. The frame for this survey was of approved construction jobs from local government administrative data on building consents. It used a postal survey to gather information on the value of construction work completed each quarter. The redesign project aimed to replace our building activity sample survey with modelled values derived from the building consents administrative data and the relationship between building consents variables and building activity variables in past data. The redesign aimed to greatly reduce the number of survey forms posted out while maintaining or improving quality. The processing and analysis for the new survey also had to be built on a new software system

because the existing one used legacy tools and software that were very difficult to maintain. This meant that many of the software tools used for coding, editing, and estimation were being updated and improved.

To guide decisions on how much reduction in survey data would be possible without putting data quality at risk, we mapped out the sources of error affecting the old and new designs using our quality framework. The framework was applied in joint collaboration with experienced subject matter analysts. They helped us to understand the issues or problems they encountered in their existing design, including any issues that did not appear to easily fit one category of error or another. The outcome of these discussions was a detailed, organised list of the known sources of error, which was used to understand the impact of the new design.

3.2. *The Three Phases Applied to the Building Activity Survey*

The final outputs of the Building Activity Survey are quarterly tables of the dollar value of work put in place in construction jobs, broken down by several variables, including the type of building and the geographic region. Both old and new designs use building consents data, the source of the building type and other variables, and a survey to collect the value of construction work done in the previous quarter. The building consents data are the selection frame for the survey in both cases, but the new design only surveys large construction jobs.

The findings of the steps described in Subsection 2.3 are summarised below.

Initial metadata collation

Table 1. Summary of the initial metadata collation for the Building Activity Survey.

Information object	Building consents	Building Activity Survey data (before redesign)
Source agency	Local government authorities	Stats NZ
Purpose of data collection	Track new construction work and provide an early indicator of building activity planned throughout New Zealand.	Provide an estimate of the value and volume of work put in place on construction jobs in New Zealand.
Target set	All building consents issued by local authorities in New Zealand with a value of NZD 5,000 or greater.	All construction jobs in New Zealand active during the reference quarter.
Main variables collected	Consent date, consent value, building type, geographic location.	Dollar value of work put in place during the reference quarter.
Mode of collection	Administrative lists requested from each local authority on a monthly basis.	Quarterly (panel) sample survey using building consents as the sampling frame.
Time span of data	1998–present in the current form, historical data from 1965.	1998–present in the current form, historical data from 1965.

Phase 1 evaluation

Phase 1 in this example relates to the building consents data that provide the number and dollar value of construction jobs granted formal approval by local administrative authorities in New Zealand (which we publish as a separate economic indicator series), and the survey data collected by Stats NZ about the construction work actually carried out in each quarter.

Some of the errors arising in this stage are:

Table 2. Examples of the Phase 1 errors which arise in Building Consents and Building Activity Survey Data.

Error type	Building consents	Building Activity Survey data
Validity error	The target concept is the amount recorded on the consent so there is no validity error.	Work done on a job is a well-defined concept easy for respondents to understand, so the question is very closely aligned with our target concept, minimising validity errors.
Measurement error	Values are often rounded down by applicants because there is a financial incentive (lower fees) to have a lower consent value.	Respondents can make mistakes or provide round numbers.
Processing error	The main errors that occur in processing are related to coding: in some cases it is extremely difficult to determine the correct building type based on the description given on the consent.	Processing errors at this point are minimal because the variable – work put in place – is scanned from the survey form. There may be some errors in capturing the information from the form (for example messy handwriting).
Frame error	Cases of consents being given the wrong consent date, and thus not being included in the data extraction for a given month provided to us by the consenting authority.	Some construction work does happen on unconsented jobs, especially small ones.
Selection error	Every consent in the frame is included in the data by definition.	Actual sample drawn from the consents can be incorrect when building consents data contains errors. This results in a building job being placed into the wrong sample stratum. Sampling errors also arise from the random sample drawn in the lower value strata in the old design.
Missing/redundancy error	We do not get missing records on the consents because the consent itself is the unit of interest – any consent issued is available in the data.	Unit and item nonresponse are difficult to distinguish on the Building Activity Survey because (aside from simple confirmations of contact details and so forth) only one statistically important variable is collected on the questionnaire. There is about 10–15% nonresponse to the survey.

Phase 2 evaluation

Phase 2 of the framework applies to the combined unit record data, which is created using the combination of building consents data and survey responses. In both the old and new Building Activity Survey designs, errors arising from data integration are minimal because survey responses are very easily matched to the consent they relate to.

The most important error sources in Phase 2 arise from the corrections for nonresponse and erroneous respondent values, and from the modelling of building work done for small construction jobs below the cut-off and hence deliberately not sampled.

Errors due to editing and item imputation fit clearly into the category of comparability errors (Zhang 2012). The question of where to place the errors arising due to modelling of building work done is more complex. These errors could be considered to be similar to inaccuracies in item imputation, because technically and methodologically the solutions are very similar. Conceptually, however, the two sources of error are quite different. Imputation errors are the result of trying to correct for nonresponse for a variable already being collected using the final harmonised measures. Modelling is a conscious decision not to collect the data in this form and to instead use statistical techniques to convert administrative data into the harmonised measure. Thinking of modelling in this way, it is more closely aligned to mapping error, which arises when "turning primary input-source measures into harmonized measures" (Zhang 2012, 51).

Phase 3 evaluation

Other types of modelling and estimation do not fit this description quite as well, though, so to talk of "modelling error" generically is difficult. This is in part the motivation for introducing a Phase 3 to the framework, which includes modelling and estimation that takes the unit record data as an input and applies adjustments, models, or other techniques to derive final outputs. For the Building Activity Survey, the application of an estimator (Horvitz-Thompson) to the old sample design would be a Phase 3 activity and sampling errors arise as errors at this point.

Estimates for the new design, such as 'total value of work done in the quarter in all of New Zealand', are simpler: a basic sum is taken of the work done for all jobs, since every job has a modelled, surveyed, or imputed value for this variable. Another one of the errors arising in this phase, though, would be from the seasonal adjustment process.

3.3. Examples of Measures Developed and Used

The changes and quality impacts of the redesign fell into two main categories:

1. Changes to existing processes that are needed in the new design.
2. New methodology that would fundamentally change the way estimates were derived.

In the first category, the most important change was in the coding of building type. Building consent forms include an open-ended text box for applicants to describe the construction job they intend to carry out. Under the old system, all building consents were manually coded by a member of the processing team, which required large amounts of

effort and was quite a tedious job. The manual process was generally assumed to have very few errors, but had limited formal evaluation of the quality.

This manual process was replaced by automatic coding using a series of rules that looked for certain words and phrases in combination. To determine whether this new solution was of sufficient quality, the project team developed a set of criteria that focussed on the outcomes of the coding process and the impact on the final estimates. These criteria were used to check the new coding method against the original manual coding for the past ten years of data. They included:

1. Checks on the proportions of building consents coded into high-level categories (residential, non-residential, non-building construction): the criterion was that on a monthly basis the proportion of consents (by dollar value and count) coded to each category should fall within the lower and upper quartile of historical proportions.
2. Proportions by count and value at lower levels of classification, also using the upper and lower quartiles of historical coding as an acceptable range.
3. Specific building types or key words which were required to be coded in a certain way, such as "prison" and "relocated".

These criteria were developed along with the expert analysts, and an iterative process of refining the rules, checking for errors, and determining fixes was carried out until the quality standards were met. Further analysis included examining differences between the time series created using the old and new methodology, such as comparing seasonal adjustment diagnostics to determine whether the seasonal patterns and trends were significantly altered by the changes. In several cases, we found discrepancies due to problems with the codes originally assigned.

For the second category of changes, which included changes in the editing, imputation, sampling, and estimation methodology, we needed to set some criteria on the allowable differences between the old and new methodologies. In the old design we had traditional sampling error estimates, and comparing the old and new time series gave a measure of the accuracy of the new design.

One major challenge was in estimating both the estimation error in the model and the risk that changes in the construction sector might result in our model parameters being outdated and inaccurate. We addressed this challenge with two measures. First, we used bootstrap estimation to produce an estimate of modelling error. This estimate allowed us to fine-tune how many units would still need to be sampled to maintain a similar level of variance as the old design. Second, we ran simulations using the widest plausible range of the modelling parameters to understand the effects on the final time series. These results could then be used to make statements such as "the parameters would have to change by *x*% before the final estimates would fall outside the sampling error range in the old design". By comparing the historic changes over time in these parameters with the impacts of those changes, we could quantify the risks of our methodological changes.

We assessed potential imputation methods in a similar way. We used simulations to develop and test several methods and understand whether the changes would be significant compared with the old sampling errors and the new modelling errors. Having a clearly defined acceptable range of error for comparison was very useful, because it showed us that the choice of a simple imputation method would be more than accurate enough. This

saved us from creating a much more complex and slow solution that would have been less suited to the tools we had available in the processing system.

An important secondary benefit was that many of the measures were suitable to be published as quality indicators for users. We presently publish measures alongside each monthly and quarterly release (see, for example, the June 2015 release of Quarterly Business Activity Survey, Statistics NZ 2015d), which include:

- estimated modelling error,
- proportion by value that is modelled (rather than surveyed),
- imputation rates and proportions.

3.4. Broader Outcomes of the Application of the Framework

The workshops and discussions we conducted to understand sources of error and apply the quality framework to the building activity survey redevelopment, also had great benefits for the wider team. Methodologists and subject matter analysts understood clearly where the most critical and important errors might arise, and where more work was needed to control or measure potential new errors introduced, for example errors in the model. Comparing the old and new designs also helped us see existing monitoring or measures that were not effective or valuable and could be removed or replaced, and points where carrying out fixes or edits earlier in the process could reduce work. From a methodological point of view, we had a very detailed picture of the quality effects of different design decisions to guide investigations.

This work helped us to understand trade-offs and make better decisions about the design, and also to prove the value of the framework and demonstrate that we had quality risks under control. The study also had other beneficial side effects.

First, the detailed and comprehensive list of the sources of error affecting the new design compared with the old meant we could alleviate the concerns of users who relied on the existing survey. We clearly described and explained the problems of the old methodology and convinced users that although some time series might be changing, most of the change was due to fixing problems in the old design rather than introducing new errors. The way the framework forces the true statistical target to be clearly stated without reference to our existing measurement of it was very valuable in these discussions.

Second, analysts involved in the discussions understood the "TSE" mindset we brought and took a larger view of the proposed changes. At times, analysts who focus on certain parts of the process are very concerned with maximising the quality of the particular step they are responsible for. While this is not necessarily a bad thing, giving them the opportunity to follow the whole process while explaining the effect of their work on the final statistical quality helped us work together to determine where their effort might have the greatest impact.

4. Case Study 2: Evaluating Administrative Data for Personal Income

This section discusses an application of our framework to income data derived from combining the 2013 New Zealand Census of Population and Dwellings and Stats NZ's Integrated Data Infrastructure (IDI). As with the previous example (redeveloping the

Business Activity Survey), this project is an example of evaluating the potential of administrative data to replace specific survey questions. This may save costs and lower respondent burden. This example also helps to illuminate the challenges that arise from imperfect linkage and coverage of administrative sources and how the limitations of an administrative source can be weighed against the limitations of survey data. It is a good example of how comparing administrative and survey data can shed light on the limitations of both sources, as long as the limitations of each are clearly understood.

The IDI is a collection of linked data sets supplied by various government agencies (including Stats NZ). A key component of the IDI, called the 'spine', is a main data source to which all other person level data sets for research link. The target population for the spine is anybody who has ever resided in New Zealand. At present, the spine is a single list of individuals created by a union of tax, birth, and long-term visa records, to which all other data sets, such as income data from administrative sources, can be linked. For further details about the structure of the IDI, see Black (2016).

Stats NZ has linked 2013 Census records to the spine of the IDI as part of a Census Transformation Programme. The aim of this work is to evaluate the potential for administrative data sources to supplement or replace some census data in the future. The evaluations so far have been relatively quick and exploratory, and used a simplified version of the quality framework, primarily focusing on the coverage of administrative sources and the accuracy of the administrative variables assessed by comparison with census.

One of the most promising studies was a comparison of personal income data collected in the census with personal income from Inland Revenue (New Zealand's tax agency). Our framework can be used to understand the differences between census records and administrative data records on personal income. The results of the investigations can also help Stats NZ understand how to improve measurement of personal income in future censuses.

Initial metadata collation

The census personal income information is collected by two questions. The first asks which sources of income a person has received in the previous year, such as wages and salaries, investment income, or government benefits. The second asks for total gross income from all the sources in the previous year, with the respondent asked to tick the income band they fall into. The bands are roughly in NZD 5,000 increments (NZD 5,000–10,000; NZD 10,000–15,000, etc).

Information on personal income is available as administrative data from Inland Revenue. The data we have access to in the IDI includes tax returns for the self-employed and records from businesses that deduct tax directly from employees' regular pay (Pay As You Earn or PAYE tax), withholding payments (usually relating to contractor's pay), and registers of the main government payments, such as government pensions and unemployment benefits. Each earner in New Zealand has an individual tax number to which their various earnings and tax payments throughout the year are attached. Generally, anybody earning a wage or salary has the amount earned recorded in the tax system, and many government payments are also included. Investment income and superannuation or pension funds other than the main government pension, are not included.

Phase 1 evaluation

In this example the relevant sources of error are on the measurement side of the Phase 1 diagram (Figure 1). We briefly discuss the census income measurement, then move to tax data. Census is a single source, so we only need to consider the Phase 1 diagram.

The concept of personal income is clearly defined in the census in a technical sense: gross annual income from all sources. This is the target concept in the framework. The questions used to operationally collect this information are very well-aligned to this concept, although they make some compromises. In particular, the banded totals mean that the results are "blurred" compared with the exact, true amount.

Including bands rather than a specific dollar value makes it easier for the respondent to answer. However, many measurement errors are still possible (and observed) in the census responses. Item nonresponse is a problem, with only about 83 per cent of working-age respondents having a valid response. Other common measurement errors include:

- confusing gross with net income,
- recall errors when someone does not remember receiving income from a particular source,
- approximations made by respondents, such as rating up their latest pay cheque to an annual figure when they also received bonus payments or pay increases; or roughly mentally rounding their income and pushing themselves into a different band,
- proxy responses where a household member responds on behalf of another and does not precisely know how much money their housemate earns,
- mistakes when summing all sources, or when rating up the net pay cheque (for example the amount actually on a bank statement) to a gross amount.

Deliberate over- or under reporting of income is also a possible source of measurement error. In our investigations limited evidence of this occurs, in part because high incomes are covered by only a small number of wide income-band checkbox options. A general trend towards underestimation at all income levels seems to be stronger than any effect from deliberately overstating incomes at low levels.

Some potential measurement errors, such as respondents making factors of 100 errors when cents are or are not included, are reduced by using income-band tick boxes. The income bands also encourage more response, since people might know their income very confidently to within a few thousand dollars but not a precise amount. This is a good example of trading off different errors against each other: the bands result in some uncertainty, but also make it easier for people to respond and hopefully reduce measurement errors.

Processing errors are minor compared with other types of error because the tick-box responses are easy to code. In the 2013 Census, few important edits were made on responses, so processing errors are small contributors to the total error.

To assess the administrative income data, we made use of both Phase 1 and Phase 2 of the framework because income data comes from different sources and is not collected to measure personal income. The general process would be to understand the precise purpose of the administrative collection and determine what can go wrong within the administrative agency with respect to that purpose. For income, the variables we are concerned with are also the most crucial for administrative purposes. For instance, pension

payments are recorded so that the government can accurately track those entitled to receive payments, and company payroll tax is audited to ensure the correct tax amount is paid to the government.

If we want to understand all the sources of error fully, we need to look at all the particular administrative processes and constraints in different agencies. For example, do systematic errors (such as under reporting or processing mistakes) occur in pension data but not in personal tax returns? Earlier studies by Stats NZ suggest these errors are small and that administrative measures are very good measures of the administrative concepts (such as amount of pension paid, amount paid to an employee during a tax period). A significant practical issue is that processing for some sources like tax data takes considerable time, which means there can be a delay of several months (or more) until full records for a given date are available.

Phase 2 evaluation

Phase 2 of the framework focuses on errors that arise when data sets from several sources are integrated to produce an output that meets a certain statistical purpose.

Data integration is done using unique identifiers (tax numbers) from administrative data sets. In order to use administrative data to impute (or completely replace) the current census income question, we need to link the administrative data belonging to each individual to the right census respondent. In our prototype linkage we were able to link about 94 per cent of people to the IDI spine, with a false positive rate of about 0.7 per cent. Low-quality linking information (primarily names and dates of birth) is the main reason for not linking to the spine, but there are also several sources of undercoverage in the administrative data which mean that some people who filled in the census are not included in the administrative data at all.

One source of undercoverage is from individuals working in the "underground" market. Roemer (2002) integrated administrative data on workers earnings with earnings data from the United States Census Bureau's March Current Population Survey (CPS) and showed missing earnings from the administrative data. Earnings missing from the administrative data are exhibited across all wage sizes but are prominent across certain occupations.

In addition to linkage error, other coverage errors result from the mismatch between the tax population and the New Zealand census night population. People can be filing tax returns from overseas in some cases, causing overcoverage, although using tax data for only those census people who we link to the administrative data will help alleviate this problem.

On the other hand, people who receive income only from investments or untaxed sources may not appear in the tax data, causing undercoverage. The same error could be better described as a relevance error in some cases, such as if a person is present in the data but has no income recorded in the tax data. Unlike the census, the tax income measure does not include all sources of income.

Phase 3 evaluation

Given the high link rates, the crucial question about using administrative data over census income data is whether the conceptual mismatch between administrative data and the standardised statistical definitions results in more error than the problems caused by measurement error in the census. Note that errors in administrative data are considered to

be relevance errors in Phase 2 of the framework, while census errors are Phase 1 measurement errors that have flowed through to the final data.

These comparisons require an appreciation that the census (or any other existing source) is subject to its own errors, and that a difference between administrative data and an existing survey is not in itself proof of error in the administrative data. The comparison (as far as possible) must be between the statistical ideal and the different data sources we have. At times we consider census data to be a 'gold standard' whose results must be exactly reproduced by administrative data. In many cases the comparison with administrative data can help us understand the limitations of the gold standard. Some findings have already resulted in suggestions for improving the census questionnaire, where we can empirically show that many respondents are making similar mistakes.

For example, the missing sources of income in administrative data will cause a systematic underestimate of total income. Is this underestimate greater than that resulting from imperfect recall by census respondents? Using the linked census–IDI data, we compared the figures from the two sources. We found that even with some income sources missing from administrative data, census income responses were typically lower. This is a good argument for administrative data, along with issues of nonresponse and the lack of precision from the banded responses in the census which prevents analysis of income distributions in more detail.

It is useful to compare this investigation to that from the *Canberra Group Handbook on Household Income Statistics* (UNECE 2011a). The handbook contains detailed and careful descriptions of errors known to arise in measuring income. It allows us to clearly define our target concepts and populations so that we have a sound basis to compare against both census and administrative data. Generally, the sources of error mentioned in the handbook are similar to what we described above. Our framework puts these errors into a TSE and statistical design context in a systematic way, helping to make the evaluation of errors more practical and allowing for comparisons of the relative influence of different error sources.

5. Case Study 3: Population Estimation in New Zealand

The aim of Stats NZ's population estimates is to produce an accurate count of the number of people who usually live in New Zealand at a certain reference date. Our published population estimates are based on a variety of sources, including the five yearly Census of Population and Dwellings and some administrative data.

It is possible to use the IDI data directly (independently of the census) to produce estimates of the size of the New Zealand population. As part of our Census Transformation initiative, assessments and studies have been carried out to assess how accurately the population can be estimated from administrative sources (Gibb et al. 2016). The goal of this work is not to replace existing estimates (at least not yet), but to understand the limitations of the administrative data so that progress can be made towards improving our own processes in combining and using available administrative data. Another major goal is to identify which sources are more reliable, and whether there are any significant issues with the administrative data which could be fixed by source agencies.

This case study is included here to demonstrate how the three phase framework can be used to understand the complex interplay between coverage and linking errors. It is also a

demonstration of the start of what we see as a very promising path for continuous improvement of estimates derived from complex combinations of administrative data where there are many known and significant sources of error.

Initial metadata collation – definition of population

The New Zealand official Estimated Resident Population (ERP), defined as the "estimate of all people who usually live in New Zealand at a given date", was about 4.8 million people at the start of 2017 (Statistics NZ 2016d).

The target population of the IDI spine lists any person who has ever lived in New Zealand, and currently contains about nine million people (Black 2016). In order to generate a list of people who usually live in New Zealand that can be compared with the official ERP, we use a set of rules to restrict the spine to only those people who reside here as of a certain date. The resulting list is called the IDI-ERP. It is derived by selecting only those people in the spine who have shown recent activity in one of the administrative data sets linked to the spine. For example, those who have filed a tax return or have interacted with the health system during the previous twelve months, or who were born less than five years ago, are included in the IDI-ERP. The rules also take other information into account, such as death registrations and data about people who have travelled overseas and not returned.

Phase 1 evaluation

In this example, Phase 1 of the framework applies to each of the source data sets integrated in the IDI. For the purposes of population estimation, many of the administrative variables are not important, but some have measurement errors that affect estimates. First, measurement errors in linking variables such as names and dates of birth result in links not being made in the IDI processing. Errors in other major demographic variables (sex, ethnicity, and address) do not affect overall population counts but cause inaccuracies in subpopulation estimates.

New Zealand Customs data is a good example of the complex effects of measurement errors. Passengers crossing New Zealand borders complete arrival/departure cards that are collected by Customs officers. To identify that someone has left the country and later returned, their departure and arrival cards must be linked. Errors in scanning or recording names on these cards, or respondents writing incorrect or changed details (such as different spellings of a name transliterated from another language) can flow through to population estimation. In many cases, the records can be linked using passport numbers, but people may travel on different passports or renew their passports resulting in a different number that is not necessarily recorded in the administrative data.

Another crucial measurement error arises when we create subnational population estimates using administrative address information to assign people to different locations. Here many problems arise, such as out-of-date addresses, missing or poor-quality addresses that cannot be accurately geocoded to a certain location, and conflicts between different administrative sources that must be resolved. Some errors might be ignored by the agency: for instance, if the tax department wants someone's address, but enters the address of their accountant instead, this could be considered a validity error (depending on

what the tax agency's true purpose for collection is). Unless it results in difficulties with getting the right amount of tax paid by the person, they are unlikely to correct it. Similarly, if an agency's usual contact with an individual is by cellphone or email, they might never check if the address in the person's file is a valid one.

Phase 2 evaluation

The most obvious and significant sources of error are in the representation side of Phase 2. The linked sets are created by identifying records across multiple data sources that we believe belong to the same individual. Identification errors and unit errors are not an issue in this case because we are not creating new statistical units, but using the linked list directly as our list of units. The only error of concern is coverage error, but this can arise in many ways. In some cases, different sources of error can cause similar net effects on population counts and yet require different treatment.

A simple coverage error is when the available data does not include a person from the target population. The visa data we have access to starts from 1997, so if a couple moved to New Zealand in 1990 and only the husband has ever paid tax, the wife might not be included in the spine at all. Overcoverage is also possible because in some situations overseas residents could be paying tax to the New Zealand tax agency, and could look like an active resident under the IDI-ERP rules. Both the IDI-ERP time-band rules for activity and the time lag in updates of the spine, create issues in classifying a person as a 'usual resident' which causes coverage errors.

Linkage errors are also a major source of error. False negative links (for example when the link between someone's birth and tax records is not made due to a name change) effectively cause duplicates in the population. Some of these duplicates will be removed by the activity rules, but there are many complex possibilities. For some reason, if a person's (active) health record is linked to their birth record, but their (also active) tax record is not, two separate and active records for one person will exist.

False positive links can have different effects. If someone who has moved overseas is erroneously linked to an accurate health record of someone with a similar name, we may get overcoverage. But if a person is falsely linked to a departing immigration record, they may be removed from the population, causing undercoverage. Depending on how the rules for inclusion and exclusion are defined and which one takes precedence, linking errors between particular data sets will result in different effects.

Phase 3 evaluation – the estimation phase

The population estimation problem highlights that the end point of Phase 2 is the final integrated microdata, rather than the final estimates derived from this data. No matter how much effort we spend on improving our processes and data, our final integrated data set will have significant amounts of undercoverage and overcoverage. Therefore, we need to devise an estimation procedure that can correct these errors. Within Stats NZ's Census Transformation project, Bryant and Graham (2015) described one attempt to construct such a model using multiple administrative data sets. However, a conclusion of this work was the need for an independent sample survey to assist with coverage estimation.

Conceptually, the problem can be described by considering a large population, the union of the total coverage of the various administrative populations with the target resident population. The problem is to construct a model that describes which individuals on the administrative data make their way into the final data set, and which target population individuals are not represented in any of the administrative data sets. The processes that lead to somebody not being present in a given data set are part of the model, as are missing or erroneous variables (for example errors in ethnicity measurement). Parameters such as coverage rates can then be estimated and used to create a final estimate of the total population correcting for undercoverage, overcoverage, and other errors.

A complete model taking all error sources into account is still a work in progress, but this approach has a clear synergy with the error framework described in this article. If we have measures for some of the errors in the administrative data, these can be used to improve this model. Conversely, if a particular source of error (for example overcoverage in a particular administrative data set) is poorly understood, the model can give us some insight into how much uncertainty this causes in the final estimates. We can then make decisions about where to target our efforts; either by helping an agency improve their data, studying the coverage in more detail, or running coverage surveys to target measures for improving our overall estimates of the population size.

