Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Improving measurement error and representativeness in nonprobability surveys

by Aditi Sen and Partha Lahiri

Release date: June 29, 2026

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Survey Methodology, June 2026
Vol. 52, No. 1, pp. 1-26
Statistics Canada, Catalogue No. 12-001-X

Improving measurement error and representativeness in nonprobability surveys

Aditi Sen and Partha Lahiri[1]

Abstract

In the age of big data, nonprobability surveys are becoming increasingly abundant. Data integration techniques involving both probability and nonprobability surveys are being extensively used for providing improved estimates for finite population estimation. While much of the existing research has focused on mitigating selection bias in nonprobability surveys, the issue of measurement error within these surveys remains relatively unexplored. Statistical methods devised with the purpose of reducing selection bias are appropriate for reliable estimation, only under the assumption of accuracy of survey responses. Motivated by a recent case study of Kennedy, Mercer and Lau (2024), our research addresses bias from both measurement and sampling errors in nonprobability surveys. In this article, we propose a new data integration method that uses multiple probability and nonprobability surveys and leverages machine learning models to construct a composite estimator. The proposed composite estimator integrates probability and nonprobability surveys, when both contain response variables of interest. We analyze the performance of this estimator in comparison to an existing composite estimator in literature, analytically as well as empirically, using multiple survey data from Kennedy et al. (2024). Finally, we identify conditions under which the proposed estimator outperforms estimators based solely on probability surveys.

Key Words: Bias correction; Composite estimator; Data integration; Measurement error; Opt-in survey; Predictive modeling.

1. Introduction

The use of design-based inference from probability samples has been pivotal for producing reliable estimates for finite population characteristics (such as means, totals, etc.) ever since the groundbreaking research of Neyman (1934). It is attractive in the sense that the same estimation methodology can be applied to any response variable of interest without the need of explicit modeling when the following are available: a sampling frame (list of all units in the targeted finite population), a sampling design with known and nonzero selection probabilities and the survey responses from the selected sample. Despite the presence of unequal selection probabilities, design-based inference remains valid in the case of probability samples; for example, the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) yields unbiased estimates by incorporating the known inclusion probabilities. In contrast, nonprobability samples present a fundamentally different challenge: the selection mechanism is unknown, and some population units may have zero probability of selection. This absence of design information renders traditional design-based methods inapplicable and motivates the development of alternative inferential strategies.

Even though it is desirable to produce estimates based on probability samples, it may not always be practically possible due to budgetary and other constraints. It can be very expensive to conduct properly devised probability surveys, on accounts of recruiting interviewers, pursuing selected respondents for a continued period with declining response rates and other reasons. Such reasons, alongside the influx of big

1. Aditi Sen is PhD Candidate, Applied Mathematics & Statistics, and Scientific Computation. E-mail: asen123@umd.edu; Partha Lahiri is Professor, Department of Mathematics and The Joint Program in Survey Methodology, University of Maryland, College Park, MD 20742, USA. E-mail: plahiri@umd.edu.

data in the age of advanced computing, contribute to the newly gained popularity of nonprobability surveys; although they have existed in the sample survey literature for a very long time. When quick responses can be obtained through web surveys or any individual can be sampled just maintaining some overall quotas (as in quota sampling), there is increased inclination towards using these more convenient options. Reflecting this growing interest, recent special issues of journals – Survey Methodology and the Calcutta Statistical Association Bulletin, have been dedicated to statistical research in nonprobability sampling.

Care must be taken while producing estimates for the targeted finite population characteristics based on such nonprobability surveys because of their obvious selection bias issue. Researchers have devised various statistical methodologies for improving representativeness of nonprobability surveys by integrating them with probability surveys (termed as "reference surveys"), which usually do not have the response variables of interest but have useful auxiliary variables. For an overview, we refer to Rao (2021), Beaumont (2020), among others. Specifically when common auxiliary variables are available for both types of surveys, some commonly used methods include matching and mass imputation methods (Rivers, 2007; Kim, Park, Chen and Wu, 2021), inverse propensity score weighting (IPW) methods (Lee and Valliant, 2009; Valliant and Dever, 2011; Wang, Valliant and Li, 2021), doubly robust methods (Chen, Li and Wu, 2020) and others. Wu (2022) provides an excellent review of these model-based approaches and related assumptions. More recently, Beaumont, Bosa, Brennan, Charlebois and Chu (2024) have investigated variable selection techniques in conjunction with propensity score weighting and post stratification of estimated weights. In contrast to these methods that require combining both probability and nonprobability surveys, Pfeffermann (2015) and Kim and Morikawa (2023) rely solely on nonprobability surveys for inference.

It is well-established in the literature that applying data integration methods makes nonprobability surveys more representative of the finite population, but such methods are developed under the assumption that the survey responses are accurate. Kennedy et al. (2024) point out that a significant amount of measurement error is present in online commercial nonprobability surveys (referred to as opt-in, in short) and bias owing to reasons other than sampling error should not be overlooked. This research is motivated by the 2021 Benchmarking Study (Mercer and Lau, 2023), designed by the Pew Research Center (referred to as Pew, in short) to determine accuracy of online surveys on general population estimates for all adults in the United States (U.S.) as well as key demographic subgroups such as age, race, gender, education, etc. The benchmarks are obtained from large government surveys such as the American Community Survey (ACS), the Current Population Survey (CPS), among others. Kennedy et al. (2024) observe that even after calibration, responses from certain subgroups in opt-in surveys remained highly biased – unlike in probability surveys, where the corresponding estimates are closely aligned with benchmark values. The reason behind this is attributed to "bogus responding", e.g. replying "Yes" regardless of the question. These responses are basically insincere/fraudulent survey answers, causing high measurement error in opt-in surveys. For details on the bogus responding literature, we refer the reader to Kennedy, Hatley, Lau, Mercer, Keeter, Ferno and Asare-Marfo (2021).

In Kennedy et al. (2024), the authors use benchmarking to identify subgroups where bogus responding is concentrated in opt-in surveys, but they do not propose a statistical solution on how to fix this problem.

A quick fix by removing the bogus respondents altogether may make matters worse. The authors highlight the need for statisticians working with opt-in samples to address both issues of bogus responding and representativeness – not just the latter. Our primary objective is to integrate two important areas of research in nonprobability sampling: improving representativeness and addressing measurement differences. In our framework, both the probability and nonprobability surveys contain the response variables of interest, as well as a set of common and informative auxiliary variables. The availability of response variables in both surveys is essential for correcting bias arising from measurement differences. In Section 1.1, we describe our main research questions and our proposed solutions to answer these questions. We provide a detailed discussion about assumptions in Section 1.2. For brevity hereon, we refer to probability surveys as *ps* and nonprobability surveys (specifically opt-in surveys) as *nps*.

1.1 Research questions

Qs.1 Does improving representativeness alone improve estimates from nps?

To answer Qs.1, we apply the IPW method of Chen et al. (2020), assuming that IPW adequately adjusts the selection bias from *nps*. We provide a review of the IPW method in Section 2. For more details on the assumptions (such as missing at random (MAR)), we refer the reader to Wu (2022). It is worth pointing out that Kennedy et al. (2024) only use benchmarking techniques on the case study and model-based methods such as IPW are not used to improve representativeness in their *nps*. It is also important to note that competing weighting methods, as discussed in earlier paragraphs in the introduction, may perform differently. We adopt IPW, which is one of the widely used methods, and do not delve further into comparisons with other methods, as it does not fall within the main focus of this article.

Qs.2 How to correct for measurement differences in nps?