5.1. *Phase 3 and Continuous Improvement of Population Estimation*

The process for producing Stats NZ's official population estimates following a Census provides a good example of the usefulness of the Phase 3 concept. We can consider the final unit record census data after all editing, imputation, and other processing (the so-called "clean unit record file") to be the outcome of Phases 1 and 2 in the census, where error arising from combining administrative data and survey data have been incorporated in Phase 2. Most output tables produced for New Zealand's 2013 Census were based on tabulating the relevant variables from this clean unit record file.

However, in deriving the base estimated resident population counts, results of the Post-Enumeration Survey (PES) were used to correct and adjust for the estimated undercount in the Census. These final results do not come directly from data integration between the PES and Census unit record data, although data integration is a part of the process. Instead, the PES allows for coverage rates to be estimated, and these rates, as well as the raw counts from the Census data, are used as part of an estimation method that aims to produce more accurate counts of the population than the raw data alone. These estimates are updated in the period between population censuses using administrative sources such as birth, death, and immigration records, which are again incorporated into an overall estimation model.

Work continues at Stats NZ to improve population estimation and understand the sources of uncertainty in population estimates and projections. See for example Bryant et al. (2016), and Statistics NZ (2016c). Evaluations of errors in individual administrative data sets, the IDI linking process, census data, and coverage surveys can all be captured in a systematic way using our framework and this adds to our understanding of the quality of our final estimates. The models developed so far can be expanded to include new sources of error as we improve our understanding of the input data and linking processes.

6. Summary and Discussion

The quality assessment framework discussed in this article facilitates the reuse of both existing data and previous quality assessments. This was successfully demonstrated in the three case studies. The framework supports Stats NZ's goal to use administrative data first. The basic idea behind the framework is that with a clear understanding of both the limitations of all source data sets, and of the way errors propagate through our statistical production processes we can obtain a complete picture of the quality of the final output. Measuring an error is the first step to correcting it. We need to separate what the collecting agency has done from our own processes and what users intend to do with the data.

Phase 1 of the framework focuses on how well a data set meets its original, intended purpose. This information is valuable for anyone who wishes to investigate whether the data can meet any other needs. The framework can provide a common language for talking about data quality issues, and be a valuable decision-making resource for the organisation. This also applies for users outside Stats NZ, when data is reused and shared for research purposes. The framework and documentation is a pedagogical instrument to help explain a data source so that researchers and other users can determine how useful or suitable it might be for their own purposes. Besides helping users, applying the framework also raises internal awareness at Stats NZ of quality and sources of errors.

Phase 2 of the framework deals with the problems that can arise when integrating data sets from different sources during processes like transforming the original variables to match statistical needs and identifying and creating statistical units from integrated data sets. The reference point in the quality assessment in Phase 2 is the statistical population we would ideally have access to, and the statistical concepts we want to measure about the units in the target population. The measurement side in Phase 2 is concerned with how variables from each source data set are reconciled. This may differ in various ways from the target concepts. The representation side is about creating a set of statistical units from the objects in the original data sets.

Phase 3 of the framework focuses on estimation, design, and evaluation. The aim is to determine the data source(s) that can minimise the cumulative effect of errors on output statistics produced from integrated or combined data in Phase 2. If there are no integrated data sets but two or more alternative data sources (thus making Phase 2 redundant), then assessments from Phase 1 can be used to determine the best statistical design. The Phase 3 investigation can also provide a list of quality risks that need to be mitigated or checked over time to ensure the consistency of the resulting statistics. For statistics that the organisation can influence, this gives valuable input into which/how production/data generation processes can be improved.

The framework provides a list of measures and indicators that can be used to quantify key aspects of data quality. The measures can be used during the design phase of a survey to determine if survey needs have been met, during statistical production to monitor the process, and for dissemination to explain the quality of a statistical output to users. They can also be used to provide feedback on the improvement of the input data sets, including suppliers of administrative data. The measures do not cover every possible situation, but give a starting point and ideas for more detailed or technically complex measures that

could be developed for specific outputs. The framework also helps us prioritise further work so that investigations can be focused on the most crucial quality issues.

From our experience, the generic or standardised lists of measures can be very useful for initial input quality evaluation and for output reporting. Stats NZ also publishes output quality reports based on a standard list of required information (for example see Statistics NZ 2015c.)

When we make technical design decisions, we often have to develop more customised measures depending on the details of the design, population, and variables. Some of these measures are important for understanding output quality, such as measuring the uncertainty in modelled estimates instead of sample survey sampling errors. In many cases, specialised measures are needed to understand particular sources of error. We advocate flexibility in the measures and indicators used, and recognise that in some cases no satisfactory way exists to measure the effect of a given source of error.

An interesting future area of work will be to develop estimation models that can work in a positive feedback loop with our error assessments. The Bayesian estimation framework (Bryant and Graham 2015) may be one way to do this. We would like to be able to use our three phase quality framework to identify sources of error that could be built into the estimation model. The model would then give us a way to isolate the effects of each error on the final estimate, so that we can focus further improvements on the areas which have the largest impact, whether that is advocating for input data set improvements or processing improvements.

In trial applications like the Building Activity Survey, we found that our quality framework was a useful tool for teaching analysts about quality and TSE concepts. Analysts responsible for statistical production may be extremely knowledgeable about the types of error that occur in their data without having a methodologist's understanding of end-to-end effects of design and data quality. The framework allows their extensive practical knowledge to be translated into standardised and structured metadata, which other people can use to investigate data reuse. It also helps the analysts think about the connection between the initial user needs that are met by their output and the effects their decisions have on data quality.

To get a full picture of the quality of statistical outputs that reuse data not originally intended for official statistics, we also need to measure the improvements in processing costs, respondent burden, and other aspects of statistical production. Issues such as public attitudes towards data integration and the risk of relying on outside data suppliers also need to be considered by decision makers. We intend our framework to be an expanding information bank as Stats NZ gains access to more administrative data. A shared understanding of what data is useful for what purposes, captured with the help of our framework, will increase the pace at which both Stats NZ and data users can get the most value from new data sources and outputs.

Appendix A

Here are definitions of terms and quality indicators and measures useful to measure Phase 1 of the quality assessment framework.

Representation Side

Target set is the set of all objects the data producer would ideally have data on. This includes, for example, people, businesses, events, and transactions.

Accessible set is the set of objects from which measurements can be taken in theory.

Accessed set is the set of objects for which measurements are obtained in practice. For example, the electoral roll doesn't include people who fail to enroll despite being legally entitled, or whose forms get lost in the mail.

Observed set is the set of objects that end up in the final, verified data set after all processing by the source agency.

Frame error is the difference between the ideal target set of objects and the accessible set. These errors refer to objects that are inaccessible even in principle. In a survey context the accessible set is the sampling frame. For an administrative source, objects may be inaccessible for a variety of reasons.

Table A1. Quality Indicators for Frame Errors

Quality indicator / measure	Definition
Lag in updating population changes	Delays in registration.
Undercoverage	When units in the target population are not on the accessible set.
Overcoverage	When units in the accessible set are not in the target population.
Authenticity	Percentage of records in the administrative data with an incorrect identifier key, including records with multiple identification keys.

Selection errors arise when objects in the accessible set do not appear in the accessed set. For example, if a store manager forgets to run the reporting tool for a week then the transactions missing from the data set due to that mistake will be selection errors: they were accessible, but were not accessed.

Table A2. Quality Indicators for Selection Errors

Quality indicator / measure	Definition
Adherence to reporting period	Proportion of units that provide data for a different period than the required reporting period for the administrative data set. This may be due to lags, delay, or non-compliance with reporting period.
Dynamics of births and deaths	Changes in birth and death rates of units in the data over time.
Readability	Proportion of records that can be accessed using existing software for reading data.
Inconsistent objects/units	Proportion of units that are (and cannot be made) internally inconsistent. Examples are objects involved in non-logical relations with other (aggregates of) objects in the data source.

Missing/redundancy errors arise from the misalignment between the accessed set and the observed set. For example, errors where an agency mistakenly rejects or duplicates objects due to their own processing could mean that objects are missing from the data set even though correct data was received about them. This category of error exists so that such errors are kept distinct from reporting-type errors.

Table A3. Quality Indicators for Missing/redundancy Errors

Quality indicator/measure	Definition
Unit nonresponse rate	Fraction of units missing in the data source.
% of duplicate records	Proportion of duplicate records present in the data.
% of units that have to be adjusted to create statistical units	Proportion of units that have to be adjusted to create statistical units. For example, the proportion of data at enterprise group level, which needs to be split to provide reporting unit data.

Measurement Side

Target concept is 'the ideal information that is sought about an object'. The target concept is usually connected to the underlying purpose of the collection and may be quite abstract. Examples could include household income, political views, advertising effectiveness, or population counts.

Target measure is the operational measurement used in practice by a source agency to capture information. A target measure can include elements such as variable definitions, classifications, a questionnaire, or rules and instructions for people filing out forms.

Obtained measures are the values initially received for specific variables against objects in the data set.

Edited measure refers to the final values that are recorded in an administrative or survey data set, after any processing, validation, and other checks.

Validity error refers to misalignment between the ideal target information and the operational 'target measure' used to collect it. The error arising from the translation from an abstract target concept or 'the ideal information sought from the administrative data set about an object' to a concrete target measure that can actually be observed in practice, and does not include issues such as misunderstanding a term used on a form.

Table A4. Quality Indicators for Validity Errors

Quality indicator / measure	Definition
% of items that deviate from target concept definition	Fraction of items from the administrative data that deviate from the target concepts. In this context, 'items' are variables or fields entered on the final unit record data set.
% of items that deviate from Stats NZ/international standards or definitions	Proportion of items from the administrative data that deviate from Stats NZ / international standards or definitions.
% of inconsistent records	Proportion of units (or records) from the administrative data that violate logical, legal, accounting, or structural relationships between variables in a record.
% of items affected by respondent comprehension of questions asked in collection process	Proportion of items from the administrative data affected by the quality of questions in the data collection process.

Measurement errors occur when the obtained measure (the value actually recorded in the data set) differs from the measurement intended. These could include people misremembering details or interpreting the questions differently from how they were designed. In more automated administrative systems, such as electronic transaction records, measurement errors could include computer system problems that corrupt some values or introduce ambiguity.

Table A5. Quality Indicators for Measurement Errors

Quality indicator / measure	Definition
Item nonresponse	Fraction of missing values for a variable.
Percentage of records from proxies	Proportion of units from the administrative data whose data were provided by proxies.
Lagged time between reference period and receipt of data	Lapsed time between the end of the reference period and the time of receipt of the data source.
% of units in administrative data which fail checks	The proportion of units that fail one or more edits.

Processing errors arise from editing and other processing carried out by the source agency to correct or change the initial values received (the obtained measures).

This kind of processing is usually intended to improve the quality of the data with respect to the target concept, but it is important to understand how much improvement the processing makes, as well as any limitations introduced by the processing.

Table A6. Quality Indicators for Processing Errors

Quality indicator / measure	Definition
% of transcription errors	The proportion of units of a variable coded or recorded incorrectly.
Modification rate	The rate of editing changes done on a variable. Editing changes refer to changes to non-missing, values being changed to other non-missing values, which in most cases will be the result of editing.
Item imputation rate	Fraction of the values of a variable modified by editing and imputation by the administrative data provider.

Appendix B

Here are definitions of terms and the quality indicators and measures that apply to the error sources from Phase 2 of the quality assessment framework.

Representation Side

Target population is the ideal set of statistical units a final data set should cover.

The **linked sets** include all the basic objects from across all source data sets that are matched together to make base units. These units will not necessarily be the final statistical units of the output.

Aligned sets are the groups of base units which have been determined (after linking and other processing) to belong to each composite unit in a final output data set. For instance, we might create household units based on dwelling units and person units. In this case, the aligned sets could be represented by a table that contains all these relationships (for example Household 1 consists of dwelling A and persons X, Y, Z, Household 2 consists of dwelling B and person W, etc.).

Statistical units are the entities for which information is sought and for which statistics are ultimately compiled. These units can, in turn, be divided into observation units and analytical units (OECD, 2007).

Coverage errors are the differences between the units actually included in the linked data sets in practice (linked set) and the full set of units included in the (ideal) target population. Coverage errors can arise in several ways. For example, the data sets themselves may not cover the whole target population, or linking errors may mean some members of the linked sets are not identified. This error may also be caused by measurement errors. For example, if the date of birth variable on an administrative data set is not of good quality and we are filtering on age to select our population, we could end up with undercoverage even though the units are not missing from the source data.

Table B1. Quality Indicators for Coverage Errors

Quality indicator / Measure	Definition
Undercoverage	The proportion of units in the target population that are missing from the final data sets.
Overcoverage	Overcoverage occurs when units that are not in the target population are present in the final linked data.
Percentage link rate	The fraction of objects in each data set that can be connected with units in other data sets.
Proportion of duplicated records in the linked data	The fraction of units duplicated in the linked data.
False positive and false negative rates	False positives are record pairs deemed to be links but are actually true non-matches. False negatives are true matches that remain unlinked.
Delay in reporting	The time difference between the period each data set relates to and when you receive the final data set.

Identification error refers to the misalignment between the linked set and the aligned set. This type of error also includes situations where the target statistical units cannot be adequately represented using combinations of base units. For example, if we wanted to measure the economic activity of all manufacturing businesses by industry, we would ideally have separate statistical units to capture different types of manufacturing done by a single company. However, in practice we might have to define statistical units via legal entities. Changes in company or legal structures might result in statistical units being absorbed into others, despite no real-world change in economic activity occurring.

Table B2. Quality Indicators for Identification Errors

Quality indicator / measure	Definition
Proportion of units with conflicting information	Proportion of linked units that contain conflicts that need to be resolved during the production process.
Proportion of units with mixed or predominance-based classifications	When assigning objects from the input data sets to composite units, a single classification may have to be assigned to the composite unit based on the properties of the base objects that make it up. If the underlying units fit under one classification code, this decision will be simple. If they don't, the decision may be based on predominance, importance, or some other decision rule. However the decision is made, the units will not completely capture the properties of the real-world object they represent. A simple indicator of the quality of the final classification is the proportion of units for which such a decision must be made.
Rates of unit change from period to period	For many statistical outputs, the target population changes relatively slowly, so significant changes in the units in the input data sets may indicate quality problems with the data, linking, or other aspects of the process. This indicator is a simple measure of the rate of change of the population.

Unit errors are introduced when the final statistical units are created for the output data set. For instance, to create household units from the aligned sets of dwellings and people we must simultaneously decide which dwellings should have a household created, and which people should go into which household unit. Because statistical units may not correspond to any of the units in the source data, a variety of errors can arise at this stage.

Table B3. Quality Indicators for Unit Error

Quality indicator / measure	Definition
Proportion of units that may belong to more than one composite unit	The fraction of units that don't have a single clear composite unit to which they can be assigned without doubt. This could be units that cannot be assigned to any composite unit for some reason, or units equally likely to belong to two different composite units.

Measurement Side

Target concept is 'the ideal information that is sought about the statistical units'. The target concept is usually connected to the underlying purpose of the collection and may be quite abstract. Examples could include household income, political views, advertising effectiveness, or population counts.

The **harmonised measures** are the operational measures decided upon in the design of the statistical output to capture the target concepts. They include elements such as questions, classifications, and variable definitions. A common example would be a survey question aligned with a standard classification.

Re-classified measures are the values of the harmonised measures.

Adjusted measures refer to the final values in an integrated microdata, after any processing, validation, and other checks.

Relevance errors are errors at a conceptual level that arise from the fact that the concrete harmonised measure usually fails to precisely capture the abstract statistical target concept. For example, if we want to find out about personal income but decide that in practice we will only measure taxable income, this creates a relevance error since non-taxable income is part of our target concept but not our harmonised measure.

Table B4. Quality Indicators for Relevance Errors

Quality indicator / measure	Definition
Percentage of items that deviate from Stats NZ / international standards or definitions	Proportion of items in the final data set that deviate from Stats NZ / international standards or definitions.

Mapping errors arise from the transformation of variables on the input data sets into output variables that have been defined (the harmonized measures). These could include transformations like:

1. Reclassification from a non-standard classification, or coding a free text field.
2. Derivation of a numerical variable from a source data set, such as removing gross sales tax from a transaction value.
3. Modelling of a target variable using a combination of several variables on a source data set and some model parameters.

In each of these cases the value of the output variable may differ from the true value, and these differences are mapping errors.

Table B5. Quality Indicators for Mapping Errors

Quality indicator / Measure	Definition
Proportion of items that require reclassification or mapping.	Fraction of the variables on the input data set that requires transformation into relevant output variables.
Proportion of units that cannot be clearly classified or mapped.	Fraction of units of which values of its target variables cannot be clearly determined using classification rules.
Inconsistency of variable definitions in linked data	Differences in variable definitions across linked data sets.
Indicators and measures of modelling error	If the output design involves modelling a target variable using one or more of the original data set variables, this introduces errors. These errors can be measured, but the method to do this depends on the chosen model. Many indicators can be applied to statistical modelling. A few examples are goodness-of-fit tests (for example R-squared), confidence intervals of model parameters, or for Bayesian models credible intervals).

Comparability error arises from editing and other treatment methods applied to values obtained from reclassified measures – to correct for missing values, inconsistencies, or invalid values.

Table B6. Quality Indicators for Comparability Errors

Quality indicator / Measure	Definition
Proportion of units failing edit checks	The fraction of units, of the total units checked, failing one or more edits.
Proportion of units with imputed values	The proportion of units that have been imputed.

18. References

Australian Bureau of Statistics. 2009. *The Australian Bureau of Statistics Data Quality Framework*. Canberra: Australian Bureau of Statistics. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0. (accessed June 2013).

Bakker, B. 2012. "Estimating the Validity of Administrative Variables." *Statistica Neerlandica* 66: 8–17. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00504.x.

Biemer, P. 2010. "Total Survey Error: Design, Implementation and Evaluation." *Public Opinion Quarterly* 74: 817–848. Doi: http://poq.oxfordjournals.org/content/74/5/817.full.pdf + html10.1093/poq/nfq058.

Biemer, P. and L. Lyberg. 2003. *Introduction to Survey Quality*. New York: Wiley.

Black, A. 2016. *The IDI prototype spine's creation and coverage*. Available at: http://www.stats.govt.nz/methods/research-papers/working-papers-original/idi-prototype-spine. (accessed August 2016).

Bryant, J., K. Dunstan, P. Graham, N. Matheson-Dunning, E. Shrosbree, and R. Speirs. 2016. *Measuring Uncertainty in the 2013-Base Estimated Resident Population* (Stats NZ Working Paper No 16-04). Available at: http://www.stats.govt.nz/methods/research-papers/working-papers-original/measure-uncertainty-2013-erp.aspx (accessed March 2017).

Bryant, J. and P. Graham. 2015. "A Bayesian Approach to Population Estimation with Administrative Data." *Journal of Official Statistics* 31: 475–487. Doi: http://dx.doi.org/10.1515/JOS-2015-0028.

Burger, J., J. Davies, D. Lewis, A. van Delden, P. Daas, and J.-M. Frost. 2013. *Deliverable 6.5/2011: Final List of Quality Indicators and Associated Guidance*, Report for Work Package 6 of the ESSnet on the Use of Administrative and Accounts Data for Business Statistics. Luxembourg: Eurostat. Available at: https://ec.europa.eu/eurostat/cros/system/files/SGA%202011_Deliverable_6.5.pdf_en. (accessed August 2016).

Daas, P.J.H., S.J.L. Ossen, and M. Tennekes. 2010. "Determination of Administrative Data Quality: Recent Results and New Developments." In Proceedings of the Q2010 European Conference on Quality in Official Statistics, May 4–6, 2010. Available at: http://www.pietdaas.nl/beta/pubs/pubs/Q2010_Session34_presentation.pdf. (accessed June 2012).

Daas, P., S. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, A. Bernardi, F. Cerroni, T. Laitila, A. Wallgren, and B. Wallgren. 2011. *Deliverable 4.1: List of Quality Groups and Indicators Identified for Administrative Data Sources*, Report for Work Package 4 of the European Commission 7th Framework program BLUE-ETS. Brussels: European Commission. Available at: http://www.blue-ets.istat.it/index.php?id=7. (accessed December 2015).

Daas, P., S. Ossen, and M. Tennekes. 2012. *Deliverable 4.3: Quality Report Card for Administrative Data Sources Including Guidelines and Prototype of an Automated Version*, Report for Work Package 4 of the European Commission 7th Framework program BLUE-ETS. Brussels: European Commission. Available at: http://www.blue-ets.istat.it/index.php?id=7. (accessed December 2015).

Gibb, S., C. Bycroft, and N. Matheson-Dunning. 2016. *Identifying the New Zealand Resident Population in the Integrated Data Infrastructure (IDI)*. Available at: http://www.stats.govt.nz/methods/research-papers/topss/identifying-nz-resident-pop-in-idi.aspx. (accessed August 2016).

Groves, R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74: 849–879. Doi: http://poq.oxfordjournals.org/content/74/5/849.full.pdf+html10.1093/poq/nfq065.

Laitila, T. and A. Holmberg. 2010. "Comparison of Sample and Register Survey Estimators via MSE Decomposition." In Proceedings of the Q2010 European Conference on Quality in Official Statistics, May 4–6, 2010. Available at: http://q2010.stat.fi/sessions/special-session-34/. (accessed December 2015).

Office for National Statistics. 2013. London: Office for National Statistics. *Guidelines for Measuring Statistical Quality*. Newport: Office for National Statistics. Available at: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html. (accessed February 2017).

Organization for Economic Cooperation and Development. 2007. *OECD Glossary of Statistical Terms*. Paris: OECD. Available at: https://stats.oecd.org/glossary/index.htm. (accessed August 2016).

Roemer, M. 2002. *Using Administrative Earnings Records to Assess Wage Data Quality in the March Current Population Survey and the Survey of Income and Program Participation*. Maryland: U.S. Census Bureau. (Technical paper No. TP-2002-22). Available at: https://www2.census.gov/ces/tp/tp-2002-22.pdf. (accessed September 2016).

Scholtus, S. and B.F.M. Bakker. 2013. *Estimating the Validity of Administrative and Survey Variables Through Structural Equation Modelling: A Simulation Study on Robustness*. The Hague / Heerlen: Statistics Netherlands. (1572-0314, no - 201302).

Smith, T.W. 2011. "Refining the Total Survey Error Perspective." *International Journal of Public Opinion Quarterly* 23: 464–484. Doi: http://ijpor.oxfordjournals.org/content/23/4/464.short/ 10.1093/ijpor/edq052.

Statistics Canada. 2009. *Statistics Canada Quality Guidelines*. Ontario: Statistics Canada. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf. (accessed June 2013).

Statistics NZ. 2015a. *Implementing Classification and Other Changes to Building Consent Statistics*. Available at: http://www.stats.govt.nz/browse_for_stats/industry_sectors/Construction/building-consent-changes-2015.aspx. (accessed January 2016).

Statistics NZ. 2015b. *Methodology and Classification Changes to Value of Building Work Put in Place Statistics*. Available at: http://www.stats.govt.nz/browse_for_stats/industry_sectors/Construction/methodology-classification-changes-value-building-work.aspx (accessed January 2016).

Statistics NZ. 2015c. *Retail Trade Survey: September 2015 Quarter, Data Quality Section*. Available at: http://www.stats.govt.nz/browse_for_stats/industry_sectors/RetailTrade/RetailTradeSurvey_HOTPSep15qtr/Data%20Quality.aspx (accessed January 2016).

Statistics NZ. 2015d. *Value of Building Work Put in Place: June 2015 Quarter*. Available at: http://www.stats.govt.nz/browse_for_stats/industry_sectors/Construction/ValueOfBuildingWork_HOTPJun15qtr/Data%20Quality.aspx. (accessed August 2016).

Statistics NZ. 2016a. *Guide to Reporting on Administrative Data Quality*. Available at: http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx (accessed at August 2016).

Statistics NZ. 2016b. *Our Strategic Direction*. Available at: http://www.stats.govt.nz/about_us/who-we-are/our-strategic-direction.aspx. (accessed August 2016).

Statistics NZ. 2016c. *How Accurate are Population Estimates and Projections? An Evaluation of Statistics New Zealand Population Estimates and Projections, 1996–2013*. Available at: http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/how-accurate-pop-estimates-projns-1996-2013.aspx (accessed March 2017).

Statistics NZ. 2016d. *Standard for population terms*. Available at: http://www.stats.govt.nz/browse_for_stats/population/standard-pop-terms.aspx (accessed May 2017).

United Nations Economic Commission for Europe. 2011a. *Canberra Group Handbook on Household. Income Statistics*. New York and Geneva: United Nations. Available at: http://www.unece.org/index.php?id=28894 (accessed December 2015).

United Nations Economic Commission for Europe. 2011b. *Using Administrative and Secondary Sources for Official Statistics – A Handbook of Principles and Practices*. New York and Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf. (accessed December 2015).

Wallgren, A. and B. Wallgren. 2014. *Register-Based Statistics:Sstatistical Methods for Administrative Data*, 2nd ed. Chichester: Wiley.

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Received January 2016
Revised March 2017
Accepted March 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 513–531, http://dx.doi.org/10.1515/JOS-2017-0024

Comparing Two Inferential Approaches to Handling Measurement Error in Mixed-Mode Surveys

Bart Buelens[1] *and Jan A. Van den Brakel*[2]

Nowadays sample survey data collection strategies combine web, telephone, face-to-face, or other modes of interviewing in a sequential fashion. Measurement bias of survey estimates of means and totals are composed of different mode-dependent measurement errors as each data collection mode has its own associated measurement error. This article contains an appraisal of two recently proposed methods of inference in this setting. The first is a calibration adjustment to the survey weights so as to balance the survey response to a prespecified distribution of the respondents over the modes. The second is a prediction method that seeks to correct measurements towards a benchmark mode. The two methods are motivated differently but at the same time coincide in some circumstances and agree in terms of required assumptions. The methods are applied to the Labour Force Survey in the Netherlands and are found to provide almost identical estimates of the number of unemployed. Each method has its own specific merits. Both can be applied easily in practice as they do not require additional data collection beyond the regular sequential mixed-mode survey, an attractive element for national statistical institutes and other survey organisations.