For Qs.2, we devise a methodology to correct for the measurement differences in *nps* due to bogus responding. The benchmarking study in Kennedy et al. (2024) reveals that when compared to large surveys (CPS, ACS, etc.), *ps* are not subject to high measurement errors, but *nps* are. Our aim in this article, is to bring the error levels in *nps* at par with *ps*. Hence, following the footsteps of Kennedy et al. (2024), we consider *ps* to be the benchmark, since it is assumed that measurement error is smallest in this data set. In this way the proposed method corrects for relative measurement bias between *ps* and *nps* (not absolute bias). While we recognize that this assumption may not hold perfectly in practice, our focus lies in enhancing the quality of *nps* by leveraging *ps*, rather than refining *ps* themselves. Consequently, our proposed method to address Qs.2 is centered on correcting "measurement differences" between the two types of surveys.

Qs.3 How to combine estimates from ps and nps for producing estimates with lower bias?

We empirically observe that even after correcting for measurement differences, the finite population mean estimates from *nps* can be further improved through carefully designed composite estimators. Such

estimators have been studied in small area estimation (Schaible, 1978; Rao and Molina, 2015), where the authors consider a weighted combination of a direct and a synthetic estimator. Elliott and Haviland (2007) also propose a weighted estimator based on an nps (web-based convenience sample) with a ps to achieve a smaller Mean Squared Error (MSE) than estimators based on ps alone. More recently, Kim and Tam (2021) propose a bias-corrected data integration estimator, leveraging big data (which can also be thought of as nps) and one ps, where one of them is subject to measurement error.

To address Qs.3, we propose a new composite estimator, which is a weighted combination of unbiased estimates from both ps and nps, aiming to reduce both selection bias and measurement differences in nps. In this paper, we show analytically that the MSE of the proposed estimator is lower than that of Elliott and Haviland (2007). Additionally, we incorporate predictive modeling using advanced machine learning (ML) models (Random Forest, Gradient Boosting) to adjust for bias in IPW-based estimates. ML models are becoming increasing popular in the sample survey context. Recent survey sampling literature, Toth and Eltinge (2011), Breidt and Opsomer (2017), Kern, Klausch and Kreuter (2019), Dagdoug, Goga and Haziza (2021), among others, use modern prediction techniques such as Regression Trees, Random Forest, etc., for model assisted survey estimation.

The estimator proposed in Kim and Tam (2021) accounts for both selection bias and measurement errors in big data without assuming MAR in their missing data approach. However, the method requires identification of overlapping units and uses calibration weighting, treating the big data as an incomplete sampling frame for the finite population. In contrast, our method is simpler (not needing unit-level comparison) but relies on the MAR assumption for IPW-based bias correction. Structurally, our estimator is more aligned with that of Elliott and Haviland (2007), but differs in the sense that it minimizes MSE over a different estimator class. Moreover, Elliott and Haviland (2007) attribute the bias mainly from the point of view of sampling error (i.e., selection bias) without considering measurement error that is also closely aligned with our framework.

Qs.4 When do composite estimators outperform estimators from ps alone?

From our data analysis, we observe that the proposed composite estimator does not uniformly outperform estimators based solely on the ps, when the ps has a sufficiently large sample size. Hence, to answer Qs.4, we draw smaller samples from an available ps and combine with an nps, to demonstrate a scenario where the proposed composite estimator outperforms the estimator created only using ps.

1.2 Conditions and assumptions

Having outlined our research questions and proposed solutions, we now proceed to a detailed discussion on the key assumptions underlying our methodology. The proposed approach primarily relies on two central assumptions, (A1) and (A2):

(A1) IPW primarily removes selection bias from nonprobability samples.

This assumption forms the basis of our approach to Qs.1, which focuses on improving the representativeness of *nps*. IPW method has been extensively used in literature for this purpose. However, this method assumes a MAR selection mechanism (Rubin, 1976). If MAR assumption doesn't hold, then estimating propensity scores is very challenging (Little and Rubin, 2019). Moreover, IPW requires strong auxiliary information to correct for selection bias. In the absence of good covariates, residual selection bias may persist in *nps* even after applying IPW. Specifically, the differences between *ps* and *nps* after correcting for selection bias will be a combination of residual selection bias in *nps*, measurement errors in *nps* and *ps*, and sampling error. However, we do not further explore such a situation for this analysis, as our main purpose is to find new techniques to integrate the two areas of bias (selection and measurement) in *nps*.

(A2) Probability samples have negligible measurement error.

This assumption is used in correcting for measurement differences in *nps*. The presence of measurement errors in *ps* is very likely, as any survey however carefully conducted, is subject to measurement error. Extensive literature has addressed measurement errors in *ps*, starting from the work of Mahalanobis (1946), in the context of agricultural surveys and crop estimation in India. Biemer (2009) in his book chapter refers to the classic work by Hansen, Hurwitz and Bershad (1961) for estimation of measurement error variance. Groves and Lyberg (2010) discuss measurement errors in *ps* within the "Total survey error" framework.

Several articles discuss that mode effects causing measurement errors in probability surveys can be reduced by good questionnaire design or adjusted by weighting/matching methods, after the survey has been conducted. Specifically, Jäckle, Roberts and Lynn (2010) discuss methods for assessing mode effects using proportional odds models, structural equation modeling, etc. Suzer-Gurtekin, Heeringa and Vaillant (2012) use multiple imputation estimation to analytically isolate mode choice and measurement effects. Schouten, Van den Brakel, Buelens, Van der Laan and Klausch (2013) propose a new experimental design that allows for a decomposition of relative mode effects. For more details see Vannieuwenhuyze, Loosveldt and Molenberghs (2010), Buelens and Van den Brakel (2015), Klausch, Schouten, Buelens and Van den Brakel (2017) and the references cited therein. In essence, evaluating and correcting measurement errors in *ps* is a vast research area, which is not focus of this article. In this article, we are mainly concerned with improving *nps* by leveraging available *ps*. Hence, we do not pursue above-mentioned methods further and leave this topic for future research.

The rest of the paper is organized as follows. In Section 2, we provide a review of the IPW estimator from Chen et al. (2020), used for improving representativeness in *nps*. In Section 3, we discuss our proposed method of treating selection bias and measurement differences together through composite estimators, devised by integrating *ps* and *nps*, when both surveys have the response variable/s of interest. In Section 4, we describe the motivating case study from Pew and provide necessary details on the surveys, including data collection method, sample size, benchmarks, weighting, etc. In Section 5, we provide the results of our data analysis, sequentially providing answers to the four research questions raised in the earlier paragraphs. Finally, we end with concluding remarks in Section 6.

2. Improving representativeness

The main objective of this section is to address the representativeness issue of *nps* and improve said selection bias. To this end, we use the IPW method described in Chen et al. (2020). IPW is computationally advantageous as it produces one set of estimated weights that can be used for any response variable. This is similar to the design-based approach of the Horvitz-Thompson estimator, the difference here being that the selection probabilities are estimated instead of being known from the sample design. To use this method, information on common auxiliary variables (from both *ps* and *nps*) and survey weights (from *ps*) are required. For completeness, we provide a review of the IPW method described in Chen et al. (2020) in the following paragraph.

Let S_A be an *nps* of size n_A and S_B be a *ps* of size n_B from a finite population of size N. d_i^B is the known design weight of the ith unit from *ps*, $i = 1, \ldots, n_B$. The selection indicator for S_A, given by R_i, is defined as follows:

$$R_i = \begin{cases} 1 & \text{if } i \in S_A, \\ 0 & \text{if } i \notin S_A, \end{cases} \qquad i = 1, \ldots, N.$$