Key words: Generalized regression; mode effects; selection bias; response mode calibration; counterfactuals.

1. Introduction

In mixed-mode sample surveys multiple modes of data collection are combined. Sequential designs apply different modes consecutively, approaching nonrespondents of one mode through a different mode. Each mode of interviewing has its own associated measurement error obstructing unbiased estimation of means or totals of true scores (Jäckle et al. 2010; Schouten et al. 2013; Buelens and Van den Brakel 2015). When different modes are administered in the same survey the total response consists of a mix of interviews obtained through the different modes, and associated therewith, a mix of mode related measurement bias. In surveys that are repeated over time, the mode composition of the mix may vary, and so may the overall measurement bias of estimated means and totals of survey variables. Confounding of true change over time of a survey statistic with change in mode composition limits the usefulness of mixed-mode surveys (Buelens and Van den Brakel 2015; Cernat 2015).

Despite this limitation, conducting surveys using a mix of interview modes has gained popularity in recent years. Benefits include cost – as a substantial number of respondents

[1] Statistics Netherlands, PO Box 4481, 6401 CZ Heerlen, The Netherlands. Email: b.buelens@cbs.nl
[2] Statistics Netherlands and Maastricht University, PO Box 4481, 6401 CZ Heerlen, The Netherlands. Email: ja.vandenbrakel@cbs.nl

are typically interviewed using cheap modes such as the internet – and more representative samples – as respondents who would refuse participation in one mode may be willing to respond in an other mode (De Leeuw 2005; Voogt and Saris 2005). A topical research question in the context of mixed-mode surveys is the influence of mode-specific measurement error on final survey estimates, see for example Lynn (2013); Vannieuwenhuyze and Loosveldt (2013); Schouten et al. (2013); Buelens and Van den Brakel (2015); Klausch et al. (2015).

In the present article two lines of research on measurement error are distinguished and their principles and merits are compared. Both are adaptations of the widely used general regression (GREG) estimator by which survey estimates of totals are expressed as $\sum_k w_k y_k$, a weighted sum of the observations y_k (Särndal et al. 1992). One approach seeks to adjust the survey weights w_k and is aimed at stabilizing total measurement error in repeated surveys (Buelens and Van den Brakel 2015). The other approach leaves the survey weights unchanged and instead proposes adjustments to the observed values y_k in order to remove measurement error (Suzer-Gurtekin et al. 2012; Suzer-Gurtekin 2013). While the two methods are motivated differently, it is shown in this article that both methods are identical for a certain parameterisation when the underlying assumptions are met. The two methods are explained and applied to a series of 36 months of the Dutch Labour Force Survey, in which three interview modes are used. This analysis provides insight into the extent to which sequential mixed-mode surveys that are repeated over time are susceptible to variations in mode composition, and how the estimation method can be adapted accordingly. Both methods are applicable to sequential mixed-mode designs and do not require the collection of additional data either by expanding the questionnaire with additional questions, for example Vannieuwenhuyze and Loosveldt (2013), or by re-interviewing respondents, for example Schouten et al. (2013).

This article contributes to the existing literature on inference with mixed-mode surveys by analytically establishing the conditions under which two different inference procedures for sequential mixed-mode surveys are equivalent. This sheds additional light on the properties of both methods. The results are illustrated by applying both methods to a series of monthly samples of the Dutch Labour Force Survey.

In Section 2 the inference methods under consideration are detailed and their assumptions discussed. Section 3 provides details of the Labour Force Survey (LFS) in the Netherlands. The results of applying the different methods to the LFS are presented in Section 4. Section 5 concludes the article.

2. Methods of Inference

2.1. GREG Estimation

The general regression estimator (GREG) of the total t_u of a variable u can be written as a weighted sum

$$\hat{t}_u = \sum_{k=1}^{n} w_k u_k \tag{1}$$

with u_k the values of u for survey respondents $k = 1, \ldots, n$ and w_k weights. The weights account for unequal inclusion probabilities associated with the sampling design and they

correct for selective nonresponse by calibrating the weights such that the sum over the weighted auxiliary variables equate the known totals in the population. Details of this method including variance estimation can be found in Särndal et al. (1992).

2.2. *Response Mode Calibration*

This paragraph summarizes an approach proposed by Buelens and Van den Brakel (2015) called response mode calibration. When measuring the variable u through a survey mode m, the measurement can be modeled as

$$y_{k,m} = u_k + b_m + \epsilon_{k,m} \tag{2}$$

with $y_{k,m}$ the observations through mode m of the true values u_k, b_m the systematic effect of mode m and $\epsilon_{k,m}$ random mode dependent error components with expected values equal to zero.

Inserting (2) in the GREG estimator for the observed total and taking the expectation with respect to the measurement error model gives

$$\hat{t}_y = \sum_{k=1}^{n} w_k y_k = \hat{t}_u + \sum_{m=1}^{p} b_m \hat{t}_m \tag{3}$$

with $\hat{t}_m = \sum_{k=1}^{n} w_k \delta_{k,m}$ and $\delta_{k,m}$ a dummy indicator equal to one if unit k responded through mode m and zero otherwise.

While the parameter p ordinarily corresponds to the number of modes applied in a survey, other conceptualizations are possible. For example p can refer to the number of interview strategies that are believed to have different associated measurement errors. Additionally, p can refer to a cross-classification of response mode or strategy, and other categorical auxiliary variables; this allows for modeling of a different measurement bias for different population subgroups.

Equation (3) expresses that the estimate of the true total, $\hat{t}_u$, is observed with error $\sum_{m=1}^{p} b_m \hat{t}_m$, a combination of mode-dependent biases. The quantity $\hat{t}_m$ can be interpreted as the estimated number of units responding through mode m in the population under the given survey design. Of the quantities in Equation (3), only $\hat{t}_y$ and $\hat{t}_m$ are observed, $\hat{t}_u$ and b_m are not.

The issue addressed by the method of response mode calibration is that in repeated surveys the response mode composition may vary between editions, leading to varying $\hat{t}_m$ and hence to a varying bias in the observed totals $\hat{t}_y$. This problem can be prevented if the bias term in Equation (3) is rendered constant. This is achieved by applying a response mode calibration as proposed byBuelens and Van den Brakel (2015). The response mode composition is calibrated to a fixed distribution, effectively requiring the $\hat{t}_m$ to equal given values. As this is exactly what the GREG estimator achieves for the other auxiliary variables, the response mode calibration is straightforwardly implemented by extending the underlying regression model with an additional covariate, response mode, and defining arbitrary but fixed response mode levels $\{\Gamma_m\}_{m=1,\ldots,p}$.

The resulting mode calibrated GREG estimator is

$$\hat{t}_y^c = \sum_{k=1}^{n} w_k^c y_k = \hat{t}_u^c + \sum_{m=1}^{p} b_m \hat{t}_m^c = \hat{t}_u^c + \sum_{m=1}^{p} b_m \Gamma_m \tag{4}$$

with w_k^c the weights resulting from the mode calibrated GREG – compare to expression (1) – and $\hat{t}_u^c = \sum_{k=1}^{n} w_k^c u_k$. By construction of the mode calibrated GREG, $\hat{t}_m^c = \Gamma_m$ for all m. The b's are the regression coefficients of response mode in the GREG weighting model. The variance of the mode calibrated GREG is obtained using the ordinary GREG variance estimation (Särndal et al. 1992), applied as if the calibration levels are known population totals. While the calibration levels Γ_m can be chosen arbitrarily, it is recommended to choose levels close to those realized in the survey. Otherwise the estimator becomes inefficient, inflating the variance unnecessarily as follows from the simulation conducted by Buelens and Van den Brakel (2015). If long-term systematic changes of the realized mode composition occur, the calibration levels Γ_m can be changed and past results can be recalibrated to the new levels to sustain a consistent time series.

A strong assumption of this method is that $\hat{t}_u = \hat{t}_u^c$. This assumption is fulfilled if response mode does not explain any selectivity of the response beyond that explained by the other covariates in the regression model of the GREG. One of the approaches to verify this assumption is suggested by Buelens and Van den Brakel (2015) and consists of applying both the usual and the mode calibrated GREG to register variables known for the survey respondents. As these variables are measured independent of the survey, mode calibration should have no effect as there cannot be a mode-dependent measurement error.

In summary, response mode calibration replaces the original weights w_k in Equation (1) by their mode calibrated version w_k^c and leaves the observations y_k unchanged. Measurement errors are not corrected for, they are merely balanced to render the total measurement bias constant across survey editions.

2.3. *Measurement Error Correction*

When measurement errors are estimated explicitly, estimates can be corrected towards a benchmark survey mode. A model based approach predicting counterfactuals – responses that would have been obtained through another mode than that actually used – has been proposed by Suzer-Gurtekin et al. (2012) and Suzer-Gurtekin (2013). A slightly modified version of their method is implemented here and summarized as follows.

Combining the linear model underpinning the GREG estimator, $u = \beta X + e$, with Equation (2) results in the regression model

$$y_{k,m} = \beta X_k + b_m \delta_{k,m} + \tilde{e}_{k,m} \tag{5}$$

with $\tilde{e}_{k,m} = \epsilon_{k,m} + e_{k,m}$, β a vector of regression coefficients for covariates other than mode, and b_m the regression coefficients for the modes $m = 1, \ldots, p$. If response mode does not explain any selectivity beyond that explained by the other covariates X, the coefficients b_m equal the measurement errors of the modes. This assumption is the same as the assumption required in the mode calibration approach.

In contrast to the mode calibration approach, the correction approach seeks to estimate the unknown parameters b_m explicitly. Fitting Model (5) using least-squares regression

results in estimates $\hat{\beta}$ and $\hat{b}_m$ of the regression coefficients. The estimated regression coefficient $\hat{b}_m$ is at the same time an estimate of the measurement error b_m in Equation (2).

Model (5) is taken to be linear here for fair comparison with the mode calibration method which employs linear models too. If desired, one could choose a generalized linear model such as a logistic regression model.

Suzer-Gurtekin et al. (2012) and Suzer-Gurtekin (2013) propose to use the fitted model to predict individual observations under an alternative mode, counterfactuals,

$$\hat{y}_{k,m}^{m'} = \hat{\beta} X_k + \hat{b}_{m'} \tag{6}$$

which can be calculated for every m' in $1, \ldots, p$. The estimate $\hat{y}_{k,m}^{m'}$ is the predicted outcome of observing unit k through mode m' while it really was observed through mode m. In this article, counterfactuals are instead obtained in a corrective rather than a predictive manner,

$$\hat{y}_{k,m}^{m'} = y_{k,m} - \hat{b}_m + \hat{b}_{m'} \tag{7}$$

which again can be computed for all m and m' in $1, \ldots, p$. The estimated measurement error of the original mode is now removed, and that of the alternative mode is added to the observations. The counterfactuals computed through (7) are closer to the initial observations than those obtained through (6).

Using the counterfactuals, a mode specific estimate of the total is obtained as

$$\hat{t}_y^{m'} = \sum_k \delta_{k,m'} w_k y_{k,m'} + \sum_k (1 - \delta_{k,m'}) w_k \hat{y}_{k,m}^{m'}, \tag{8}$$

the sum over measurements of units observed in mode m' and counterfactuals of units observed in other modes. This estimator would typically be applied if one of the modes is the preferred mode towards which other measurements are benchmarked.

Using the counterfactuals as obtained in (7), Expression (8) can be written as

$$\hat{t}_y^{m'} = \sum_k w_k \hat{y}_{k,m}^{m'}. \tag{9}$$

The variance of $\hat{t}_y^{m'}$ has two sources, associated with the two terms in Equation (8). The first source is the design variance due to sampling. The second is model-based and due to model uncertainty. Suzer-Gurtekin (2013) adopt a multiple imputation approach to capture the model induced variance. Here, a bootstrap approach is followed instead, capturing the design and model variances simultaneously. Through repeated sampling with replacement from the original sample, a bootstrap distribution of $\hat{t}_y^{m'}$ is obtained, from which the total variance is calculated.

If there is no benchmark mode or preference for one mode specifically, different counterfactuals can be combined. As the models are linear this can be done at aggregate level,

$$\hat{t}_y^{combi} = \sum_{m=1}^{p} \alpha_m \hat{t}_y^m \tag{10}$$

with α_m mixing coefficients summing to one, defining the mode composition of the final estimator. The variance of this combined estimator is again estimated through bootstrapping.

Using (9) and Expressing (10) as

$$\hat{t}_y^{combi} = \sum_k w_k \left(\sum_{m'=1}^{p} \alpha_{m'} \hat{y}_{k,m}^{m'} \right) \tag{11}$$

it is clear that this estimator involves adjustments to the observed values y_k and leaves the original weights unchanged. For the calibration estimator (4) the reverse holds: the weights are adjusted and the measurements are kept unchanged.

Suzer-Gurtekin (2013) propose to choose values for α_m through an optimization procedure, for example minimizing the variance or MSE. In the present study, a comparison with the calibration approach is the primary goal. Therefore the most sensible choice is to choose the mixing proportions α_m such that they correspond to the calibration levels Γ_m in Subsection 2.2. For each mode m, α_m and Γ_m are chosen so that $\alpha_m = \Gamma_m/N$ with N the known population total. With this choice, the calibration estimator (4) and the correction estimator (10) are both composed of the same mixing composition of modes, facilitating comparative analyzes.

2.4. Relation Between the Two Methods

When setting the levels in the calibration approach to Γ_m and the mixing proportions in the correction approach to $\alpha_m = \Gamma_m/N$, it can be shown analytically that the two methods are approximately equal. The relation between the two methods has not been addressed before in earlier research.

Using Expression (7), the combined measurement error correction estimator (11) can be written as

$$\hat{t}_y^{combi} = \sum_k w_k \left(\sum_{m'=1}^{p} \alpha_{m'} (y_{k,m} - \hat{b}_{m(k)} + \hat{b}_{m'}) \right). \tag{12}$$

with $\hat{b}_{m(k)}$ denoting the actual response mode of respondent k.

According to measurement error model (2), $y_{k,m} - b_m = u_k + \epsilon_{k,m}$. Expression (12) can be elaborated as

$$\hat{t}_y^{combi} = \sum_k w_k \sum_{m'=1}^{p} \frac{\Gamma_{m'}}{N} (u_k + b_{m(k)} - \hat{b}_{m(k)} + \hat{b}_{m'} + \epsilon_{k,m}).$$

Taking the expectation with respect to the measurement error model gives

$$\begin{aligned} \hat{t}_y^{combi} &= \sum_k w_k \sum_{m'=1}^{p} \frac{\Gamma_{m'}}{N} (u_k + \hat{b}_{m'}) \\ &= \sum_k w_k u_k + \sum_k w_k \sum_{m=1}^{p} \frac{\Gamma_m}{N} \hat{b}_m \\ &= \hat{t}_u + \sum_{m=1}^{p} \Gamma_m \hat{b}_m. \end{aligned}$$

It is assumed that $\alpha_m = \Gamma_m/N$ and that $\sum_{k=1}^{n} w_k = N$. The former is a choice one can make, as said before. The latter equality holds if the weighting model at least uses the target

population size as an auxiliary variable, which is the case if at least one categorical variable dividing the population in two or more poststrata is included – which is almost always the case in practice. Finally the equality holds only approximately since $b_{m(k)} \approx \hat{b}_{m(k)}$.

Comparing Expressions (4) and (13) shows that both estimators are equal if in (4) the assumption holds that $\hat{t}_u^c = \hat{t}_u$, which is the case if the response mode does not explain any selectivity beyond that explained by the other auxiliary variables. In addition, it has been assumed that the GREG models used in both approaches are the same, that the model in Expression (5) is identical to the GREG model extended with response mode, and that it is this model that is used in both the calibration and correction approaches.

3. The Labour Force Survey

Statistics Netherlands conducts the Labour Force Survey (LFS) using a rotating panel design consisting of five waves. Since April 2012, data collection in the first wave follows a sequential mixed-mode strategy. Respondents are invited by regular mail to complete the survey online via the web. Nonrespondents are approached through telephone interviewing if they have a known telephone number and are a household with fewer than three people, and through face-to-face interviewing otherwise. Interviews in the second to fifth waves are conducted by telephone only – a contact telephone number is asked for in the first interview.

The LFS is a household survey. The target population is the non-institutionalized population aged 15 years or over residing in the Netherlands. The sampling frame is obtained from municipal registrations and consists of all known occupied addresses in the country. Each month, a stratified two-stage cluster design of addresses is selected, with strata formed by geographic regions. Municipalities are primary sampling units and addresses secondary. All households residing at an address, up to a maximum of three, are included in the sample and can be regarded as the ultimate sampling units. Each year approximately 140,000 households are in the LFS sample. In 2014, approximately 30,000 households responded via the web, 12,000 via face-to-face interviewing, and 9,000 by telephone. Not all of the web-nonrespondents are re-approached by a different mode; approximately 28,000 addresses are approached for face-to-face interviewing and 24,000 for telephone. These and other details can be found in reports published by Statistics Netherlands, such as the LFS 2014 report (Centraal Bureau voor de Statistiek 2015).

The response data are weighted to account for the survey design and for selective nonresponse using a GREG procedure, see Subsection 2.1. Weighting is conducted for each of the five waves independently. The GREG weighting model used for production of the regular unemployment statistics contains the variables listed in Table 1. All variables are categorical with the number of categories for each variable given in brackets. Age and sex are included as an interaction and the remaining variables as main effects. The variable 'registered unemployed' indicates registration with the Employment Agency and does not coincide with the LFS definition of being unemployed. Registration at the Employment Agency is not compulsory for the unemployed – it is required only to be eligible for unemployment benefits or to receive training or coaching. Given the survey design, it would be sensible to include a dichotomous variable indicating whether households can be

Table 1. Variables used in the regular monthly GREG estimates of the LFS.

Variable (number of categories)	Definition
Sex (2)	Male or female
Age (21)	Age classes
Household type (3)	With children, single-person, other
Region (43)	NUTS-3 areas and largest cities
Registered unemployed (5)	Duration of registration (0 meaning not registered)
Income class (6)	Standardised household income
Income type (3)	Salary, welfare benefit, unknown
Ethnicity (3)	Native, western immigrant, non-western immigrant

reached by telephone. Unfortunately no such population frame data are available; a third-party provides telephone numbers of households in the sample only.

The GREG results are used as input for a structural time series model. Through the use of such model, the precision of the estimates is increased as the model allows for borrowing strength from previous time periods. In addition, the model takes into account rotation group bias and discontinuities due to the survey redesigns in 2012 and before, see Van den Brakel and Krieg (2015). The structural time series model explicitly accounts for the systematic differences between the first and subsequent waves by benchmarking the outcomes for the second, third, fourth and fifth waves to the level of the first. The level estimates resulting from the first wave of the survey are therefore crucial. To avoid

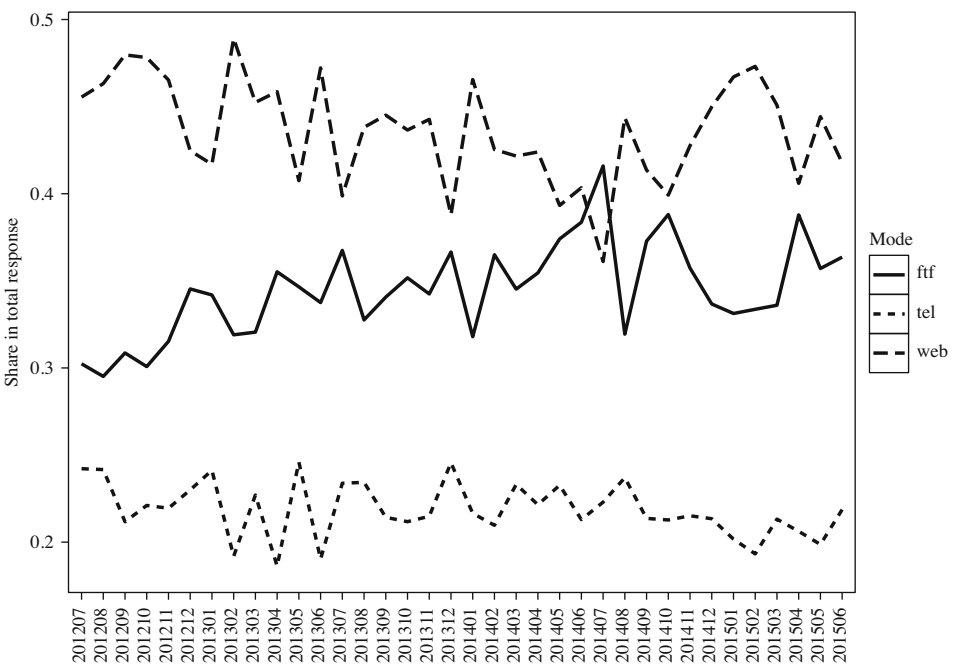

Fig. 1. Response mode composition of the LFS response during the 36 month study period; the three modes are face-to-face (ftf), telephone (tel), and web.

additional technical complications with this time series modeling approach, only the level estimates obtained in the first wave are used in this research.

In this article, first wave GREG weighted estimates from the LFS from July 2012 through June 2015 are studied, a period of 36 months. In the remainder of this article, this series is referred to as the regular approach – not applying any of two adjustment methods. Data collected in the subsequent telephone-only waves are not used in the present research. Issues pertaining to the redesigns of April 2012 and earlier are not discussed as they precede the study period. Executing the sequential mixed-mode strategy and applying the GREG procedure results in a weighted survey response composed of a mix of three modes, web, telephone and face-to-face. The composition varies from month to month and is shown in Figure 1. The share of telephone is rather constant. Face-to-face and web are exchanged in that months with relatively low web shares exhibit relatively high face-to-face shares and vice versa. The average mode composition over the study period is web 44%, telephone 22%, and face-to-face 34%.

4. Results

4.1. Response Mode Calibration

The calibration method of Subsection 2.2 is applied to the LFS, independently for each month of the 36 month study period. Four different calibration schemes are executed. The first, calBalanced, is the scheme that would ordinarily be applied based on recommendations in earlier research (Buelens and Van den Brakel 2015), taking the proportions for the three modes to be the averages over the study period, 44% web, 22% telephone, and 34% face-to-face interviews. The other three schemes are more extreme, each suppressing the contribution of one of the modes: two modes are calibrated to 45% each, and the third mode to ten per cent. These alternative schemes are executed to assess robustness and to illustrate the mode calibration technique.

The resulting estimates of the number of unemployed are shown in Figure 2. The mode calibrated estimates are presented relative to the number of unemployed estimated using the regular approach, which consists of the GREG estimates obtained from the survey weights, without applying mode-related calibration adjustments. The calBalanced alternative does not deviate a lot from the regular approach. The more extreme alternatives exhibit larger deviations. Estimates that are five per cent higher or lower than the regular estimates occur often. Table 2 lists the estimated monthly number of unemployed averaged over the whole study period. The calLessWeb and calLessFtf approaches result in systematically lower estimates, while the calLessTel results in a systematically higher estimate. Under the assumptions of the method, these differences are due to measurement error. In this case the telephone mode must measure lower than the other two modes.

The estimated standard errors of the point estimates are obtained with the standard analytic approximation for the variance of the GREG estimator and are shown in Figure 3 and are relative to the standard errors of the regular approach. The errors of the calBalanced approach are similar to those of the regular approach. The alternative approaches have larger standard errors, as expected, as they use the sample in a less efficient manner due to up or down weighting of respondents of certain modes. Of the

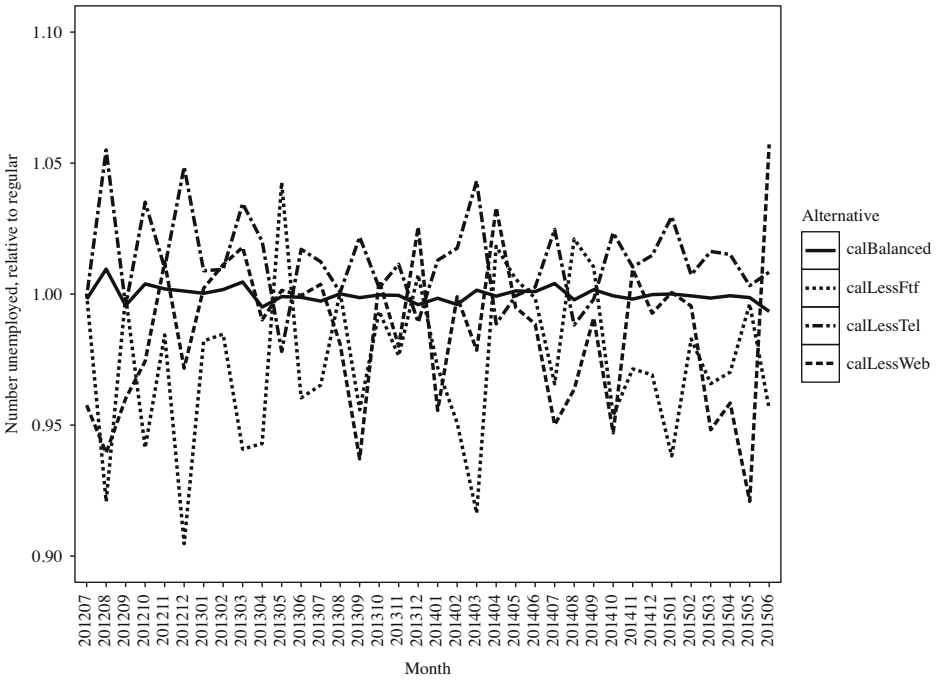

Fig. 2. Estimates of the total number of unemployed obtained through the calibration approach, relative to the regular approach.

three alternatives, the calLessWeb is the least efficient. This is expected, as the share of Web respondents is largest, so suppressing them has the most extreme adverse effect on the efficiency.

If one were to apply the mode calibration method to the LFS for production purposes, the recommendation would be in accordance with Buelens and Van den Brakel (2015) to use calibration levels that are close to the levels realized in the survey. In this case, this would be the calBalanced approach.

4.2. *Measurement Error Correction*

The measurement error correction approach presented in Subsection 2.3 is applied to the same LFS data. Measurement errors are estimated using a regression model with survey

Table 2. Number of unemployed averaged over the 36 month study period, under the various schemes. The composition is the percentage share of Web-Tel-Ftf.

Scheme	Mode composition	Unemployed	SE
regular	variable	678,126	5,211
calBalanced	44-22-34	677,863	5,202
calLessWeb	10-45-45	668,539	6,482
callLessTel	45-10-45	686,634	5,555
calLessFtf	45-45-10	660,369	5,847

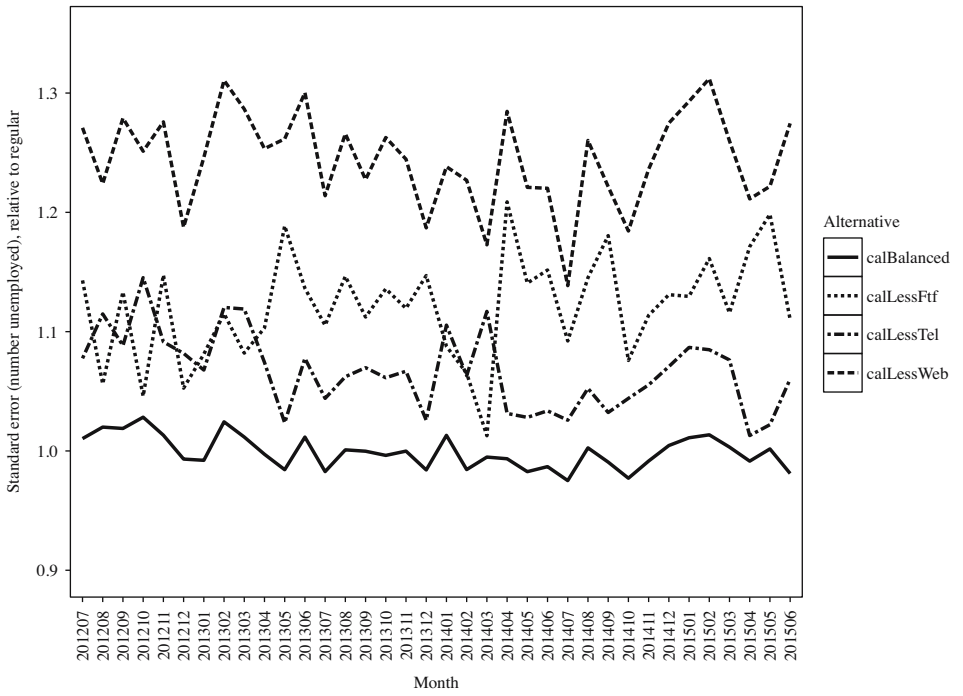

Fig. 3. Standard errors of estimates of the total number of unemployed obtained through the calibration approach, relative to the regular approach.

mode as an explanatory variable in addition to the variables in the GREG model (see Table 1). Since it can be expected that the measurement error does not change during the study period the model is fitted with all data pooled. To allow for between-month variance not explained by the other covariates, month itself is added to the model as a covariate. Corrections are applied in an additive manner using the estimated regression coefficients, which correspond to estimates of the measurement errors, see Equation (7).

Four estimators are considered. One for each mode, corFtf, corTel, and corWeb, which correct the measurements towards face-to-face, telephone, and web modes respectively. A combined correction estimator, corCombi, is a mix of the other three with mixing coefficients in line with the calibration levels of the calBalanced estimator, ie. 44% web, 22% telephone, and 34% face-to-face.

The resulting estimates are shown in Figure 4, again relative to the level of the regular approach. The corCombi estimates are almost equal to the regular estimates. The corFtf and corWeb estimates are higher and the corTel estimates are lower than the regular estimates. Under the assumptions of the applied method, these level differences are due to relative measurement bias between the modes. The finding that telephone interviewing measures at a level below that of the other modes confirms the results of the calibration approach.

The standard errors of these estimates are obtained with a bootstrap procedure and are shown in Figure 5. They are all relatively small compared to the standard errors of the calibration estimators other than the balanced version, see Figure 3. The corFtf and corTel

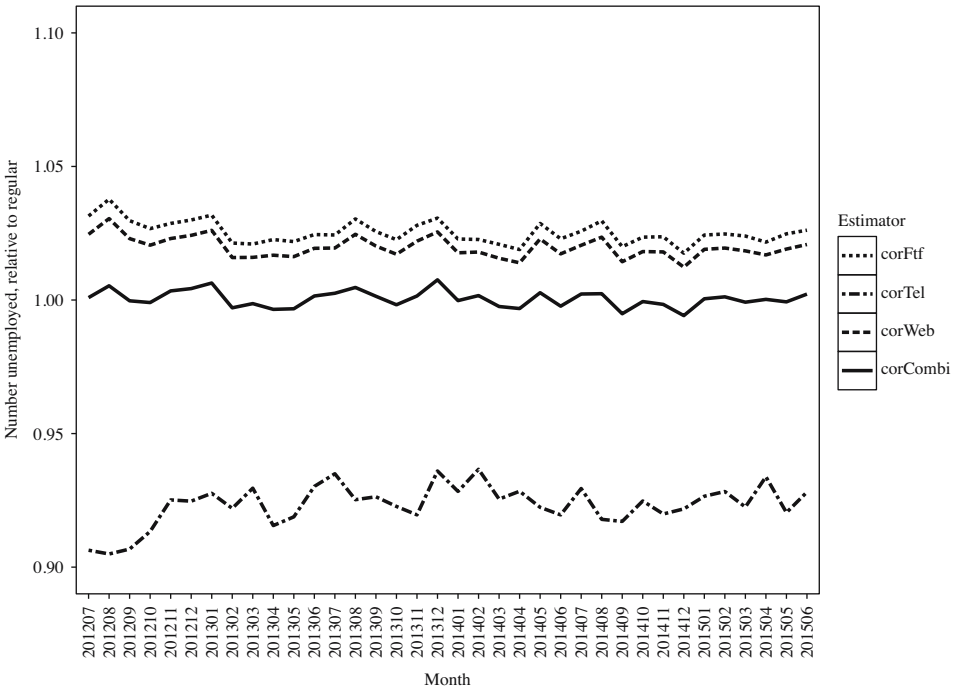

Fig. 4. Estimates of the total number of unemployed obtained through the correction approach, relative to the regular approach.

standard errors are largest as they both require more unit observations to be corrected. The corWeb estimates have standard errors that are only marginally larger than the corCombi estimates, which are similar to the standard errors of the regular approach.

Similar to the annual results for the calibration estimator (see Table 2), the annual results for the correction estimators are shown in Table 3. Of the three estimators that are corrected towards a single mode, the web and face-to-face estimators give comparable results, while the telephone estimator results in a substantially lower estimated number of unemployed. Consequently, the combined estimator results in a level estimate above telephone and below web and face-to-face. The combined estimate is almost equal to the estimate obtained with the regular approach. It is important to stress again that selection bias that is not explained by the model might contribute to the differences seen in Table 3.

It is an empirical result that the estimates corFtf and corWeb are comparable and that both are higher than corTel. The differences are due to mode-dependent measurement errors. The difference between telephone and face-to-face interviewing found here is in line with earlier research; with a randomized experiment embedded in the Dutch LFS, Van den Brakel (2008) showed that the unemployment rate under telephone interviewing is significantly lower than under face-to-face interviewing. The Dutch LFS is a household survey where a response is required from all adult household members. Proxy responses are allowed and are much more frequent in telephone interviews than in face-to-face, which may explain at least a part of the observed differences. Other explanations could be offered by cognitive models of the survey response process. Such models provide a

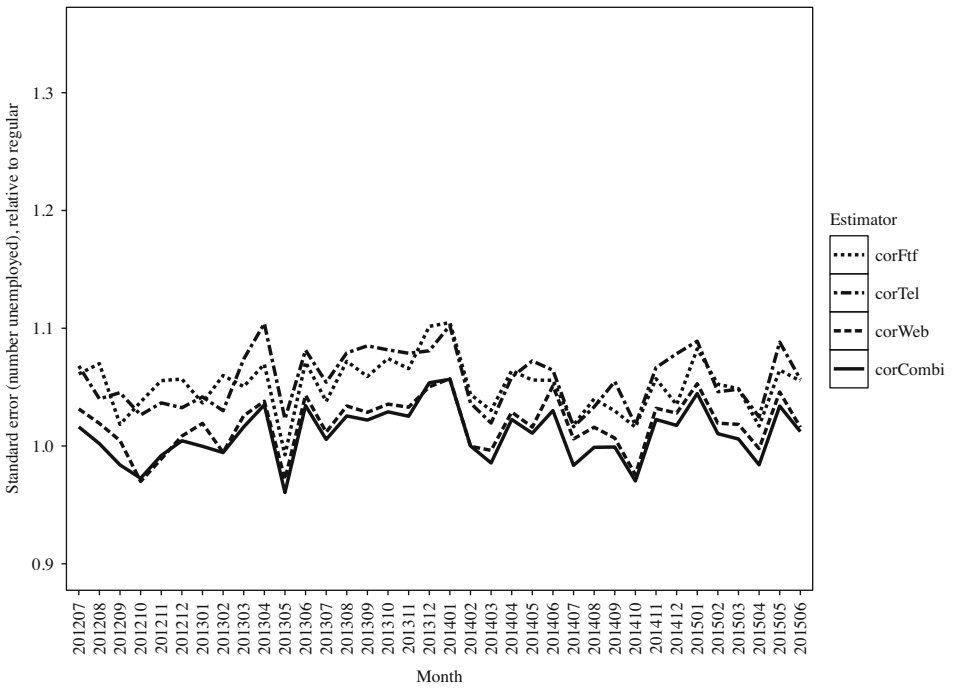

Fig. 5. Standard errors of estimates of the total number of unemployed obtained through the correction approach, relative to the regular approach.

framework for describing the process by which respondents interpret questions, retrieve the required information, make judgements about an adequate response, and provide an answer (Cannel et al. 1981; Tourangeau et al. 2000). A complicating factor in understanding the effects seen in the present analysis is that labour status is derived from a set of questions to determine whether a respondent is working, or willing to work, and is actively looking for work, among other elements. Respondents are generally more likely to give socially desirable answers and demonstrate acquiescence in the presence of an interviewer than in self-administered modes (Dillman et al. 2009; Holbrook et al. 2003). Satisficing (Krosnick 1991) occurs more frequently in self-administered modes than in interviewer modes, and within interviewer modes satisficing occurs more in telephone interviews than in face-to-face interviews, due to the higher speed of the former (Holbrook

Table 3. Number of unemployed averaged over the 36 month study period, using the various correction estimators. The composition is the percentage share of Web-Tel-Ftf.

Estimator	Mode composition	Unemployed	SE
regular	variable	678,126	5,211
corCombi	44-22-34	678,394	5,267
corWeb	100-0-0	691,374	5,311
corTel	0-100-0	626,581	5,507
corFtf	0-0-100	695,122	5,482

et al. 2003). Primacy and recency effects are factors that may explain differences between visual and aural modes (Krosnick and Alwin 1987). They do not completely explain the observed differences, since for some of the questions used to derive the labour market status of respondents, the answer categories are not read out loud by the interviewer. In these cases the interviewer asks an open question and chooses an appropriate answer category based on the answer provided by the respondent. Under the web mode, the respondent can read the different answer categories.

The explanations for the differences between the modes offered by these theories are tentative only. It is not possible to draw conclusions about the validity of the estimates under the different modes, or to choose one of the modes as the benchmark best approximating the true level of unemployment.

4.3. *Calibration versus Correction*

Comparing the preferred calibration approach, where a mode composition is chosen that resembles that actually realized in the survey, to the correction approach with mixing coefficients that are chosen accordingly, gives rise to Figures 6 and 7. All three estimation methods result in virtually the same series of unemployed (Fig. 6) with very similar standard errors (Fig. 7). This is in agreement with the established relation between the two methods, see Subsection 2.4. The empirical outcome that the calibration and correction methods give the same results is reassuring as they are largely based on the same assumptions and models, albeit motivated differently. The small differences between both

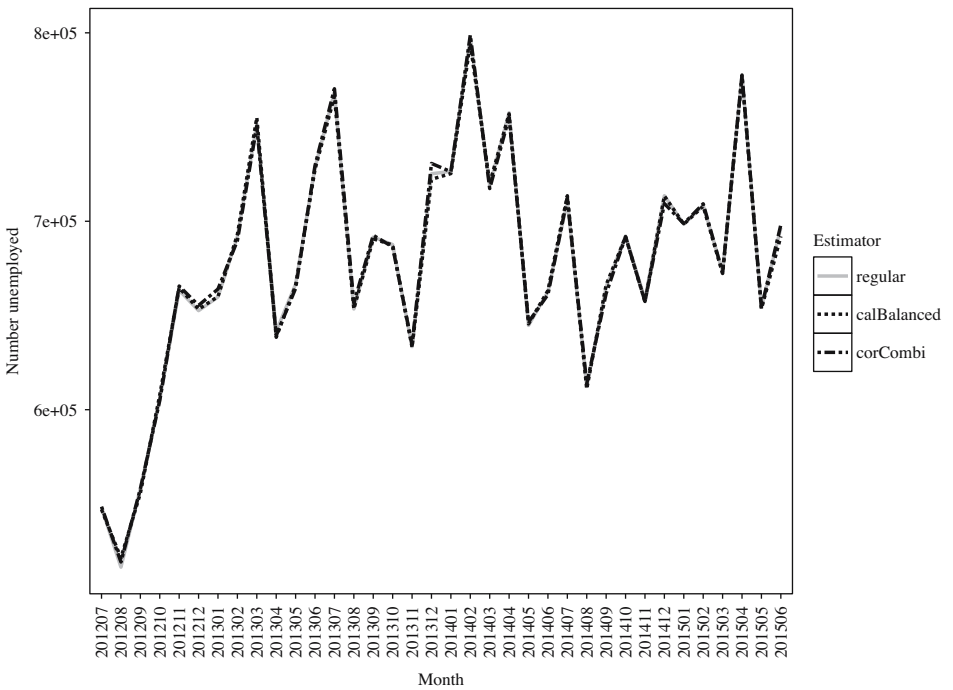

Fig. 6. Estimates of the total number of unemployed obtained through the regular, calibration and correction approaches.

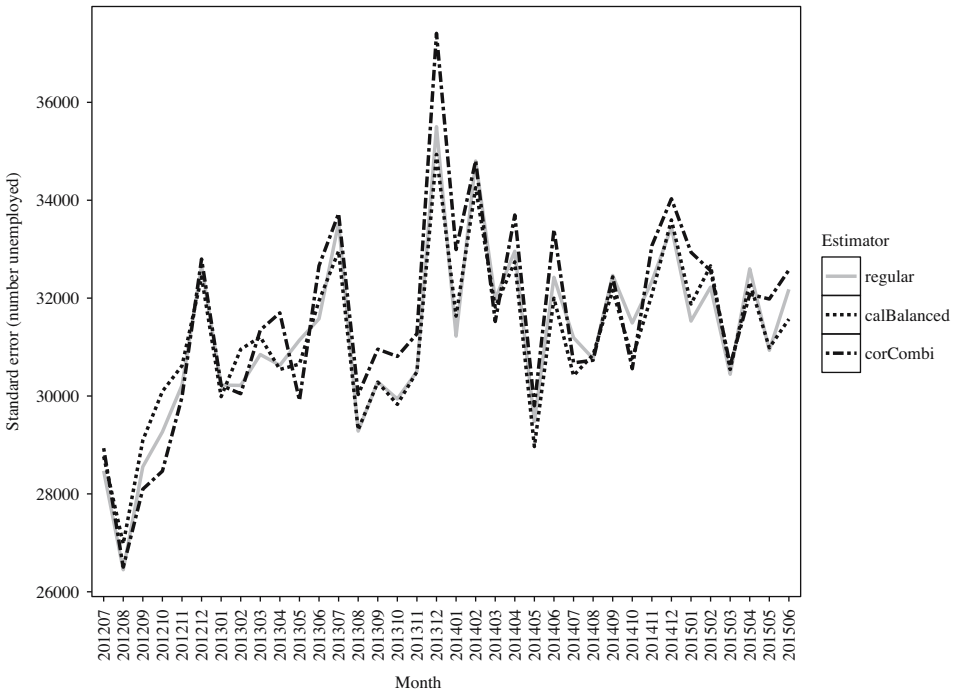

Fig. 7. Standard errors of estimates of the total number of unemployed obtained through the regular, calibration, and correction approaches.

approaches observed in the application can be explained by the fact that the underlying assumption that the auxiliary variables in the GREG estimator apart from the response mode do not completely correct for selective nonresponse. The fact that both almost coincide with the original series is specific to the case at hand, and is due to the relative insensitivity of the results to the realized variations in the mix of survey modes in the LFS. In this specific case, there is no pressing need to apply any of the two methods. However, since there are no adverse effects of the methods, it might be desirable to apply one of the methods nevertheless, as a protective measure against potential future instabilities in the mode composition.

In survey statistics where change over time is strongly confounded with changes in survey mode composition, the calibration and correction methods have a stabilizing effect. This is the case specifically for survey variables that suffer from large mode-dependent measurement effects, such as attitudes or answers to questions susceptible to social-desirability bias. An example where the mode composition varies extremely is the Crime Victimization Survey in the Netherlands, discussed in Buelens and Van den Brakel (2015).

5. Discussion

Estimates from repeated mixed-mode sample surveys can be unstable when the mode composition of the response varies over time. Two recently proposed methods of inference are compared in the present article. The calibration method adjusts the survey weights to

balance the response with respect to the survey modes, while the correction approach adjusts measurements using predicted counterfactuals. While motivated differently, it is shown that both estimators are equal if the mixing parameters for the combined measurement error correction approach mirror the mode distribution assumed for the mode calibration estimator; the remaining auxiliary variables of the weighting schemes of both estimators must be equal too. The two methods rely on the following assumptions (Buelens and Van den Brakel 2015):

i) the weighting model removes mode-dependent selectivity with respect to the survey variables;
ii) time-indepence of the measurement error model;
iii) constant population size – only required when estimating population totals.

When (iii) does not hold, a residual measurement error bias remains, which is not affected by fluctuations in mode composition. Condition (iii) is not required for population means. Violations of (i)–(iii) will lead to biased estimates. It must be emphasized that this would be the case too in uni-mode designs employing a single mode of data collection. Some issues could be resolved by more advanced modeling, for example allowing for time-dependent measurement errors.

In the present research, thirty-six monthly editions of the Dutch LFS are used as a case study. Small deviations between both approaches are observed and can be explained by not meeting the underlying assumption that the auxiliary variables in the weighting model, apart from the mode distribution, completely correct for selective nonresponse. Both approaches produce similar standard errors for the unemployed labour force in the case that the mixing parameters for the combined measurement error correction approach resemble the distribution of the respondents over the modes observed in the sample. In the case of extreme distributions, where the contribution of one of the modes is suppressed, the differences in standard errors under the two approaches are large. The standard error of the mode calibration estimator increases rapidly with increasing discrepancies between the distribution in the sample and in the population. Under the measurement error correction approach, the standard errors increase only slightly, even when the outcomes are corrected to a single mode. The explanation for this difference between the two methods is that the measurement error correction estimator uses additional information by explicitly relying on Model (7) to correct the actual observations for a measurement error component. Unlike the calibration method, the measurement error correction method does not have a built-in protection against strong deviations of the sample and population distributions, unless the mixing coefficients are chosen by minimizing the MSE as proposed by Suzer-Gurtekin (2013), or by choosing them close to the observed mode distribution, as proposed in this article.

The results in Subsection 4.2 indicate that if the LFS were conducted by telephone and the same respondents were reached as currently with the mixed-mode strategy, the estimated average unemployed during the study period would drop from 678000 to 627000. Had the same respondents been interviewed face-to-face, the estimated average would have been 695000. It is a disconcerting thought that the true number of unemployed could be anywhere in this range, or even outside the range, as all three modes can be biased with only relative bias observable. This stresses the inadequacy of traditional measures of

uncertainty only taking into account the uncertainty due to random sampling. This issue is also present in single-mode surveys where it is not as manifestly visible as in mixed-mode surveys. Further research into quantifying measurement related uncertainty is important and could possibly follow the strand of research of the Total Survey Error paradigm, see, for example Groves and Lyberg (2010) for a review.

The observed differences between the three modes are in line with the results of a mode experiment with the LFS obtained in the past and can be partially explained with cognitive models of the survey process. Observed relative mode-effects are nevertheless empirical results and explaining their direction or making statements which mode can be used as a benchmark remains highly speculative. The two methods studied in this article are intended to stabilize the mode distribution in repeated surveys to avoid fluctuations in mode-dependent measurement bias obscuring measurements of change over time. As is the case in single mode surveys, mixed-mode surveys may measure at a level different from the true level in the population. As long as the level difference remains constant through time, change over time can be estimated without bias, both in single mode and mixed-mode surveys. It is recommended to choose the distribution for the mode calibration or the mixing proportions for the correction approach close to the observed distribution of the respondents over the modes in the samples. This avoids unnecessary increase of fluctuations in the weights and in the standard errors. The techniques applied in this article are practically useful as they do not require additional questions, questionnaires, or repeated interviewing.

6. References

Buelens, B. and J.A. Van den Brakel. 2015. "Measurement Error Calibration in Mixed-Mode Sample Surveys." *Sociological Methods & Research* 44(3): 391–426. Doi: http://dx.doi.org/10.1177/0049124114532444.

Cannel, C., P. Miller, and L. Oksenberg. 1981. "Research on Interviewing Techniques." In *Sociological Methodology*, edited by S. Leinhardt. 389–437. San Fransisco: Jossey-Bass.

Centraal Bureau voor de Statistiek. 2015. *Methoden en Definties Enquête Beroepsbevolking 2014*. Technical report, Statistics Nederlands, Heerlen. Available at: https://www.cbs.nl/NR/rdonlyres/1BB3C645-47CC-4F58-9031-89F490AEE981/0/methodenendefinitiesebb2014.pdf (accessed March 2017).

Cernat, A. 2015. "Impact of Mode Design on Measurement Errors and Estimates of Individual Change." *Survey Research Methods* 9(2): 83–99. Doi: http://dx.doi.org/10.18148/srm/2015.v9i2.5851.

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

Dillman, D., G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B. Messer. 2009. "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response and the Internet." *Social Science Research* 39: 1–18. Doi: http://dx.doi.org/10.1016/j.ssresearch.2008.03.007.

Groves R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5): 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065.

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.

Jäckle, A., C. Roberts, and P. Lynn. 2010. "Assessing the Effect of Data Collection Mode on Measurement." *International Statistical Review* 78: 3–20. Doi: http://dx.doi.org/10.1111/j.1751-5823.2010.00102.x.

Klausch, T., J. Hox, and B. Schouten. 2015. "Selection Error in Single- and Mixed Mode Surveys of the Dutch General Population." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(4): 945–961. Doi: http://dx.doi.org/10.1111/rssa.12102.

Krosnick, J. 1991. "Response strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. Doi: http://dx.doi.org/10.1002/acp. 2350050305.

Krosnick, J. and D. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement." *Public Opinion Quarterly* 51: 201–219. Doi: http://dx.doi.org/10.1086/269029.

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

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

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.

Suzer-Gurtekin, Z.T. 2013. *Investigating the Bias Properties of Alternative Statistical Inference Methods in Mixed-Mode Surveys*. Ph.D. thesis, University of Michigan. Available at: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/102471/tsuzer_1.pdf (accessed March 2017).

Suzer-Gurtekin, Z.T., S. Heeringa, and R. Vaillant. 2012. "Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys." In Proceedings of the JSM, Section on Survey Research Methods, San Diego, July 28–August 2, 2012. American Statistical Association, 4711–25.

Tourangeau, R., L. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Van den Brakel, J. 2008. "Design-Based Analysis of Embedded Experiments with Applications in the Dutch Labour Force Survey." *Journal of the Royal Statistical Society, Series A* 171: 581–613. Doi: http://dx.doi.org/10.1111/j.1467-985X.2008.00532.x.

Van den Brakel, J.A. and S. Krieg. 2015. "Dealing with Small Sample Sizes, Rotation Group Bias and Discontinuities in a Rotating Panel Design." *Survey Methodology* 41(2): 267–296. Available at: http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14231-eng.pdf (accessed March 2017).

Vannieuwenhuyze, J.T.A. and G. Loosveldt. 2013. "Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement

Effects." *Sociological Methods & Research* 42(1): 82–104. Doi: http://dx.doi.org/10.1177/0049124112464868.

Voogt, R. and W. Saris. 2005. "Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects." *Journal of Official Statistics* 21: 367–387.

Received January 2016
Revised February 2017
Accepted March 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 533–550, http://dx.doi.org/10.1515/JOS-2017-0025

Adjusting for Measurement Error and Nonresponse in Physical Activity Surveys: A Simulation Study

Nicholas Beyler[1] *and Amy Beyler*[2]

Adult Americans are encouraged to engage in at least 150 minutes of moderate to vigorous physical activity (MVPA) each week. National surveys that collect physical activity data to assess whether or not adults adhere to this guideline use self-report questionnaires that are prone to measurement error and nonresponse. Studies have examined the individual effects of each of these error sources on estimators of physical activity, but little is known about the consequences of not adjusting for both error sources. We conducted a simulation study to determine how estimators of adherence to the guideline for adults to engage in 150 minutes of MVPA each week respond to different magnitudes of measurement and nonresponse errors in self-reported physical activity survey data. Estimators that adjust for both measurement and nonresponse errors provide the least amount of bias regardless of the magnitudes of measurement error and nonresponse. In some scenarios, the naïve estimator, which does not adjust for either error source, results in less bias than estimators that adjust for only one error source. To avoid biased physical activity estimates using data collected from self-report questionnaires, researchers should adjust for both measurement error and nonresponse.

Key words: Measurement error model; moderate to vigorous physical activity; response propensity model; total survey error.

1. Introduction

The U.S. Department of Health and Human Services establishes physical activity guidelines for Americans (Department of Health and Human Services 2008). A key guideline is for adults to engage in at least 150 minutes of moderate to vigorous physical activity (MVPA) each week. Intensity of physical activity is measured in terms of metabolic equivalents (METs), where an individual at rest is at a baseline level of one MET (Welk 2002). For adults, moderate physical activity requires three to six METs of intensity, and vigorous activity requires more than six METs of intensity (Ainsworth et al. 2000; Crouter et al. 2006; Troiano et al. 2008).

One key component of physical activity research is determining the percentage of adult Americans who adhere to this guideline, using information collected via self-report

[1] 2M Research Services, LLC, 1775 Eye Street, NW, Suite 1150, Washington, D.C., U.S.A. Email: nbeyler@2mresearch.com
[2] Mathematica Policy Research, 1100 1st Street, NE, Washington, D.C., U.S.A. Email: abeyler@mathematica-mpr.com

Acknowledgments: We would like to thank Mark Timms for his help developing the figures presented in the article and our colleagues from Mathematica's Data Science and Statistics Methodology Working Group (DSSMWG) for their helpful comments on the article. We would also like to thank the editors and reviewers from JOS for their thoughtful comments on the article.

surveys or questionnaires that ask individuals about their physical activity or using monitoring devices such as accelerometers. Troiano et al. (2008) estimated that approximately 51 percent of adult Americans engage in at least 150 minutes of MVPA each week, using physical activity questionnaire data collected for the 2003–2004 National Health and Nutrition Examination Survey (NHANES). Alternatively, using accelerometers from NHANES 2003–2004, which tend to underestimate the amount of physical activity in individuals (Matthews 2005), Troiano et al. (2008) estimated that less than five percent engage in at least 150 minutes of MVPA each week. Tucker et al. (2011) estimated that about 60 percent of adult Americans adhere to the same guideline, using NHANES questionnaire data from 2005–2006, and about ten percent adhere to the guideline using accelerometer data from NHANES 2005–2006. A study using data from the National Health and Interview Survey (NHIS) estimated that closer to 42 percent of adults adhere to the guideline of 150 minutes of MVPA each week (Schoenborn and Stommel 2011), whereas a study of data from the Behavioral Risk Factor Surveillance System (BRFSS) found that approximately 65 percent of adults adhere to the same guideline (Loustalot et al. 2009). Table A in the Appendix provides more summative information about these surveys and the physical activity self-report questionnaires they used.

These studies provide estimates of the same parameter (percentage of individuals who adhere to the guideline for minutes of MVPA) with data collected from the same population (adult Americans), yet yield different results. This is due, in part, to nonsampling errors. Unlike sampling errors, which result from the conscious choice to collect data from a sample instead of an entire population, nonsampling errors are unintentional and can exist in a number of forms (Lessler and Kalsbeek 1992). If there is a difference between a reported or measured value and the true, unknown value for a survey outcome, the consequence is a form of nonsampling error known as measurement error (or response error). Measurement error exists in self-reported physical activity data because of cognitive limitations associated with recalling activities, which may result in unintentional misreporting about the frequency and duration of activity (Bassett et al. 2000; Matthews 2002). Social desirability effects may also entice some respondents to overreport their physical activity from the recent past (Adams et al. 2005; Warnecke et al. 1997). In addition, the terminology used in self-report questionnaires may be confusing or unclear, which can result in misreporting (Sallis and Saelens 2000). Self-reporting errors like these often result in much larger estimates of physical activity than those taken from more objective instruments like accelerometers or other monitoring devices and can fundamentally bias estimates of physical activity in adult populations (Ferrari et al. 2007; Nusser et al. 2012; Tooze et al. 2013).

If some sampled participants fail to respond to survey questions and those individuals would have responded in a systematically different way than sampled participants who do respond to survey questions, there is the potential for nonresponse error (or bias). Like many surveys, those that collect physical activity information through self-reporting are prone to nonresponse. National surveys attempt to reduce the effects of nonresponse bias by using weighting adjustments or other adjustment procedures, but it is unclear whether such approaches are actually beneficial, since physical activity data are unavailable for nonrespondents. The limited research available on the relationship between nonresponse

and physical activity self-reporting in national surveys is inconclusive about whether nonresponse leads to biased estimates of physical activity outcomes. Hill et al. (1997) found that initial survey respondents reported more physical activity relative to survey nonrespondents who were part of a responding follow-up subsample. Van Loon et al. (2003) found that a smaller percentage of survey respondents had low physical activity compared to nonrespondents. However, other studies did not find significant biases in physical activity estimates due to nonresponse (Smith and Nutbeam 1990; Vink et al. 2004).

To obtain accurate estimates for the percentage of adult Americans that adhere to physical activity guidelines, appropriate adjustments for both measurement error and nonresponse are prudent. However, to the best of our knowledge, no study currently exists in which a total survey error framework (Groves and Lyberg 2010) is utilized to account for both measurement error and nonresponse in physical activity self-report surveys to see if and when such an approach is superior to one that adjusts for only one error source or neither error source. Our goal, in this article, is to determine what the combined impacts of measurement error and nonresponse are on estimates of physical activity generated from self-report data. Because no study or data currently exist which focus on both of these error sources, simulation procedures are needed to help achieve this goal. We present results from a simulation study which measures the bias from not adjusting for one or more of these nonsampling errors when estimating adherence to the physical activity guideline for adult Americans to engage in at least 150 minutes of MVPA each week. We developed our simulation models using established statistical methods from the literature for analyzing physical activity data. We present results for simulated scenarios that cover a range of situations in which measurement and nonresponse errors may exist in self-reported physical activity data collected from adult Americans. For each scenario, we consider estimators that account and adjust for both measurement and nonresponse errors, measurement error only, nonresponse error only, and neither error source.

2. Methods

In this section we describe the procedures used for the simulation study. A more technical description of the simulation procedures is provided in the Appendix. For the simulation, we assume that a random sample of 1,000 individuals is selected from the population of all adult Americans, and sampled participants are asked to provide information about their physical activity using a self-report survey during two separate weeks, randomly selected, over the course of a year. We assume that each sampled individual has some true, unknown amount of time he or she spends engaging in MVPA during a typical week which is different than the amount he or she reports from the self-report survey. This difference is due to measurement error, which we assume to be random (see Equation 1 in the Appendix). We also assume that some sampled individuals will not provide all the necessary information to measure the amount of time they spend engaged in MVPA from the self-report survey. Each sampled individual is assumed to have some propensity to respond to the survey which is a function of his or her true, unknown physical activity level (see Equation 2 in the Appendix). For the simulation, individuals are randomly assigned as either respondents or nonrespondents according to their response propensity. So, for

example, if an individual's propensity to respond is 0.75, he or she has a 75 percent probability of being assigned as a respondent.

Using the simulated data for the 1,000 sampled individuals, we estimate the parameter, the percentage of adults in the population that adhere to the physical activity guideline of 150 or more minutes of MVPA per week (which we denote θ) using four different estimators. One estimator accounts and adjusts for both the measurement error and nonresponse (which we denote $\hat{\theta}_{full}$), one accounts for nonresponse only, ignoring the measurement error (which we denote $\hat{\theta}_{nr}$), one accounts for the measurement error only, ignoring the nonresponse (which we denote $\hat{\theta}_{me}$), and one accounts for neither measurement error nor nonresponse (which we denote $\hat{\theta}_{naive}$). (These estimators are defined more explicitly in the Appendix along with other simulation model parameters.) It is important to focus on all four of these estimators so that we understand not only the combined impact of not adjusting for measurement error and nonresponse (which is represented by the naïve estimator) but also what the impact is if adjustments for only one of these error sources is ignored. Once we calculate estimates for these four estimators based on the sample of 1,000 individuals we repeat the simulation process again (randomly selecting another 1,000 individuals) and calculate four new estimates. In total, we run 1,000 simulations, generating 1,000 replicate estimates for each of the four estimators. We calculate the average bias for each estimator by averaging the differences between the 1,000 simulated estimates and the true value being estimated (the percentage of adults that engage in 150 or more minutes of MVPA each week).

As with any simulation, a number of assumptions are made about the parameter values. The parameter values selected for the simulations (which are provided along with the simulation models in the Appendix) were based on research which studied measurement error and nonresponse in data from physical activity self-report surveys (Ferrari et al. 2007; Tooze et al. 2013; Hill et al. 1997). We assumed that the true percentage of adults that adhere to the physical activity guideline of 150 or more minutes of MVPA each week was $\theta = 25$ percent. We consider this value reasonable since the available self-report-based estimates of this quantity, ranging from 42 to 65 percent (Troiano et al. 2008; Tucker et al. 2011; Schoenborn and Stommel 2011; Loustalot et al. 2009), are likely too high due to reporting biases, and the accelerometer-based estimates of this quantity, ranging from five to ten percent (Troiano et al. 2008; Tucker et al. 2011), are likely too low, since accelerometers are unable to accurately measure all types of activity in free-living conditions (Matthews 2005).

To investigate the combined impacts of measurement error and nonresponse on the estimators, we considered a range of scenarios. We considered nine different measurement error scenarios (ranging from no random measurement error to very large random measurement error). We also considered nine different nonresponse scenarios (ranging from a strong negative relationship between propensity to respond and physical activity to a strong positive relationship between propensity to respond and physical activity). In total, we considered 81 different simulation scenarios based on the combinations of nine measurement error scenarios and nine nonresponse scenarios. In the results section, we conduct a more detailed examination of six of these scenarios summarized in Table 1 and then summarize findings from all 81 scenarios at the end of the section.

Table 1. Model parameter specifications for six simulation scenarios.

Scenario	Empirical values of key model parameters for Models (1) and (2) in the Appendix
Large measurement error and large negative nonresponse bias	$\sigma_e^2 = 8$ and $\alpha_1 = -2$
Large measurement error and no nonresponse bias	$\sigma_e^2 = 8$ and $\alpha_1 = 0$
Large measurement error and large positive nonresponse bias	$\sigma_e^2 = 8$ and $\alpha_1 = 2$
No measurement error and large negative nonresponse bias	$\sigma_e^2 = 0$ and $\alpha_1 = -2$
No measurement error and no nonresponse bias	$\sigma_e^2 = 0$ and $\alpha_1 = 0$
No measurement error and large positive nonresponse bias	$\sigma_e^2 = 0$ and $\alpha_1 = 2$

Note: σ_e^2 represents measurement error variability and α_1 represents nonresponse bias. Both terms are explicitly defined in the Appendix.

3. Results

The simulation results for the six scenarios given in Table 1 are presented in terms of average biases in Table 2. There are two separate panels in Table 2 that provide results for scenarios with large measurement error (top panel) and no measurement error (bottom panel). Within each panel, three columns represent three different nonresponse bias conditions – negative nonresponse bias, no nonresponse bias, and positive nonresponse bias. Each table cell includes four average bias estimates for the four estimators of θ.

3.1. Simulation Scenarios with Large Measurement Error

Average biases for a scenario in which there is large measurement error and negative nonresponse bias are given in the first column of the top panel of Table 2. The estimators

Table 2. Average biases for estimators of θ for six simulation scenarios.

Large measurement error ($\sigma_e^2 = 8$ in Model (1) in the Appendix)		
Negative nonresponse bias ($\alpha_1 = -2$ in Model (2) in the Appendix)	No nonresponse bias ($\alpha_1 = 0$ in Model (2) in the Appendix)	Positive nonresponse bias ($\alpha_1 = 2$ in Model (2) in the Appendix)
$\hat{\theta}_{naive}$: 1.0	$\hat{\theta}_{naive}$: 11.5	$\hat{\theta}_{naive}$: 21.5
$\hat{\theta}_{nr}$: 11.5	$\hat{\theta}_{nr}$: 11.5	$\hat{\theta}_{nr}$: 11.6
$\hat{\theta}_{me}$: −18.0	$\hat{\theta}_{me}$: −0.4	$\hat{\theta}_{me}$: 17.1
$\hat{\theta}_{full}$: −1.3	$\hat{\theta}_{full}$: −0.4	$\hat{\theta}_{full}$: 0.3
No measurement error ($\sigma_e^2 = 0$ in Model (1))		
Negative nonresponse bias ($\alpha_1 = -2$ in Model (2))	No nonresponse bias ($\alpha_1 = 0$ in Model (2))	Positive nonresponse bias ($\alpha_1 = 2$ in Model (2))
$\hat{\theta}_{naive}$: −19.5	$\hat{\theta}_{naive}$: 0.0	$\hat{\theta}_{naive}$: 15.1
$\hat{\theta}_{nr}$: −0.0	$\hat{\theta}_{nr}$: 0.0	$\hat{\theta}_{nr}$: −0.1
$\hat{\theta}_{me}$: −18.5	$\hat{\theta}_{me}$: 0.1	$\hat{\theta}_{me}$: 17.1
$\hat{\theta}_{full}$: −1.0	$\hat{\theta}_{full}$: 0.0	$\hat{\theta}_{full}$: −0.3

that account for only one error source do not perform well under this scenario. The average bias for $\hat{\theta}_{nr}$ is 11.5 percentage points. This means that when accounting for nonresponse bias but not for measurement error, under this scenario estimates of θ will be around 36 or 37 percent for a true value of θ that is 25 percent. The average bias for $\hat{\theta}_{me}$ is -18.0 percentage points under this same scenario, so that when accounting for only measurement error, estimates of θ will be around seven percent (well below the true value of 25 percent). The naïve estimator, $\hat{\theta}_{naive}$, has an average bias of 1.0 percentage points and performs as well as the full-adjustment estimator, $\hat{\theta}_{full}$, which has an average bias of -1.3 percentage points.

In the scenario where there is large measurement error and no bias due to nonresponse, the measurement error estimator performs well, as does the full-adjustment estimator. Both have average biases that are 1.0 percentage points in absolute value. Neither the nonresponse adjustment estimator nor the naïve estimator perform well. Both have average biases of 11.5 percentage points.

The final scenario in the top panel of Table 2, with large measurement error and positive nonresponse bias, provides the largest absolute bias in the naïve estimator: 21.5 percentage points. This means that, under this scenario, the naïve estimator provides estimates of θ close to 50 percent, almost double the true value of 25 percent. Like the naïve estimator, the nonresponse adjustment and measurement error adjustment estimators do not perform well. The full-adjustment estimator does perform well, with an average bias of only 0.3 percentage points.

3.2. *Simulation Scenarios with no Measurement Error*

The bottom panel of Table 2 provides three scenarios where there is no measurement error. As one would expect, under these scenarios the nonresponse adjustment estimator performs as well as the full-adjustment estimator. The naïve and measurement error adjustment estimators do not perform as well, except for the scenario where there is no nonresponse bias. In the scenario with negative nonresponse bias and no measurement error, the average negative bias for the naïve and measurement error adjustment estimators are both about 19 percentage points. So if these estimators were used under this scenario, the estimates would be around six percent, much less than the true parameter value of 25 percent.

3.3. *Graphical Representations for Biases in the Naïve and Full Estimators*

A more comprehensive picture of the simulation results across all 81 scenarios for the naïve and full estimators is provided in Figures 1 and 2, respectively. In each figure, the x-axis represents the magnitude of the nonresponse bias, which ranges from large negative nonresponse bias (when $\alpha_1 = -2$) to large positive nonresponse bias (when $\alpha_1 = 2$). Along the y-axis, the measurement error ranges from no measurement error (when $\sigma_e^2 = 0$) to large positive measurement error (when $\sigma_e^2 = 8$). The shading in the plot is darker when there is larger bias and the color key in the top left corner of each figure shows the numerical values associated with different shading (little to no shading means that there is little to no bias in the estimator). The "+" and "−" signs indicate whether the bias is positive or negative, respectively. For example, a box that is a dark shade with a "+" sign

means that the estimator was positively biased under the specific scenario the box represents in terms of the measurement error and nonresponse bias scenarios.

In Figure 1, we see that the bias in the naïve estimator varies greatly depending on the magnitudes of the measurement and nonresponse errors. In the bottom left portion of the plot, there is negative bias in the naïve estimator; in the top right portion there tends to be more positive bias in the naïve estimator. There is lighter shading in the middle of the plot, which represent scenarios where there is little to no bias in the naïve estimator. This was also observed in the scenario presented in the top left cell of Table 2, where we saw the naïve estimator perform relatively well.

In comparison to Figure 1, the shading in Figure 2 is consistently much lighter, ranging from only about -2 to $+2$ percentage points in bias across all scenarios, suggesting that there is always little to no bias in the full estimator regardless of the magnitudes of measurement error and nonresponse bias. This was also reflected in Table 2 where the bias in the full estimator was always about one percentage point or less in absolute value.

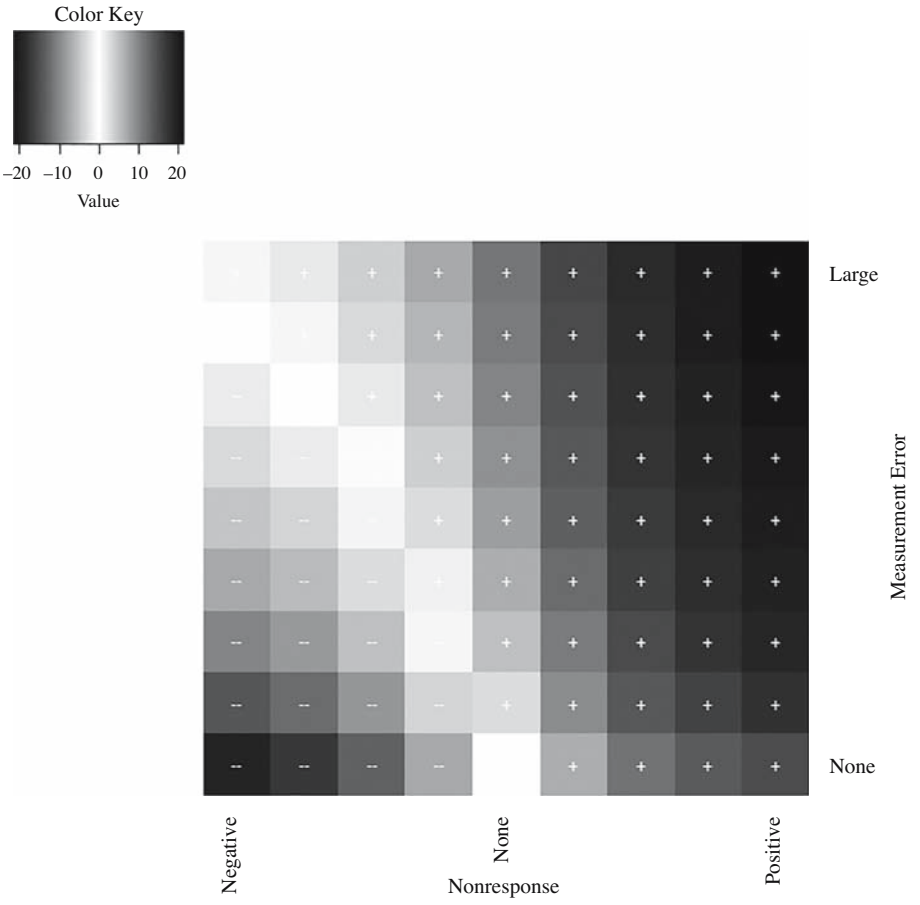

Fig. 1. Average biases in the naïve estimator for simulation scenarios with negative to positive nonresponse bias and no to large measurement error. Darker shades of grey with "−" signs indicate greater negative bias in the naïve estimator; darker shades of grey with "+" signs indicate greater positive bias in the naïve estimator.

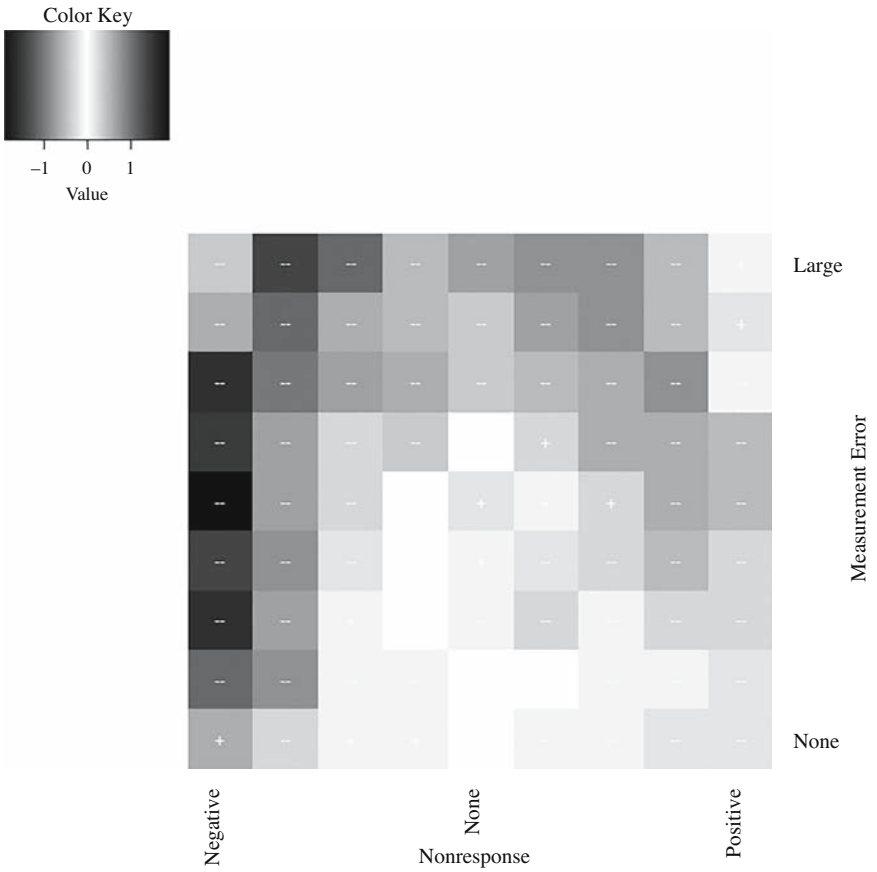

Fig. 2. Average biases in the full estimator for simulation scenarios with negative to positive nonresponse bias and no to large measurement error. Darker shades of grey with "−" signs indicate greater negative bias in the full estimator; darker shades of grey with "+" signs indicate greater positive bias in the full estimator.

4. Discussion

Across all scenarios in our simulation study, which investigate the effects of 81 different combinations of measurement error and nonresponse on estimators of the percentage of adults that adhere to physical activity guidelines, the results suggest that adjusting for both measurement error and nonresponse is preferable to adjusting for one error source or neither error source. The average bias for the full estimator, $\hat{\theta}_{full}$, is close to zero regardless of the magnitudes of the measurement error and nonresponse biases (Figure 2). In contrast, the average bias for all other estimators ranges from -19.5 to 21.5 percentage points, depending on the scenario being considered (Table 2). There are a few other findings that deserve further discussion.

In scenarios where there is large measurement error and negative nonresponse bias, the naïve estimator, which does not adjust for measurement error or nonresponse, performs better than the estimators that adjust for only one error source (Table 2). This is due to a cancelling-out effect in the error sources. With negative nonresponse bias, more inactive individuals respond than active individuals. But with large measurement error, these

individuals are also misreporting their activity, which creates a positive bias that negates the effect of the negative nonresponse bias in the naïve estimator. The likelihood of scenarios in which there is a negative nonresponse bias in large-scale survey settings may be small. Intuitively, one may actually expect inactive individuals to be less likely to respond to a physical activity survey – not more likely – because they are not as aware of or as interested in their physical activity compared to more active individuals. Moreover, the current, albeit limited, research that looks at relationships between physical activity and nonresponse suggests that the correlation between physical activity and propensity to respond is either positive (Hill et al. 1997; Van Loon et al. 2003) or close to zero (Smith and Nutbeam 1990; Vink et al. 2004).

The more likely simulation scenarios that we could observe in practice may be those in which there is large measurement error and positive nonresponse bias. In these scenarios, the full-adjustment estimator performs well, but the other three estimators do not. The naïve estimator and those that adjust for only one error source produce average biases ranging from 12 to 22 percentage points (Table 2). Therefore, in the more extreme cases, estimates for the percentage of adults that adhere to the physical activity guideline of 150 minutes of MVPA per week that do not adjust for both error sources reach almost 50 percent, whereas the true parameter being estimated is 25 percent. In other words, only one in four adult Americans adheres to the guideline (in our simulation study), whereas estimates that do not properly adjust for measurement error and nonresponse suggest that anywhere between one in three and one in two adult Americans adheres to the guideline. Estimates based on self-report data from national surveys are not far off from 50 percent (Troiano et al. 2008; Tucker et al. 2011; Schoenborn and Stommel 2011; Loustalot et al. 2009). In these studies, weighted estimates are presented that account for the sample design and nonresponse, but nonresponse adjustments are based primarily on demographic characteristics of respondents and nonrespondents, like gender and race, and not on physical activity outcomes. For this reason, the methods considered in this simulation study cannot be directly applied to data from national surveys such as NHANES or NHIS because these surveys do not collect two or more independent measures at different points in time from a subset of individuals. We recommend that future surveys be designed with these specifications to allow for both adjustments. So although it is unclear whether the true value of 25 percent used in the simulations is accurate, the simulations in which we see overestimation of this parameter may be good proxies for what is happening in practice with the self-report data from these national surveys: a combination of positive, nonresponse bias and large measurement error. Respondents in these surveys may tend to be more physically active individuals and may also misreport their activity.

The findings presented in this article are not without their limitations. The findings are based on simulations and should be interpreted as such. The relationship between propensity to respond and physical activity may not be as simplistic as we depict it in Model (2). There may be different response patterns that depend on the gender, age group, racial or ethnic status, or other characteristics of the adult Americans in the population. Moreover, in some subgroups of the population, there may be no association between physical activity and propensity to respond, as some studies suggest (Smith and Nutbeam 1990; Vink et al. 2004). Additional research that sets out to capture the physical activity of

individuals who tend to be nonrespondents in national surveys through other avenues, like field follow up, is needed.

Another caution to note is that the measurement error model (1) used in the simulations assumes that there is no systematic bias in the self-report measurements. This is likely too optimistic an assumption and was used in our simulations to focus, specifically, on random measurement error. Other studies have shown evidence of systematic biases in self-reports of physical activity outcomes (Ferrari et al. 2007; Nusser et al. 2012; Tooze et al. 2013), although they focused on data analyzed from convenience samples that are not representative of all adult Americans. To better understand measurement error properties of self-reported physical activity data in national surveys, multiple measurements from different points in time should be considered for at least a subsample of survey participants in order to distinguish measurement error and other forms of intra-individual variation in the data from actual variability in physical activity across individuals. Such an approach would not require a major change in a survey's design; at a minimum, a single follow-up data collection on a randomly selected subsample of the individuals at a later point in time (after the initial survey data are collected) would be required. As an alternative to this measurement error approach, structural equation models which model multiple indicators of physical activity from national surveys could be considered to better understand the measurement error properties in physical activity measurements.

The study's limitations should not diminish the importance of the findings to the study of physical activity measurement. As part of a comprehensive simulation study, we established 81 different scenarios in which measurement error and nonresponse influence self-reported physical activity data. Although there are a number of other scenarios we could have considered, including those that simulated systematic measurement error or more complex patterns of nonresponse, the ones we focus on provide a good starting point for this line of research focusing on the impacts of multiple sources of nonsampling error on self-reported physical activity outcomes.

Appendix

Technical Documentation for the Simulation Study

In this Appendix, we provide a more technical presentation of the simulation study set up provided in the methods section.

Model Development

For the simulation, we assume that a random sample is selected from the population of all adult Americans, and sampled participants are asked to provide information about their physical activity using a self-report survey. Each individual i in the sample has some true, unknown amount of time he or she spends engaging in MVPA during a typical week. We define T_i to be the time spent in MVPA during a typical week for individual i in minutes. Without loss of generality we assume that in the population, individuals' T_i are independent and normally distributed with mean μ_T and variance σ_T^2. In practice, a log transformation (Ferrari et al. 2007; Tooze et al. 2013) or power transformation (Beyler

et al. 2015) may be required to achieve normality. The main parameter of interest, θ, is the percentage of adult Americans that engage in 150 minutes or more MVPA during a typical week. If we knew the value of T_i for each individual in the sample, we could estimate θ as

$$\hat{\theta} = n^{-1}\sum_{i=1}^{n} I\{T_i \geq 150\},$$

where n is the number of individuals in the sample and $I\{T_i \geq 150\}$ is an indicator function that takes a value of 1 if T_i is greater than or equal to 150, and a value of 0 otherwise. However, in practice, T_i is unknown and subject to nonsampling errors; therefore, alternative estimation methods must be considered.

We assume that the survey asks each (simulated) individual in the sample to report on his or her physical activity during two weeks randomly selected over the course of a year; from these reports, we obtain two measures for the time he or she spends engaging in MVPA (in minutes) during the course of a typical week. We define X_{ij} to be the self-report measure for individual i from week j ($j = 1, 2$). We assume that X_{ij} is subject to measurement error and define the relationship between X_{ij} and T_i for each individual as

$$X_{ij} = T_i + e_{ij}, \tag{1}$$

where e_{ij} is a random measurement error term for individual i from week j. We assume that the e_{ij} are independent and normally distributed with mean 0 and variance σ_e^2. We also assume that T_i and $e_{i'j}$ are independent for all i, i', and j. These assumptions are necessary for model identifiability. Model (1) is often referred to as the classical measurement error model (Carroll et al. 2006) and assumes that X_{ij} is an unbiased measurement of T_i. Such an assumption is considered for the purposes of model development for this simulation study but in practice is often not justifiable for self-report measures and more robust model assumptions and model equations should be considered (Ferrari et al. 2007; Nusser et al. 2012; Tooze et al. 2013).

We assume that all individuals sampled will not respond and hence we will not obtain X_{ij} values from all of our sampled individuals. We assume that each sampled individual i has a propensity to respond to the survey, p_i, defined by

$$log\left\{\frac{p_i}{1-p_i}\right\} = \alpha_0 + \alpha_1 T_i, \tag{2}$$

where α_0 and α_1 are fixed intercept and slope coefficients, respectively. In model (2) we assume that for $\alpha_1 \neq 0$ there is a relationship between an individual's propensity to respond and how physically active he or she is. The full simulation model, accounting for the influence of measurement error and nonresponse on true activity (T_i), is represented by model equations (1) and (2).

Simulation Procedures

In our simulation study, we consider four estimators of θ, the true percentage of adult Americans who engage in 150 or more minutes of MVPA each week. We define $\hat{\theta}_{full}$ to be the estimator of θ that accounts and adjusts for both measurement error and nonresponse, $\hat{\theta}_{nr}$ to be the estimator of θ that accounts and adjusts for nonresponse only, $\hat{\theta}_{me}$ to be the

estimator of θ that accounts and adjusts for measurement error only, and $\hat{\theta}_{naive}$ to be the estimator of θ that does not account for either error source. We develop these four estimators in a series of steps:

1. First we randomly generate n values of T_i from a normal distribution with mean μ_T and variance σ_T^2.
2. For each T_i we randomly generate two values of e_{ij} (e_{i1} and e_{i2}) from a normal distribution with mean 0 and variance σ_e^2.
3. Next, we calculate values of X_{ij} and p_i from Models (1) and (2), respectively.
4. For each individual i, we randomly generate R_i from a Bernoulli distribution with success probability p_i. If R_i is 1, individual i is a respondent; if R_i is 0, individual i is a nonrespondent.
5. Because p_i is not known, in practice, for the simulation we estimate p_i for all respondents. For individuals with $R_i = 1$, we let

$$\hat{p}_i = \frac{exp\{\alpha_0 + \alpha_1 T_i + \varphi_i\}}{1 + exp\{\alpha_0 + \alpha_1 T_i + \varphi_i\}},$$

   where φ_i is randomly generated from a normal distribution with mean 0 and variance σ_φ^2. Each respondent is then assigned a nonresponse adjustment weight of $w_i = \hat{p}_i^{-1}$.
6. Next, we fit the measurement error model (1) to the data from responding individuals using method of moments. (For this measurement error model, method of moments provides similar estimates to those calculated using maximum likelihood.) We estimate both unweighted and weighted model parameters. The unweighted estimator of μ_T is

$$\hat{\mu}_T^{unwt} = n_R^{-1} \sum_{i=1}^{n_R} \bar{X}_i,$$

   where n_R is the number of responding individuals (those with $R_i = 1$) and $\bar{X}_i$ is the average of X_{i1} and X_{i2}. The weighted estimator of μ_T is

$$\hat{\mu}_T^{wt} = \frac{\sum_{i=1}^{n_R} w_i \bar{X}_i}{\sum_{i=1}^{n_R} w_i},$$

   where w_i is the nonresponse adjustment weight defined in Step 5. The unweighted and weighted estimators of σ_e^2 are

$$\hat{\sigma}_{e,unwt}^2 = n_R^{-1} \sum_{i=1}^{n_R} \sum_{j=1}^{2} \left(X_{ij} - \bar{X}_i\right)^2,$$

   and

$$\hat{\sigma}_{e,wt}^2 = \frac{\sum_{i=1}^{n_R} \sum_{j=1}^{2} w_i \left(X_{ij} - \bar{X}_i\right)^2}{\sum_{i=1}^{n_R} w_i},$$

respectively. The unweighted and weighted estimators of σ_T^2 are

$$\hat{\sigma}^2_{T,unwt} = (n_R - 1)^{-1} \sum_{i=1}^{n_R} \left(\bar{X}_i - \bar{X}_{..}\right)^2 - \frac{1}{2}\hat{\sigma}^2_{e,unwt},$$

and

$$\hat{\sigma}^2_{T,wt} = \frac{\sum_{i=1}^{n_R} w_i \left(\bar{X}_i - \bar{X}_{..}\right)^2}{\left(\sum_{i=1}^{n_R} w_i\right) - 1} - \frac{1}{2}\hat{\sigma}^2_{e,wt},$$

respectively, where $\bar{X}_{..}$ is the overall mean of the X_{ij} for responding individuals.

7. We estimate $\hat{\theta}_{naive}$ as

$$\hat{\theta}_{naive} = n_R^{-1} \sum_{i=1}^{n_R} I\{\bar{X}_i \geq 150\},$$

where $I\{\bar{X}_i \geq 150\}$ is an indicator function that takes a value of 1 if $\bar{X}_i$ is greater than or equal 150, and a value of 0 otherwise.

8. We estimate $\hat{\theta}_{nr}$ as

$$\hat{\theta}_{nr} = \frac{\sum_{i=1}^{n_R} w_i I\{\bar{X}_i \geq 150\}}{\sum_{i=1}^{n_R} w_i},$$

where w_i is the nonresponse adjustment weight defined in Step 5.

9. To estimate $\hat{\theta}_{me}$ we first generate 1,000 values of $\ddot{X}_i$ from a normal distribution with mean $\hat{\mu}_T^{unwt}$ and variance $\hat{\sigma}^2_{T,unwt}$. Then, the estimator is

$$\hat{\theta}_{me} = \frac{\sum_{i=1}^{1,000} I\{\ddot{X}_i \geq 150\}}{1,000}.$$

This approach, which adjusts for the measurement error in X_{ij}, is based on methods described in Dodd et al. (2006) and elsewhere.

10. To estimate $\hat{\theta}_{full}$, we use a similar approach to that described in Step 9, but use the weighted measurement error model parameter estimates instead of the unweighted estimates. We generate 1,000 values of $\tilde{X}_i$ from a normal distribution with mean $\hat{\mu}_T^{wt}$ and variance $\hat{\sigma}^2_{T,wt}$. Then, the estimator $\hat{\theta}_{full}$ is

$$\hat{\theta}_{full} = \frac{\sum_{i=1}^{1,000} I\{\tilde{X}_i \geq 150\}}{1,000}.$$

This entire process (Steps 1–10) is repeated M times to account for simulation variation. For each of the M simulations and each of the four estimators, we calculate the bias as the difference between an estimate and the true value of the parameter θ. The average bias across the M simulations is then calculated for each of the four estimators.

For the simulation, we set $\mu_T = 4.20$ and $\sigma_T^2 = 1.44$ which represent the mean and variance of time spent in MVPA during a typical week in the log scale. We use the log scale because physical activity data are often analyzed in the log scale (Ferrari et al. 2007;

Tooze et al. 2013). For these choices of μ_T and σ_T^2, the true value of θ is about 25 percent and the ratio of μ_T and σ_T^2 is similar to the ratio found in other studies (Ferrari et al. 2007; Tooze et al. 2013).

We set $\sigma_\varphi^2 = 0.04$ to estimate the response propensities (see Step 5 above). Values of σ_φ^2 larger than 0.04 introduce increased uncertainty in the estimated response propensities which diminishes the effectiveness of the nonresponse adjustments. Although this may very well occur in practice, the goal of these simulations is to compare the effects of adjusting and not adjusting for measurement error and nonresponse when the adjustments are considered effective, without model misspecifications that could result in poorly estimated response propensities or measurement error model parameters.

For the remaining model parameters that subject the data to measurement error and nonresponse bias, we considered a range of scenarios. We let the measurement error variance, σ_e^2, range from 0 (no measurement error), to 8 (large measurement error) by 1-unit increments. With larger measurement error there will be more positive bias when estimating θ without proper adjustments for the measurement error because there will be more variability in the data and consequently a higher percentage of cases above the threshold of 150 minutes of MVPA per week. We let the response propensity parameter, α_1, range from -2 to 2 by 0.5-unit increments. Negative values of α_1 will result in negative bias when estimating θ (without accounting and adjusting for nonresponse) and positive values of α_1 will result in positive bias. The empirical values of α_0 are chosen to correspond with values of α_1, so that each set of response propensity parameters gives an expected response rate of about 60 percent. In total, we considered 81 different simulation scenarios based on the nine values of σ_e^2 and nine combinations of α_0 and α_1.

Table A. National surveys with physical activity self-reports.

National survey	Estimated percentage of adults that adhere to physical activity guideline	Survey Year(s) in which estimate is based	Data collection mode	Link to questionnaire used to estimate MVPA	Source
BRFSS	65%	2007	Landline telephone	http://www.cdc.gov/brfss/questionnaires/pdf-ques/2007brfss.pdf	Loustalot et al. (2009)
NHANES	51%	2003–2004	Field interview	https://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/sp_paq_c.pdf	Troiano et al. (2008)
	60%	2005–2006		https://www.cdc.gov/nchs/data/nhanes/nhanes_05_06/sp_paq_d.pdf	Tucker et al. (2011)
NHIS	42%	1997–2004	Field interview	ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2004/english/QADULT.pdf	Schoenborn and Stommel (2011)

Note: All physical activity measurements were self-administered. The NHIS estimated percentage of adults that adhere to physical activity guideline was averaged across the 1997–2004 survey years.

5. References

Adams, S.A., C.E. Matthews, C.B. Ebbeling, C.G. Moore, J.E. Cunningham, J. Fulton, and J.R. Hebert. 2005. "The Effect of Social Desirability and Social Approval on Self-Reports of Physical Activity." *American Journal of Epidemiology* 161: 389–398. Doi: https://doi.org/10.1093/aje/kwi054.

Ainsworth, B.E., W.L. Haskell, M.C. Whitt, M.L. Irwin, A.M. Swartz, S.J. Strath, W.L. O'Brien, D.R. Bassett, K.H. Schmitz, P.O. Emplaincourt, and D.R. Jacobs. 2000. "Compendium of Physical Activities: An Update of Activity Codes and MET Intensities." *Medicine and Science in Sports and Exercise* 32: S498–S504. Available at: https://pdfs.semanticscholar.org/314e/dc8553c9a5920a14eb799b67c2a11e07b8bf.pdf (accessed April 2017).

Bassett, D.R., A.L. Cureton, and B.E. Ainsworth. 2000. "Measurement of Daily Walking Distance: Questionnaire versus Pedometer." *Medicine and Science in Sports and Exercise* 32: 1018–1023. Available at: http://journals.lww.com/acsm-msse/pages/articleviewer.aspx?year=2000&issue=05000&article=00021&type=abstract (accessed April 2017).

Beyler, N., S. James-Burdumy, M. Bleeker, J. Fortson, and M. Benjamin. 2015. "Estimated Distributions of Usual Physical Activity During Recess". *Medicine and Science in Sports and Exercise* 46: 1197–1203. Doi: https://doi.org/10.1249/MSS.0000000000000535.

Carroll, R.J., D. Ruppert, and L.A. Stefanski. 2006. *Measurement Error in Nonlinear Models*. New York: CRC Press.

Crouter, S.E., K.G. Clowers, and D.R. Bassett. 2006. "A Novel Method for Using Accelerometer Data to Predict Energy Expenditure." *Journal of Applied Physiology* 100: 1324–1331. Doi: http://dx.doi.org/10.1152/japplphysiol.00818.2005.

Dodd, K.W., P.M. Guenther, L.S. Freedman, A.F. Subar, V. Kipnis, D. Midthune, J.A. Tooze, and S.M. Krebs-Smith. 2006. "Statistical Methods for Estimating Usual Intake of Nutrients and Foods: A Review of the Theory." *Journal of the American Dietetic Association* 106: 1640–1650. Doi: http://doi.org/10.1016/j.jada.2006.07.011.

Ferrari, P., C. Friedenreich, and C.E. Matthews. 2007. "The Role of Measurement Error in Estimating Levels of Physical Activity." *American Journal of Epidemiology* 166: 832–840. Doi: https://doi.org/10.1093/aje/kwm148.

Groves, R.M. and L. Lyberg. 2010. "Total Survey Error: Past: Present, and Future." *Public Opinion Quarterly* 74: 849–879. Doi: https://doi.org/10.1093/poq/nfq065.

Hill, A., J. Roberts, P. Ewings, and D. Gunnell. 1997. "Non-Response Bias in a Lifestyle Survey." *Journal of Public Health* 19: 203–207. Doi: https://doi.org/10.1093/oxfordjournals.pubmed.a024610.

Lessler, J.T. and W.D. Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York: Wiley.

Loustalot, F., S.A. Carlson, J.E. Fulton, J. Kruger, D.A. Galuska, and F. Lobelo. 2009. "Prevalence of Self-Reported Aerobic Physical Activity among US States and Territories–Behavioral Risk Factor Surveillance System, 2007." *Journal of Physical Activity and Health* 6: S9–S17. Doi: http://dx.doi.org/10.1123/jpah.6.s1.s9.

Matthews, C.E. 2002. "Use of Self-Report Instruments to Assess Physical Activity." In *Physical Activity Assessments for Health Related Research*, edited by G.J. Welk, 107–123. Champaign: Human Kinetics.

Matthews, C.E. 2005. "Calibration of Accelerometer Output for Adults." *Medicine and Science in Sports and Exercise* 37: S512–S522. Doi: http://dx.doi.org/10.1249/01.mss.0000185659.11982.3d.

Nusser, S.M., N.K. Beyler, G.J. Welk, A.L. Carriquiry, W.A. Fuller, and B. King. 2012. "Modeling Errors in Physical Activity Recall Data." *Journal of Physical Activity and Health* 9: S56–S67. Doi: http://dx.doi.org/10.1123/jpah.9.s1.s56.

Sallis, J.F. and B.E. Saelens. 2000. "Assessment of Physical Activity by Self-Report: Status, Limitations and Future Directions." *Research Quarterly for Exercise and Sport* 71: 1–14. Doi: http://dx.doi.org/10.1080/02701367.2000.11082780.

Schoenborn, C.A. and M. Stommel. 2011. "Adherence to the 2008 Adult Physical Activity Guidelines and Mortality Risk." *American Journal of Preventive Medicine* 40: 514–521. Doi: http://doi.org/10.1016/j.amepre.2010.12.029.

Smith, C. and D. Nutbeam. 1990. "Assessing Non-Response Bias: A Case Study from the 1985 Welsh Heart Health Survey." *Health Education Research* 5: 381–386. Doi: https://doi.org/10.1093/her/5.3.381.

Tooze, J.A., R.P. Troiano, R.J. Carroll, A.J. Moshfegh, and L.S. Freedman. 2013. "A Measurement Error Model for Physical Activity Level as Measured by a Questionnaire with Application to the 1999–2006 NHANES Questionnaire." *American Journal of Epidemiology* 177: 1199–1208. Doi: http://dx.doi.org/10.1093/aje/kws379.

Troiano, R.P., D. Berrigan, K.W. Dodd, L.C. Masse, T. Tilert, and M. McDowell. 2008. "Physical Activity in the United States Measured by Accelerometer." *Medicine and Science in Sports and Exercise* 40: 181–188. Doi: http://dx.doi.org/10.1249/mss.0b013e31815a51b3.

Tucker, J.M., G.J. Welk, and N.K. Beyler. 2011. "Physical Activity in U.S. Adults: Compliance with the Physical Activity Guidelines for Americans." *American Journal of Preventative Medicine* 40: 454–461. Doi: http://dx.doi.org/10.1016/j.amepre.2010.12.016.

U.S. Department of Health and Human Services. 2008. *2008 Physical Activity Guidelines for Americans*. Atlanta: HHS. Available at: http://health.gov/paguidelines (accessed November 2015).

Van Loon, A.J.M., M. Tijhuis, H.S.J. Picavet, P.G. Surtees, and J. Ormel. 2003. "Survey Non-Response in the Netherlands: Effects on Prevalence Estimates and Associations." *Annals of Epidemiology* 13: 105–110. Doi: http://doi.org/10.1016/S1047-2797(02)00257-0.

Vink, J.M., G. Willemsen, J.H. Stubbe, C.M. Middeldorp, R.S. Ligthart, K.D. Baas, H.J. Dirkzwager, E.J. de Geus, and D.I. Boomsma. 2004. "Estimating Non-Response Bias in Family Studies: Application to Mental Health and Lifestyle." *European Journal of Epidemiology* 19: 623–630. Doi: http://dx.doi.org/10.1023/B:EJEp.0000036814.56108.66.

Warnecke, R.B., T.P. Johnson, N. Chavez, S. Sudman, D.P. O'rourke, L. Lacey, and J. Horm. 1997. "Improving Question Wording in Surveys of Culturally Diverse

Populations." *Annals of Epidemiology* 7: 334–342. Doi: https://doi.org/10.1016/S1047-2797(97)00030-6.

Welk, G.J. 2002. "Introduction to Physical Activity Research." In *Physical Activity Assessments for Health Related Research*, edited by G.J. Welk, 3–18. Champaign: Human Kinetics.

Received January 2016
Revised March 2017
Accepted April 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 551–570, http://dx.doi.org/10.1515/JOS-2017-0026

Effect of Missing Data on Classification Error in Panel Surveys

Susan L. Edwards[1], *Marcus E. Berzofsky*[2], *and Paul P. Biemer*[3]

Sensitive outcomes of surveys are plagued by wave nonresponse and measurement error (classification error for categorical outcomes). These types of error can lead to biased estimates and erroneous conclusions if they are not understood and addressed. The National Crime Victimization Survey (NCVS) is a nationally representative rotating panel survey with seven waves measuring property and violent crime victimization. Because not all crime is reported to the police, there is no gold standard measure of whether a respondent was victimized. For panel data, Markov Latent Class Analysis (MLCA) is a model-based approach that uses response patterns across interview waves to estimate false positive and false negative classification probabilities typically applied to complete data.

This article uses Full Information Maximum Likelihood (FIML) to include respondents with partial information in MLCA. The impact of including partial respondents in the MLCA is assessed for reduction of bias in the estimates, model specification differences, and variability in classification error estimates by comparing results from complete case and FIML MLCA models. The goal is to determine the potential of FIML to improve MLCA estimates of classification error. While we apply this process to the NCVS, the approach developed is general and can be applied to any panel survey.

Key words: Survey error; full information maximum likelihood; measurement error; Markov latent class analysis; national crime victimization.

1. Introduction

Social and behavior science researchers often collect data using questionnaires or instruments consisting of items that purport to measure some underlying construct that is difficult to measure accurately. For example, it is well known that employment status is difficult to measure because it relies on misunderstood concepts such as "looking for work," "temporary layoff" versus "job termination," "temporary work" versus "permanent employment," and so on (see Biemer 2004). Employment classifications are typically based on responses to a series of questions that must be combined to categorize an individual as "employed," "unemployed," or "not in the labor force." Because of the fine

[1] RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709, U.S.A. Email: sedwards@rti.org
[2] RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709, U.S.A. Email: berzofsky@rti.org
[3] RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709, U.S.A. Email: ppb@rti.org
Acknowledgments: The authors would like to thank the National Science Foundation (NSF) for sponsoring this research under award number 1229222. However, we would like to note that the views expressed in this article are those of the authors only and do not reflect the views or position of NSF.

distinctions among these categories or classes, misclassifications that lead to unstable and biased estimates of the class sizes are not uncommon.

A mixture modeling technique called Markov Latent Class Analysis (MLCA) can be used in panel surveys to correct the estimates for misclassification bias. It models wave-to-wave transitions and treats inconsistencies between the data and the model as measurement error or other model errors. MLCA provides estimates of the probabilities of misclassifying people in each labor category, the Wave 1 class probabilities, and the probabilities of transitioning from class to class across waves that have been corrected for misclassification.

A common problem in panel surveys that may limit this analysis is that some respondents fail to respond at one or more panel waves, resulting in an incomplete longitudinal record. This incompleteness poses a problem not only for MLCA but also for standard longitudinal modeling techniques that delete observations with missing time points and analyze only records with no missing values (referred to as case-wise deletion; see, for example, Allison 2001). Two different, although somewhat equivalent, modeling approaches are available to address this missing data problem: imputation and Full-Information Maximum Likelihood (FIML) estimation. One key difference between the two is that imputation replaces the missing values in the record with model-derived values to obtain a complete record that can then be used in a full data set estimation process. FIML, the focus of this article, obtains parameter estimates by maximizing the incomplete data likelihood using completely observed and partially observed cases; that is, all available (full) information. *Multiple* imputation (see, for example, Schafer and Graham 2002; Little and Rubin 2002) is an extension of single imputation that multiply-imputes each missing value to facilitate the computation of imputation variance. It has been shown in Allison (2012) that FIML is equivalent to multiple imputation in the limit as the number of imputations per missing value approaches infinity.

Equally as important as the choice of approach is the assumption that is made for the missing data mechanism itself. Assuming that the data are Missing Completely At Random (MCAR) will lead to bias inferences if response propensities are correlated with the classification error probabilities, which seems common (see, for example, Vermunt 1997; Hess et al. 2013). For example, Biemer (2004) showed that, in the Current Population Survey, people who misreport unemployment may tend to be nonrespondents whose information is often collected by proxy response. Likewise, people who under-report victimizations or who provide erroneous information about their victimizations may also be more likely to fail to respond at some panel wave.

This article demonstrates the importance of compensating for nonresponse in the Latent Class Analysis (LCA) of panel survey data, particularly when making inferences about the measurement components of the model. It shows the importance of including observations that contain missing values on some variables, not only for variance reduction, but also to reduce the bias. We will also show how it is possible to model data that are Missing At Random (MAR) using MLCA combined with FIML models.

Thus, the focus of this article is to explore the effects of methods for compensating for wave nonresponse on the classification error rates in each panel survey wave under the alternative assumptions about the nature of missing data. For this purpose, data collected between 2007 and 2013 from a long-standing national panel survey with indicators of

violent and household-level crime victimization, the National Crime Victimization Survey (NCVS) (U.S. Department of Justice 2015), will be used to fit MLCA models for two types of victimizations: property crimes and violent crimes. Missing data will be modeled simultaneously in an MLCA model under MCAR and MAR missing data assumptions to address two key aims:

(1) Demonstrate the importance of using full information in modeling the structural and measurement components of an MLCA model by determining the effect that missing data have on the MLCA model determined to best fit the data.
(2) Evaluate the effects of alternative assumptions about the missing data mechanism (i.e., MCAR or MAR) on the estimates of misclassification and prevalence.

The remainder of this section provides a brief overview of MLCA models and the basic FIML approach to compensate for nonresponse. Section 2 describes the study data and modeling approach used to address the key aims of this article. In Section 3, the final MLCA model under each missing data mechanism is presented, along with estimates of classification error and crime victimization prevalence over time under MCAR and MAR missing data assumptions. The article concludes in Section 4 with a discussion of the differences across these models, their impact on classification error, thoughts on which model is most appropriate for the NCVS, and ideas for future analysis in this area. Although we apply this process to the NCVS, the approach we develop is general and can be applied to any panel survey.

1.1. Methods for Assessing Measurement Error in Panel Data

Markov Latent Class Models (MLCMs) adjust a panel survey's substantive estimates for the effects of misclassification and, as a byproduct of this process, produce estimates of the "response probabilities". In this application, response probabilities are referred to as classification error parameters because of the interpretation that the latent variable is the true classification. Rather than relying on external realizations of the true or "gold standard" values to estimate measurement error, MLCMs assume a model of the population structure and the measurement distribution parameters to provide maximum likelihood estimates of the parameters of this model. This approach was first introduced with cross-sectional data by Paul Lazarsfeld (1950) as LCA. In 1973, a modification of LCA, MLCA, was proposed by Wiggins (1973) to extend LCA techniques to panel data. Since then, MLCA methodology has been further developed by Poulsen (1982), Van de Pol and De Leeuw (1986), Van de Pol and Langeheine (1990), Dias and colleagues (2008), and Di Mari and colleagues (2016).

Using the notation in Biemer (2011), let X and Y denote two arbitrary random variables having values x and y, respectively. Denote $\Pr(X = x)$ by π_x^X and $\Pr(Y = y|X = x)$ by $\pi_{y|x}^{Y|X}$. Extensions of this notation to three or more variables are straightforward. The MLCM assumes that observations on a latent categorical variable X are subject to classification errors. These models require a minimum of three time points with each time point consisting of a latent variable and an indicator of that latent variable. Let the variable X_t denote the true value of the latent variable (X) at time t and let the observed value Y_t be an indicator of X_t. For purposes of this article, X_t and Y_t are assumed to have the same number

of categories for all time points t. However, extensions to situations where the number of latent and manifest classes differ are straightforward.

The general MLCM contains two components: (1) the *structural component*, which describes the interdependencies between the X_t and the model covariates (referred to as grouping variables because they are categorical variables), and (2) the *measurement component*, which describes the interdependencies among the observations Y_t at each wave $t = 1,...,T$ and their interactions with X_t and other model covariates. Later in the article, a model employing four panel waves will be used in the analysis. However, to simplify the exposition, fix the ideas and establish the notation, here we present the model for three panel waves (i.e., $T = 3$) – the minimum number of panel waves for a MLCM to be identifiable. Extensions to four or more waves are straightforward.

The standard MLCM assumptions for three waves are as follows:

1. *First-Order Markov Property*. $\pi^{X_3|X_1X_2}_{x_3|x_1x_2} = \pi^{X_3|X_2}_{x_3|x_2}$ (i.e., a unit's latent state at Wave 3 (X_3), given its state at Wave 2 (X_2) is independent of its state at Wave 1 (X_1)).
2. *Independent Classification Errors (ICE)*. $\pi^{Y_1Y_2Y_3|X_1X_2X_3}_{y_1y_2y_3|x_1x_2x_3} = \pi^{Y_1|X_1}_{y_1|x_1}\pi^{Y_2|X_2}_{y_2|x_2}\pi^{Y_3|X_3}_{y_3|x_3}$ (i.e., classification errors for the three indicators are mutually independent across waves).
3. *Time-Invariant Classification Errors*. $\pi^{Y_t|X_t}_{y_t|x_t} = \pi^{Y|X}_{y|x}$ for $y = y_t, x = x_t, t = 1,2,3$; classification errors for the indicator Y_t are assumed to be the same for all waves $t = 1,2,3$.
4. *Group-Homogeneous Error Probabilities*. $\pi^{Y_t|X_t}_{y_t|x_t}$ for $t = 1,2,3$ is the same for all units in class $X_t = x_t$ (i.e., within the same latent class, individuals in the same class have equal misclassification probabilities).

Thus, the likelihood kernel for an MLCM with three time points with latent variables X_1, X_2, and X_3 with corresponding indicators Y_1, Y_2, and Y_3 and a single grouping variable G can be expressed as:

$$\mathcal{L}(\pi) = \pi^{GY_1Y_2Y_3}_{gy_1y_2y_3} = \pi^{G}_{g}\sum_{x_1,x_2,x_3}\left(\pi^{X_1|G}_{x_1|G}\pi^{X_2|GX_1}_{x_2|Gx_1}\pi^{X_3|GX_2}_{x_3|Gx_2}\right)\left(\pi^{Y_1|GX_1}_{y_1|Gx_1}\pi^{Y_2|GX_2}_{y_2|Gx_2}\pi^{Y_3|GX_3}_{y_3|Gx_3}\right) \quad (1)$$

where $\pi^{G}_{g}\sum_{x_1,x_2,x_3}(\pi^{X_1|G}_{x_1|g}\pi^{X_2|GX_1}_{x_2|gx_1}\pi^{X_3|GX_2}_{x_3|gx_2})$ is the structural component of the model and $\sum_{x_1,x_2,x_3}\pi^{Y_1|GX_1}_{y_1|gx_1}\pi^{Y_2|GX_2}_{y_2|gx_2}\pi^{Y_3|GX_3}_{y_3|gx_3}$ is the measurement component of the model with $\pi^{Y_t|GX_t}_{y_t|gx_t}$ representing the classification error probabilities at time t with $t = 1,2,3$.

The likelihood kernel presented in (1) can be expressed succinctly using Goodman's (1973) notation for hierarchical models, whereby the model terms for the structural, measurement and nonresponse (if applicable) components are specified in braces using only the highest order interactions. For example, in (1), the structural component can be expressed as a log-linear model $\{GX_1\ GX_1X_2\ GX_2X_3\}$, or as a modified path model as $\{X_1|G\ X_2|X_1G\ X_3|X_2G\}$, and the measurement component as $\{GX_1Y_1\ GX_2Y_2\ GX_3Y_3\}$ or $\{Y_1|X_1G\ Y_2|X_2G\ Y_3|X_3G\}$. Thus Goodman's notation for the likelihood kernel presented in (1) may be expressed either as the log-linear form: $\{GX_1\ GX_1X_2\ GX_2X_3\}\{GX_1Y_1\ GX_2Y_2\ GX_3Y_3\}$ or the modified path model form: $\{X_1|G\ X_2|X_1G\ X_3|X_2G\}\{Y_1|X_1G\ Y_2|X_2G\ Y_3|X_3G\}$. Goodman's notation will be used throughout the rest of the article because it is more succinct. Figure 1 graphically depicts this model in the form of a path diagram.

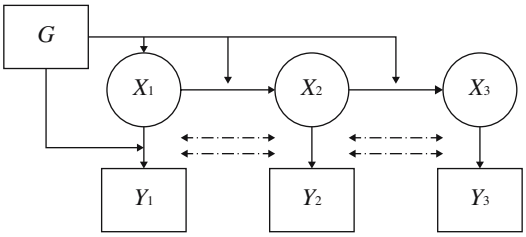

Fig. 1. Illustration of a Markov latent class model with one grouping variable, G. Double arrow denotes equivalence of the response probabilities.

At the tth wave, an indicator of the event (Y_t) is collected, which is a representation of the true value or latent construct X_t. In addition to measurement error, the indicators at Waves 2 and 3 are also subject to attrition (wave nonresponse) and item nonresponse. In Figure 1, circles represent the latent variables, squares represent manifest variables, and arrows denote relationships. An ignorable nonresponse mechanism, defined in more detail below, is assumed for the model.

1.2. Methods for Accounting for Wave Nonresponse in MLCA

When wave nonresponse exists in the indicators or item nonresponse exists in the grouping variables, then the exclusion of cases with one or both types of nonresponse may introduce bias into the model results. When dealing with nonresponse, it is important to understand the nonresponse mechanism and account for it appropriately. Nonresponse is often classified according to one of three missing data mechanisms: MCAR, MAR, or Missing Not At Random (MNAR), also referred to as "nonignorable." Originally defined by Rubin (1976), MCAR occurs when the missing data do not depend on the observed or unobserved data; MAR is less restrictive in that the missing data depend on only the observed data; MNAR is the least restrictive mechanism where the missing data depend on the unobserved data.

Recent work with cross-sectional data suggests benefits of using FIML techniques over listwise deletion, listwise deletion with reweighting, and hot deck imputation to fit a single hypothesized model. FIML methods provide better estimates of variance and are recommended when nonresponse is more than 5% and missing is dependent on the outcome (Allison 2012; Iannacchione 1982). FIML has shown promising results in LCA under a MAR and MNAR missing data mechanism to estimate inmate victimizations over complete case analysis, suggesting that respondents with missing indicators are more likely to be victims (Berzofsky, Biemer, Edwards 2015).

FIML can maintain unbiased inferences on the estimates (Graham 2009; Little and Rubin 2002; Enders 2010). For categorical data analysis, FIML approaches are similar to those developed to handle continuous data – partially observed information is used when fitting log-linear models under the assumption of a multinomial sampling distribution (Vermunt 1997). FIML can handle an MCAR, MAR, or MNAR missing data mechanism (Fay 1986; Vermunt 1997). However, handling MNAR missing data requires knowledge of the MNAR mechanism that is unobservable; this requirement leaves the researcher to formulate a model for the MNAR mechanism for which methods of testing have not been developed (Enders 2010).

In 1982, Fuchs (1982) extended the methodology of FIML to estimate the parameters of a saturated log-linear model using the Estimation-Maximization (EM) algorithm when nonresponse is MAR (Vermunt 1997), thus providing the fifth assumption for the models presented in this article:

5. *Nonresponse is Ignorable.* Nonresponse at each wave is "MAR" in the sense of Little and Rubin (2002).

Thus, the likelihood kernel for the MLCM detailed in Equation 1 can be modified to include dichotomous (0/1) response indicators R_1, R_2, and R_3 that correspond to indictors Y_1, Y_2, and Y_3, respectively, under a MAR mechanism, as follows:

$$\begin{aligned}\mathcal{L}(\pi) &= \pi^{GY_1Y_2Y_3}_{gy_1y_2y_3} \\ &= \pi^{G}_{g}\sum_{x_1,x_2,x_3}\sum_{r_1,r_2,r_3}\left(\pi^{X_1|G}_{x_1|G}\,\pi^{X_2|GX_1}_{x_2|Gx_1}\,\pi^{X_3|GX_2}_{x_3|Gx_2}\right)\left(\pi^{Y_1|GX_1}_{y_1|Gx_1}\,\pi^{Y_2|GX_2}_{y_2|Gx_2}\,\pi^{Y_3|GX_3}_{y_3|Gx_3}\right)\left(\pi^{R_1R_2R_3|GY_1Y_2Y_3}_{r_1r_2r_3|gy_1y_2y_3}\right) \qquad (2)\end{aligned}$$

where the terms $\pi^{R_1R_2R_3|GY_1Y_2Y_3}_{r_1r_2r_3|gy_1y_2y_3}$ determine the response mechanism assumed for the model. Under the assumption of ignorable nonresponse, the log likelihood function can be separated into two additive terms – one involving the parameters of the model in (1) and the other involving the nonresponse parameters. Thus, maximizing the likelihood associated with (1) will produce valid estimators of the MLCM.

In the case of MLCA, the default method for handling nonresponse in LatentGOLD (Vermunt and Magidson 2013) is Fuchs' approach for wave nonresponse and stochastic mean imputation for item nonresponse, making applying this technique straightforward and accessible to researchers. LatentGOLD 5.0 was used for all presented analyses.

2. Methods

2.1. Data: National Crime Victimization Survey

The NCVS is a nationally representative, probability-based household survey of the United States sponsored by the Bureau of Justice Statistics and conducted by the U.S. Census Bureau that gathers information on criminal victimization, reported and not reported to police (Truman and Morgan 2016). The NCVS incorporates a rotating panel design, which uses a stratified multistage cluster sample that includes roughly 50,000 households per sample group with each household interviewed every six months for a total of seven interviews. All households and people aged twelve or older in a rotation group are interviewed about the number and characteristics of victimizations experienced during the previous six months.

For this article, we focused on property crime and violent crime victimizations. For property crime, there is a single household respondent. For violent crime, each eligible person in the household responds. Because of the rareness of certain crimes and structure of the NCVS, multiple crime types were collapsed into a single violent or property victimization indicator at each wave. Violent crimes consisted of rape and sexual assault, aggravated assault, robbery, and simple assault; property crimes consisted of household

burglary, motor vehicle theft, and theft. By collapsing, we gained model stability and avoided sparseness in the grouping classifications (Berzofsky and Biemer 2017).

The NCVS implements a two-phase approach to identify and enumerate victimizations. During the first phase of the interview, a screener is used to identify experiences with crime during the six-month reference period. The second phase of the interview is a detailed follow-up for each victimization identified during the screening phase. Indicators of specific types of victimization are created as a composite from various questions. Regarding household crimes, the respondent was asked about property break-ins or attempts and motor vehicle theft in the last six months in various scenarios (e.g., "did anyone steal gas from (it/them)"). At the person level, respondents are asked questions about victimization attacks and provided cues (e.g., location, weapon). For example, the question on theft with location cues was worded "since _____, were you attacked or threatened or did you have something stolen from you," and some of the cues provided were "at home including porch or yard" and "at work or school."

The amount of wave nonresponse observed in the crime victimization indicators at each wave of the study is more than 35% for violent crimes and less than 13% for household crimes. Wave nonresponse rates observed during the first four waves for the property and violent crime victimization indicators are presented in Table 1. Among typical reasons for nonresponse, the NCVS has two special considerations that may cause nonresponse during an individual wave. First, people may move out of a household. If an address is empty during the time of the interview, then the household and its members will have missing values for the wave. Second, new or newly eligible people may move into an existing household (e.g., a child turning twelve, a college graduate moving in with his or her parents at some point after the initial wave). In this case, the new or newly eligible person will have missing values for previous waves when they were either not in the household or ineligible. The NCVS does have unit-level response rates in the high 80% range at the person level (see, for example, Truman and Morgan 2016).

For our analysis, we limited the NCVS data to include panel and rotation groups for which all seven waves had occurred, resulting in data collected between 2007 and 2013. For these panel and rotation groups, all people and households in which at least one wave was completed were included in the analysis. Typically, multiple years of data would be pooled to reduce the standard errors of estimates, making the estimates more reliable. However, the NCVS public use files limit the number of years that can be pooled, because the household identifier was scrambled in 2006 when the new census primary sampling

Table 1. Crime victimization indicator wave nonresponse.

	Any	Wave 1	Wave 2	Wave 3	Wave 4
Violent					
Missing	110,236	58,935	58,960	58,520	56,955
Non-missing	51,635	102,936	102,911	103,351	104,916
Wave nonresponse rate (%)	68.10	36.41	36.42	36.15	35.19
Property					
Missing	47,713	8,037	7,929	8,560	8,779
Non-missing	34,678	56,423	57,339	58,360	59,434
Wave nonresponse rate (%)	57.91	12.47	12.15	12.79	12.87

units were integrated into the sample design. As MLCA requires linking households and people across time, the scrambling of the identifier limits the number of years that can be pooled. The issue of sparse cell sizes (i.e., model cells with zero or near-zero counts) can cause difficulties with model convergence (Biemer 2011; Bartolucci et al. 2013). Therefore, only the first four waves were used for the violent and property crime victimization analysis. This focus resulted in a total of 161,871 people and 68,213 households. Among these people, the number with an observed violent crime victimization was less than 1.5 percent, and among these households, the number with an observed property crime victimization was less than nine percent. Observed crime victimization prevalence is presented in Table 2.

It is expected that the initial interview would have larger victimization rates compared with the later waves because it is unbounded and respondents may "telescope" by recalling incidents that occurred before the six-month reference period. Despite this consideration, Wave 1 data were included to be consistent with the NCVS, which, beginning in 2006, included Wave 1 responses in the published estimates (Rand and Catalano 2007). Data gathered in Wave 2 may be considered the most accurate because they are from the first bounded interview with the least amount of fatigue; however, being the first bounded interview does not imply a gold standard because the data can still suffer from other sources of measurement error (e.g., interviewer bias, questionnaire wording).

2.2. *Modeling Approach*

We followed the modeling strategy that worked best on most tested models as discussed by Berzofsky and Biemer (2017) (see also Biemer 2011), which consisted of two main steps. First, grouping variables were identified with a forward selection approach using the Bayesian Information Criterion (BIC) to identify when each grouping variable should

Table 2. Observed crime victimizations in the NCVS.

	Wave			
	1	2	3	4
Violent victimization				
Unweighted				
Victims	1,295	844	801	710
Non-victims	101,641	102,067	102,550	104,206
Weighted				
% Victimization	1.36	0.89	0.83	0.73
Standard error	0.05	0.04	0.03	0.03
Property victimization				
Unweighted				
Victims	5,199	3,524	3,299	3,173
Non-victims	59,261	61,744	63,621	65,040
Weighted				
% Victimization	8.19	5.48	4.99	4.68
Standard error	0.17	0.12	0.12	0.10

enter the model. Grouping variables create mutually exclusive groups whereby the classification error rates are homogenous within each group; grouping variables are further discussed in the following paragraph. These variables were added to the structural and measurement models. Likelihood ratio tests were used to determine the most parsimonious base model that removed group heterogeneity and met the MLCM assumptions: first-order Markov, ICE, time-invariant classification errors, and group-homogeneous error probabilities. Second, using the base model from step 1, all remaining MLCM assumptions were tested and relaxed according to the following procedure: (1) models with boundary or convergence issues that might make the model unstable were rejected, and (2) for models without estimation issues, results from likelihood ratio tests for nested models and BIC for non-nested models were used to select the final model.

The NCVS collects information on 14 grouping variables: twelve personal or household-level variables and two para-data variables (U.S. Department of Justice 2015). These 14 grouping variables formed the foundation of grouping variables considered for our models. Grouping variables were classified as time varying or time invariant (Bartolucci et al. 2013). Time-invariant grouping variables were those where fewer than five percent of respondents changed status from the first observed value to the last observed value; age category was an exception to this rule. Time-invariant grouping variables were defined by the first observed value. To reduce the complexity of the model and get parsimony without sacrificing fit, time-varying grouping variables were defined by the creation of an additional category to capture the "movers" who, regardless of movement direction, had similar classification error rates (Berzofsky and Biemer 2017). Because of low item nonresponse rates in all but one of the grouping variables (less than four percent), grouping variables were imputed before MAR analysis with a stochastic mean imputation technique, the default imputation method for covariates in LatentGOLD. Grouping variables considered for the violent and property victimization models with item nonresponse rates are detailed in Table 3.

One challenge of conducting MLCA with complex survey data is that one or more assumptions may be violated because of the sample design (Biemer 2011). For the structural component assumptions, first-order Markov models were tested against second-order Markov models, models where transition probabilities are assumed to depend on the previous two time points, and Mover-Stayer models, models with an additional latent construct to identify persons or households whose victimization status is constant (stayer) or changes (mover) over time (Goodman 1961). Time-invariant classification error rates were tested by relaxing assumptions on the coefficients for each time point. For the measurement component assumptions, group-homogeneous error probabilities were tested by relaxing assumptions on the coefficients for each indicator; ICE assumptions were tested using bivariate residual analysis to identify dependent indicators (Vermunt and Magidson 2013).

Table 4 highlights the various models that were compared using Goodman's notation for hierarchical models. In Table 4, X_1 to X_4 represent the latent construct of victimization (violent or property) at each wave; Y_1 to Y_4 represent indicator 1 through indicator 4, respectively; A represents marital status, B represents age category, C represents household ownership, D represents household size category, E represents age category of the oldest person in the household, F represents urbanity, and M is a latent construct to capture movement.

Table 3. Crime victimization grouping variable item nonresponse.

	Missing	Non-Missing	Item Nonresponse Rate (%)
Violent			
Age category[1,3]	0	161,871	0.00
Education[1]	3,756	158,115	2.32
Gender[1]	0	161,871	0.00
Household size category[2]	5,437	156,434	3.36
Household ownership[1,3]	0	161,871	0.00
Interview type (in person/phone)	0	161,871	0.00
Marital status[1,3]	1,506	160,365	0.93
Number of in person interviews	0	161,871	0.00
Proxy answered interview	0	161,871	0.00
Race category[1]	0	161,871	0.00
Urbanity[1]	0	161,871	0.00
Property			
Age category for oldest in household[1,3]	0	82,391	0.00
Household income[2]	18,768	63,623	22.78
Household size category[2,3]	0	82,391	0.00
Household ownership[1]	0	82,391	0.00
Interview type – all in person	0	82,391	0.00
Interview type – all/some/none in person	0	82,391	0.00
Number of in person interviews	0	82,391	0.00
Race category for oldest in household[1]	0	82,391	0.00
Urbanity[1,3]	0	82,391	0.00

[1] First observed value used for analysis.
[2] Time varying variable with single "mover" category.
[3] Grouping variable used in violent or property victimization model.

Once the final model was determined from MCAR and MAR analysis according to the approach detailed previously, models were fit to each category of victimization – violent and property. Each model was applied to two data sets:

(1) MCAR analysis using complete case data (e.g., listwise deletion) and
(2) MAR analysis using the Fuchs FIML approach on the outcome (victimization) and mean imputation on the grouping variables.

Thus, a total of four models were used to address the aims of this article:

(1) violent victimization MCAR model applied to the person-level MCAR data set,
(2) violent victimization MAR model applied to the person-level MAR data set,
(3) property victimization MCAR model applied to the household-level MCAR data set, and
(4) property victimization MAR model applied to the household-level MAR data set.

LatentGOLD software was used for all analyses in this report; LatentGOLD addresses the issue of clustering and weighting through a pseudo-maximum likelihood technique and

Table 4. Models considered.

	Violent models	Property models
Base grouping variable model with all MLC assumptions	$\{X_1A\ X_1B\ X_1C\ X_tX_{t-1}A\ X_tX_{t-1}B\ X_tX_{t-1}C\}$ (for t = 2, 3, 4) $\{Y_tX_tA\ Y_tX_tB\ Y_tX_tC\}$ (for t = 1, 2, 3, 4)	$\{X_1D\ X_1E\ X_1F\ X_tX_{t-1}D\ X_t\ X_{t-1}E\ X_tX_{t-1}F\}$ (for t = 2, 3, 4) $\{Y_tX_tD\ Y_tX_tE\ Y_tX_tF\}$ (for t = 1, 2, 3, 4)
Time-invariant classification error assumption relaxed	$\{X_1A\ X_1B\ X_1C\ X_tX_{t-1}A\ X_tX_{t-1}B\ X_tX_{t-1}C\}$ (for t = 2, 3, 4) $\{Y_tX_tA\ Y_tX_tB\ Y_tX_tC\}$ (for t = 1, 2, 3, 4)	$\{X_1D\ X_1E\ X_1F\ X_tX_{t-1}\ X_tX_{t-1}D\ X_tX_{t-1}E\ X_tX_{t-1}F\}$ (for t = 2, 3, 4) $\{Y_tX_tD\ Y_tX_tE\ Y_tX_tF\}$ (for t = 1, 2, 3, 4)
First-order markov property assumption relaxed		
Second order markov	$\{X_1A\ X_1B\ X_1C\ X_2X_1A\ X_2X_1B\ X_2X_1C\ X_tX_{t-1}X_{t-2}A\ X_tX_{t-1}X_{t-2}B\ X_tX_{t-1}X_{t-2}C\}$ (for t = 3, 4) $\{Y_tX_t\ Y_tX_tA\ Y_tX_tB\ Y_tX_tC\}$ (for t = 1, 2, 3, 4)	$\{X_1D\ X_1E\ X_1F\ X_2X_1D\ X_2X_1E\ X_2X_1F\ X_tX_{t-1}X_{t-2}D\ X_tX_{t-1}X_{t-2}E\ X_tX_{t-1}X_{t-2}F\}$ (for t = 3, 4) $\{Y_tX_tD\ Y_tX_tE\ Y_tX_tF\}$ (for t = 1, 2, 3, 4)
Mover stayer	$\{X_1A\ X_1B\ X_1C\ X_1M\ X_tX_{t-1}A\ X_tX_{t-1}B\ X_tX_{t-1}C\ X_tX_{t-1}M\}$ (for t = 2, 3, 4) $\{Y_tX_tA\ Y_tX_tB\ Y_tX_tC\}$ (for t = 1, 2, 3, 4)	$\{X_1D\ X_1E\ X_1F\ X_1M\ X_tX_{t-1}D\ X_tX_{t-1}E\ X_tX_{t-1}F\ X_tX_{t-1}M$ (for t = 2, 3, 4) $\{Y_tX_tD\ Y_tX_tE\ Y_tX_tF\}$ (for t = 1, 2, 3, 4)
Group-homogeneous error probabilities assumption relaxed		
Wave 1 different	$\{X_1A\ X_1B\ X_1C\ X_2X_1A\ X_2X_1B\ X_2X_1C\ X_tX_{t-1}X_{t-2}A\ X_tX_{t-1}X_{t-2}B\ X_tX_{t-1}X_{t-2}C\}$ (for t = 3, 4) $\{Y_tX_t\ Y_tX_tA\ Y_tX_tB\ Y_tX_tC\}$ (for t = 1, 2, 3, 4)	$\{X_1D\ X_1E\ X_1F\ X_1M\ X_tX_{t-1}D\ X_tX_{t-1}E\ X_tX_{t-1}F\ X_tX_{t-1}M\}$ (for t = 2, 3, 4) $\{Y_tX_tD\ Y_tX_tE\ Y_tX_tF\}$ (for t = 1, 2, 3, 4)
All waves different	$\{X_1A\ X_1B\ X_1C\ X_2X_1A\ X_2X_1B\ X_2X_1C\ X_tX_{t-1}X_{t-2}A\ X_tX_{t-1}X_{t-2}B\ X_tX_{t-1}X_{t-2}C\}$ (for t = 3, 4) $\{Y_tX_t\ Y_tX_tA\ Y_tX_tB\ Y_tX_tC\}$ (for t = 1, 2, 3, 4)	$\{X_1D\ X_1E\ X_1F\ X_1M\ X_tX_{t-1}D\ X_tX_{t-1}E\ X,X_{t-1}F\ X_tX_{t-1}M\}$ (for t = 2, 3, 4) $\{Y_tX_tD\ Y_tX_tE\ Y_tX_tF\}$ (for t = 1, 2, 3, 4)

MLC = Markov latent class.

addresses nonresponse through applying FIML and stochastic mean imputation to categorical data analysis. The ability to apply Fuchs' FIML approach is built into the software as the default method for addressing nonresponse on the dependent variables. For independent variables, LatentGOLD applies stochastic mean imputation by default. In regard to MLCA, FIML is used to address wave nonresponse in the indicators, and stochastic mean imputation is used to address item nonresponse in the grouping variables.

3. Results

The aims of this article are (1) to demonstrate the importance of using full information in an MLCM and (2) to evaluate the effect MCAR and MAR missing data assumptions have on MLCA model estimates of misclassification and prevalence. Subsection 3.1 provides details on the model fitting process and final models used in our analysis for both victimizations (violent and property). Subsection 3.2 compares estimates of misclassification from MCAR and MAR MLCMs for each type of victimization. Subsection 3.3 compares prevalence estimates from the structural component of MCAR and MAR MLCMs for each type of victimization.

3.1. Modeling Results

With respect to victimization type, models with and without missing data identified the same grouping variables and relaxed the same MLCA assumptions, resulting in identical final models. Table 5 presents victimization model diagnostics for violent and property crime victimization. The dissimilarity index indicates the percentage of data that would need to change cells for the model to fit perfectly; it is an alternative way to assess the fit of the model. Full measurement models for violent and property crime victimization are given in Supplemental data, Appendix A (available online at: http://dx.doi.org/10.1515/JOS-2017-0026). Complete LatentGOLD syntax for model estimation of violent and property crime victimizations is given in Supplemental data, Appendix B (available online at: http://dx.doi.org/10.1515/JOS-2017-0026). Subsections 3.1.1 and 3.1.2 provide specific details on the base and final models for violent and property crime victimization, respectively.

3.1.1. Violent Crime Victimization Modeling Results

The violent crime victimization final model without missing data (i.e., after listwise deletion, respondents without missing indicators or grouping variables) included 51,528 cases, 31.8% of all respondents. The base model found three grouping variables to be significant in the measurement component of the MLCM – first observed value of marital status, household ownership, and first observed value of categorized age. When missing data were included, the same grouping variables were found to be significant. The identified grouping variables are listed in Table 3.

For violent crime victimizations, a full model with interaction terms between the grouping variables and the latent wave indicator of victimization status was deemed appropriate, and several MLCA assumptions were relaxed, regardless of the missing data assumption. Our models were able to relax model assumptions because four time points were used in the models. Based on the bivariate residual test, the ICE assumption was not

Table 5. Model fitting statistics.

Attribute of model	Missing data approach	# of parameters	Degrees of freedom	Log-likelihood	BIC	Dissimilarity index[1]
VIOLENT VICTIMIZATION						
Base model	MCAR	35	325	−7161	14703	0.0029
	MAR	35	1812	−20439	41299	0.0075
Final model	MCAR	93	267	−7099	15207	0.0019
	MAR	93	1754	−20363	41841	0.0071
PROPERTY VICTIMIZATION						
Base model	MCAR	45	630	−38196	76878	0.0166
	MAR	45	2594	−56442	113393	0.0392
Final model	MCAR	101	574	−38147	77384	0.0146
	MAR	101	2538	−56367	113876	0.0370

[1]Formula for dissimilarity index. $D = \sum_i |n_i - \hat{m}_i|/(2N)$, where n_i = observed cell count, $\hat{m}_i$ = estimated expected cell count, N = # of observations, i = cell identifier.

violated. The final MCAR and MAR violent victimization models consisted of a second-order Markov model with varying covariates for the observed victimizations and varying classification errors between the first wave and all following waves.

3.1.2. Property Crime Victimization Modeling Results

The property crime victimization final model without missing data included 48,590 cases, 59.0% of all responding households. The base model found three grouping variables to be significant in the measurement component of the MLCM – categorized household size, first observed value of categorized age of oldest household member, and urbanity. As with violent crime victimization when missing data were included, the same grouping variables found to be significant in the MCAR models were also identified in the MAR models.

The property crime victimization model experienced similar MLCM assumption violations as the violent crime victimization model. The base model for property crime victimization consisted of a full model with main effects and interaction terms between the grouping variables and the latent wave indicator of victimization status. The time homogeneous classification error, first-order Markov assumptions, and group homogeneous classification error assumptions were relaxed. The final MCAR and MAR property crime models consisted of a mover-stayer full model with varying covariates for the observed victimizations and varying classification errors between the first wave and all following waves.

3.2. Estimated Misclassifications

Now we use the second-order MLCM and the mover-stayer MLCM to create estimates of misclassification and prevalence by fitting the measurement and structural components with each missing data assumption (MCAR, MAR). The measurement component provides estimates of false positive and false negative rates at each time point. False positive rates measure the probability of respondents identifying as victims when in truth they are nonvictims (i.e., $P(Y_t = 1|X_t = 2)$). False negative rates result from respondents identifying as nonvictims when in truth they are victims (i.e., $P\big(Y_t = 2|X_t = 1\big)$). Trends of estimated false positive and false negative rates for violent and property crime victimization at each wave of the NCVS are presented in Figures 2 and 3, respectively, with 95% confidence intervals represented by error bars.

From Figure 2, it is clear that regardless of model type, the false positive rates for violent victimizations are larger for the first interview. These larger rates are probably the

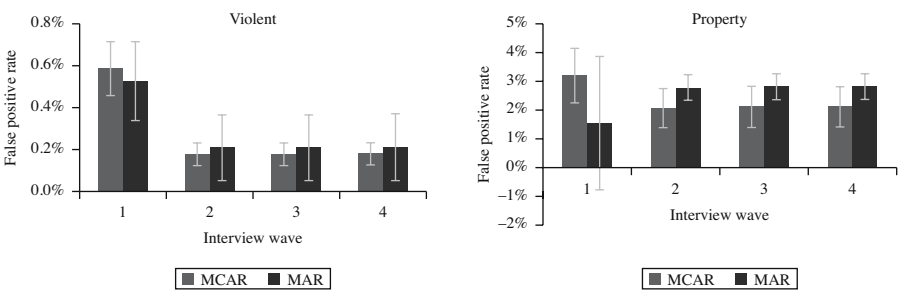

Fig. 2. False positive rates for violent and property crime victimization.

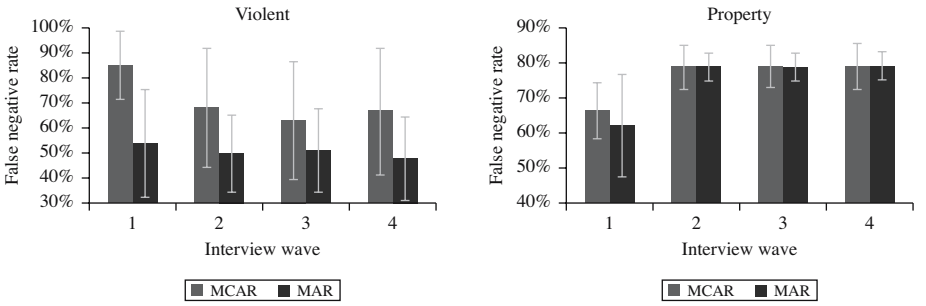

Fig. 3. False negative rates for violent and property crime victimization.

result of telescoping (as noted earlier, the first interview used for estimation is unbounded, whereas all follow-up interviews are bounded). This finding also held for the MCAR property victimization model, but not the MAR property victimization model. False positive rates are low regardless of model and victimization type: less than one percent for violent and less than five percent for property victimizations.

Based on false positive rates, victimization type appears to affect the results from the MCAR and MAR models after the first wave in different manners. Both MAR models yielded higher estimates of false positive rates compared to the corresponding MCAR estimates. For violent crimes, the MCAR and MAR models yield similar estimates. For property crimes, the MAR model yields estimates near the upper end of the 95% confidence interval of the MCAR model estimates; MAR estimates for false positive rates are larger than the MCAR estimates at every wave, except Wave 1, by roughly 0.7%.

From Figure 3, the manner in which false negative rates change over time for violent victimizations differs depending on the mechanism for missing response. Under the MCAR model, the false negative rate is significantly higher in the first interview (85%) compared with the later interviews ($\approx$65%); however, for the MAR model, the false negative rate is statistically unchanged across the four periods ($\approx$51% in all waves). This result is an indication that the inclusion of those who do not respond helps control for differences in the false negative rate over time. For property victimization, although there appears to be an increase in the false negative rate from interview Wave 1 (66% for MCAR and 62% for MAR) compared with the later waves ($\approx$80% for all waves for MCAR and MAR) regardless of the missing data mechanism, these differences are not statistically significant.

Interestingly, our results do not detect an increase in the false negative rate in interview wave 4. Some research (see, for example, Hart et al. 2005) has shown that respondent fatigue occurs in later waves of the NCVS. Respondent fatigue is likely to increase the classification error rates over time. One possible reason that our models do not demonstrate this pattern is because we limited our analysis to the first four interview waves. Hart and colleagues (2005) looked at all seven waves, finding respondent fatigue to have its greatest effects in Waves 6 and 7, which are not included in our current analysis.

Perhaps due to the less sensitive nature of property crimes and a more engaged respondent, the estimated false negatives for property crime victimization at each wave of the NCVS by model type show different trends than those for violent crimes. Estimates

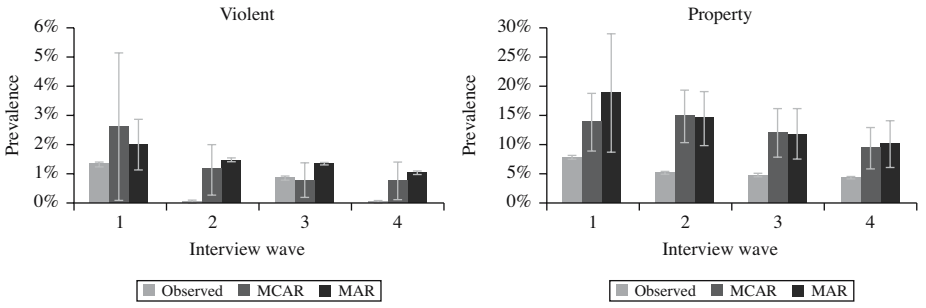

Fig. 4. Prevalence rates for violent and property crime victimization.

differed by 4.1% at the first wave but were similar during following waves, with MAR estimates being slightly higher by at most 0.2%. The false positive and false negative rate estimates are somewhat consistent across waves with respect to model type.

3.3. Estimated Prevalence

Victimization prevalence is measured in the structural component of the MLCM. Estimates of violent and property victimization were computed three ways:

(1) based on observed responses (i.e., direct estimates from the data set),
(2) based on the MCAR model, and
(3) based on the MAR model are presented in Figure 4, with 95% confidence intervals represented by error bars.

As with the classification error rates, standard errors are larger for the FIML methods compared with the complete case analysis. If the missingness is MCAR, then we would expect the standard errors from the MAR model to be smaller, because the MAR model uses more information than the MCAR model. If the missingness is MAR, increased standard errors are to be expected because missing values contribute more variance to the final model. Estimates differ between model types for violent crime victimizations, suggesting that missing data do affect estimates of prevalence. FIML models estimate violent crime prevalence to be higher than the observed and complete case analysis. Prevalence rates for either victimization are highest during the first wave; this finding may be attributed to telescoping because the initial wave is unbounded, which inflates the number of reported victimizations (U.S. Census Bureau 2014).

Overall estimates of property crime victimization are similar across model types, with the MAR model differing the most from the MCAR model during the first wave by 5.1%. For violent crime victimization, for all but the first wave, the MAR estimates are higher than the MCAR estimates. The FIML MAR model appears to be correcting for respondent fatigue by keeping the violent crime prevalence consistent across waves.

4. Discussion

In this article, we fit two different types of models for the response mechanism in NCVS data. One model (MCAR) was fit in a complete case analysis that included only records

with no missing values on all the victimization indicators or grouping variables across all four waves. This model excluded about 70% of the cases for violent victimization and about 40% of the cases for property victimization. The other model (MAR) used FIML techniques to account for missing data in the indicators and mean imputation to account for missing data in the grouping variables. This model included all cases that responded in one or more waves. Estimates of classification error rates and prevalence rates were produced from both models. The MAR model attempts to compensate for any bias that could be introduced into the analysis by excluding the missing cases. MAR models assume that the missing data mechanism does not depend on the variable that is missing but may depend on other influencing factors that can be modeled using additionally observed variables.

A third type of missing data mechanism, MNAR, can also be modeled using FIML techniques. This type of model assumes that the missing data mechanism associated with the outcome variable (i.e., victimization indicator) depends on that same outcome variable. However, MNAR FIML models are difficult to apply with existing software, and there are trade-offs in doing so. MNAR model estimates will often have larger variances, which may offset any gains in reducing nonresponse bias. The software we used experienced issues with EM convergence and local minima, leading to model instability. Besides being more difficult to program, MNAR models that may be specified can be limited by the computer's memory capabilities. In our case, 16 gigabytes of RAM were not sufficient to run some models. As a result, the MNAR models we fit resulted in implausible estimates, which were most likely due to weak identifiability and local minima (Bartolucci et al. 2013; Biemer 2011). Given the poor performance of the MNAR models, those results were not included in this article.

MAR estimates that differ considerably from MCAR estimates usually indicate that the MCAR assumption is untenable; thus, excluding the cases with missing data from the analysis will yield biased estimates. For violent and property crime, MAR models produced substantially different estimates from the MCAR models. For violent crime victimization, FIML MAR estimates of prevalence were higher than MCAR estimates at all but the first wave. For property crime victimization, MAR and MCAR estimates of prevalence were similar in all but the first wave. This result suggests that nonrespondents are more likely to be victims of violent crime but perhaps not property crime.

As previously noted, the purpose of this article was to demonstrate that MLCMs can be used to account for measurement error and nonresponse and to evaluate the differences between MLCMs with and without missing data. However, further research is needed. For example, the nonresponse bias implied by the MAR models for violent crimes presents an intriguing finding, namely, omitting respondents with wave and/or item nonresponse from the analysis of violent crime could substantially bias the results. These models would benefit from further refinement and verification. Although item nonresponse was minimal in our final models, future research could treat the two types of nonresponse (wave and item) differently because the mechanism that causes an individual to opt out at a wave may be different than that which causes an individual to not respond to an item in the interview.

One opportunity to develop qualitative research to support our findings would be through cognitive interviewing. We hypothesize that much of the measurement error from wave to wave is due to comprehension error, recall error, or respondent fatigue (also

known as "satisficing"). Perhaps evidence of these errors can be found using cognitive interviewing techniques where respondent conditions that give rise to these error sources could also be explored. Another method for verifying our findings would be via a simulation study. Using Monte Carlo simulation, multiple data sets, each with a unique and known nonresponse mechanism, could be generated from the current NCVS data to determine how various types of nonresponse errors manifest themselves as biases in victimization estimates. In addition, the simulations could also investigate the extent to which measurement errors affect the application of MAR and MNAR nonresponse models. A variation on the simulation study could explore the validity of the models by generating data sets with varying levels of nonresponse and measurement errors. Then the results from the models could be compared with the known model-generating parameters.

The data themselves presented a few unique challenges. Because of the design of the NCVS, it is difficult to pool data from a larger time period. We could pool only panels that started data collection in 2007 because of issues of household and person ID linkage related to the scrambled household identifiers introduced in 2006. This small sample could be contributing to the large standard errors observed for the model-based estimates. Pooling data from a larger time period would increase the sample size, which could then result in more stable models.

The problems with small samples were compounded by the fact that crime victimization is a rare event, which resulted in few positives in the sample on which to build a model for misclassification of positives. We addressed this issue by combining crime types into two distinct categories – property crimes and personal crimes – to build up their prevalence. In 2014, the overall rate of violent crimes was 20.1 per 1,000 people aged twelve or older; rape and sexual assault crimes accounted for just 1.1 per 1,000 people aged twelve or older of the overall rate (Langton and Truman 2015). For this reason even with 15 years of data, standard errors could still be large, particularly for false negative estimates. In addition, analyzing 15 years of data may expose other issues such as temporal changes in definitions of certain types of crime or crime reporting over time. Our analysis excluded data collected from the later waves (i.e., Waves 5, 6, and 7). This exclusion was done primarily to reduce the number of sparse cells due to cross-classifying responses from seven waves, which could be compounded because of potentially greater respondent fatigue in later waves.

The goal of this analysis is not to quantify all of the errors present in the NCVS, but to show a model-based way of addressing two types of errors and the effect nonresponse can have on model estimates. Despite the limitations of the data, our findings demonstrated that excluding respondents with missing data may bias estimates of prevalence.

5. References

Allison, P.D. 2001. "Missing Data." In *Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136*. Thousand Oaks, CA: Sage.

Allison, P.D. 2012. "Handling Missing Data by Maximum Likelihood." In Proceedings of SAS Global Forum 2012, Statistics and Data Analysis, April 22–25, 2012. 312. Haverford, PA: SAS Institute. Available at: http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf (accessed August 2016).

Bartolucci, F., A. Farcomeni, and F. Pennoni. 2013. *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: CRC Press.

Berzofsky, M.E., P.P. Biemer, and S.L. Edwards. 2015. "Latent Class Analysis with Missing Data under Complex Sampling: Results of a Simulation Study." Presented at 60th World Statistics Conference, July 26–31, 2015. Rio de Janeiro, Brazil: World Statistics Conference.

Berzofsky, M. and P.B. Biemer. 2017. "Classification Error in Crime Victimization Surveys: A Markov Latent Class Analysis." In *Total Survey Error in Practice*, edited by P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West, 387–412. Hoboken, NJ: Wiley.

Biemer, P.P. 2004. "An Analysis of Classification Error for the Revised Current Population Survey Employment Questions." *Survey Methodology* 30(2): 127–140.

Biemer, P.P. 2011. *Latent Class Analysis of Survey Error*. Hoboken, NJ: Wiley.

Di Mari, R., D.L. Oberski, and J.K. Vermunt. 2016. "Bias-Adjusted Three-Step Latent Markov Modeling with Covariates, Structural Equation Modeling." *Structural Equation Modeling* 23(5): 649–660. Doi: http://dx.doi.org/10.1080/10705511.2016.1191015.

Dias, J.G., J.K. Vermunt, and S. Ramos. 2008. "Heterogeneous Hidden Markov Models." In *Compstat 2008 Proceedings*, August, 2008. City, State: Compstat. Available at: http://members.home.nl/jeroenvermunt/dias2008.pdf (accessed March 2015).

Enders, C.K. 2010. *Applied Missing Data Analysis*. New York: Guilford Press.

Fay, R.E. 1986. "Causal Models for Patterns of Nonresponse." *Journal of the American Statistical Association* 81(394): 354–365. Doi: http://dx.doi.org/10.1080/01621459.1986.10478279.

Fuchs, C. 1982. "Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data." *Journal of the American Statistical Association* 77(378): 270–278. Doi: http://dx.doi.org/10.2307/2287230.

Goodman, L.A. 1961. "Statistical Methods for the Mover-Stayer Model." *Journal of the American Statistical Association* 56(296): 841–868. Doi: http://dx.doi.org/10.2307/2281999.

Goodman, L.A. 1973. "The Analysis of Multidimensional Contingency Tables when Some Variables are Posterior to Others: A Modified Path Analysis Approach." *Biometrika* 60(1): 179–192. Doi: http://dx.doi.org/10.2307/2334920.

Graham, J.W. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology* 60: 549–576. Doi: http://dx.doi.org/10.1146/annurev.psych.58.110405.085530.

Hart, T.C., C.M. Rennison, and C. Gibson. 2005. "Revisiting Respondent 'Fatigue Bias' in the National Crime Victimization Survey." *Journal of Quantitative Criminology* 21(3): 345–363. Doi: http://dx.doi.org/10.1007/s10940-005-4275-4.

Hess, S., N. Sanko, J. Dumont, and A. Daly. 2013. "A Latent Variable Approach to Dealing with Missing or Inaccurately Measured Variables: The Case of Income." In Proceedings of the Third International Choice Modelling Conference, July 3–5, 2013. Sydney, Australia: ICM Conference. Available at: http://www.icmconference.org.uk/index.php/icmc/ICMC2013/paper/viewFile/744/233 (accessed August 2015).

Iannacchione, V. 1982. "Weighted Sequential Hot Deck Imputation Macros." In Proceedings of the SAS Users Group International Conference, February 14–17, 1982.

759–763. San Francisco, CA. Available at: http://www.sascommunity.org/sugi/SUGI82/Sugi-82-139%20Iannacchione.pdf (accessed March 2015).

Langton, L. and J. Truman. 2015. *Criminal Victimization, 2014*. Washington, DC: Bureau of Justice Statistics. (NCJ 248973).

Lazarsfeld, P.F. 1950. "The Logical and Mathematical Foundation of Latent Structure Analysis." In *Studies on Social Psychology in World War II, Vol. 4, Measurement and Prediction*, edited by S. Stauffer, E.A. Suchman, P.F. Lazarsfeld, S.A. Starr, and J. Clausen. Princeton, NJ: Princeton University Press.

Little, R.J. and D.B. Rubin. 2002. *Wiley Series in Probability and Statistics: Statistical Analysis with Missing Data*. 2nd ed. Somerset, NJ: Wiley.

Poulsen, C.A. 1982. *Latent Structures Analysis with Choice Modeling Applications*. Aarhus, Denmark: Aarhus School of Business Administration and Economics.

Rand, M. and S. Catalano. 2007. *Criminal Victimization, 2006*. Washington, DC: U.S. Department of Justice, Office of Justice Programs. (NCJ 219413).

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63(3): 581–592. Doi: http://dx.doi.org/10.1093/biomet/63.3.581.

Schafer, J.L. and J.W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7(2): 147–177. Doi: http://dx.doi.org/10.1037//1082-989x.7.2.147.

Truman, J.L. and R.E. Morgan. 2016. *Criminal Victimization, 2015*. Washington, DC: Bureau of Justice Statistics. (NCJ 250180).

U.S. Census Bureau. 2014. *National Crime Victimization Survey: Technical Documentation*. Washington, DC: U.S. Census Bureau. (NCJ 247252).

U.S. Department of Justice. 2015. *Bureau of Justice Statistics. National Crime Victimization Survey, 2014*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Van de Pol, F. and J. de Leeuw. 1986. "A Latent Markov Model to Correct for Measurement Error." *Sociological Methods & Research* 15: 118–141. Doi: http://dx.doi.org/10.1177/0049124186015001009.

Van de Pol, F. and R. Langeheine. 1990. "Mixed Markov Latent Class Models." In *Sociological Methodology*, edited by C.C. Clogg, 213–247. Oxford: Blackwell.

Vermunt, J.K. 1997. *Log-Linear Models for Event Histories*. London: Sage.

Vermunt, J.K. and J. Magidson. 2013. *Technical Guide to Latent Gold 5.0: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations.

Wiggins, L.M. 1973. *Panel Analysis, Latent Probability Models For Attitude And Behavior Processing*. Amsterdam: Elsevier SPC.

Received January 2016
Revised April 2017
Accepted April 2017