The propensity score for the ith unit is defined as $\pi_i^A = P_q(R_i = 1 | x_i)$, $i = 1, \ldots, N$, where q is the model for the selection mechanism for S_A and x is the vector of covariates. Following assumptions from Chen et al. (2020), we have $\pi_i^A > 0$, R_i and response y_i are independent given x_i, R_i and R_j are independent given x_i and x_j, for all $i \neq j$. The propensity score takes the following form for the logistic model:

$$\pi_i^A = \pi(x_i, \theta_0) = \frac{\exp(x_i'\theta_0)}{1 + \exp(x_i'\theta_0)}, \quad i = 1, \ldots, N,$$

where θ_0 is the true value of the unknown model parameters. Maximum Likelihood Estimate (MLE) of the propensity score is $\hat{\pi}_i^A = \pi(x_i, \hat{\theta})$, where $\hat{\theta}$ maximizes the following pseudo-log-likelihood function:

$$l^*(\theta) \equiv \sum_{i \in S_A} x_i'\theta - \sum_{i \in S_B} d_i^B \log\{1 + \exp(x_i'\theta)\}.$$

Solution of the above equation is obtained using the Newton-Raphson iterative procedure. Finally, for a response variable y, the IPW estimator for the population mean $\mu_Y = N^{-1} \sum_{i=1}^{N} y_i$ is given by

$$\hat{\mu}_{CLW} \equiv (\hat{N}^A)^{-1} \sum_{i \in S_A} (y_i / \hat{\pi}_i^A), \tag{2.1}$$

where $\hat{N}^A = \sum_{i \in S_A} (\hat{\pi}_i^A)^{-1}$. For detailed expression of the variance of $\hat{\mu}_{CLW}$, we refer to Chen et al. (2020).

For *ps*, we use the survey weighted estimator for response variable y, denoted by

$$\hat{\mu}_{cal} \equiv (\hat{N}^B)^{-1} \sum_{i \in S_B} d_i^B y_i, \tag{2.2}$$

where $\hat{N}^B = \sum_{i \in S_B} d_i^B$. The uncertainty measure of this estimator, can be derived from the design information. The estimator defined in (2.2) can also be used for constructing a weighted estimator from *nps*, using calibrated weights instead of design weights. Hence, to avoid confusion, we use the abbreviation "cal" in the estimator defined in (2.2), as we will be using the same estimator for *ps* (with design weights) and *nps* (with calibrated weights by Pew) in our data analysis.

3. Composite estimators

In this section, we introduce a method to combine estimates from *ps* and *nps*, given that the response variable of interest is present in both the surveys. To formulate this composite estimator, we adopt the model described in Elliott and Haviland (2007), where bias in *nps* is assumed to be known. Mathematically, let $\bar{y}_1$ be a survey weighted estimator of population mean from *ps* ($\hat{\mu}_{cal}$ in our case), $\bar{y}_2$ an estimator from *nps* ($\hat{\mu}_{CLW}$ in our case). Let μ be the true mean of the *ps* estimator, whereas *nps* estimator has an additional bias term ϵ. In this model, ϵ is assumed to be known and μ is the unknown parameter of interest. v_1 and v_2 are known variances of $\hat{\mu}_{cal}$ and $\hat{\mu}_{CLW}$, respectively. The correlation between estimates from *ps* and *nps* is assumed to be 0, as the *ps* and *nps* are independently drawn. The sampling model of Elliott and Haviland (2007) can be written as:

$$\begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \sim \left(\begin{pmatrix} \mu \\ \mu + \epsilon \end{pmatrix}, \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix} \right). \tag{3.1}$$

In practice, a plug-in estimator of ϵ, obtained from historical data or subject matter knowledge, can be used. Elliott and Haviland (2007) propose to estimate bias in practice as the difference between the mean estimates of *ps* and *nps*. v_1 and v_2 can be obtained from the estimates of variance of a design-based estimator (such as our $\hat{\mu}_{cal}$) and the estimate of variance of *nps* (such as $\hat{\mu}_{CLW}$ derived in Chen et al., 2020), respectively.

In Elliott and Haviland (2007), the authors propose a composite estimator of the form:

$$\hat{\mu}_{EV} \equiv \frac{(\epsilon^2 + v_2)\bar{y}_1 + v_1 \bar{y}_2}{\epsilon^2 + v_1 + v_2}, \tag{3.2}$$

which is a biased estimator, with remaining bias given by

$$b_{EV} \equiv E(\hat{\mu}_{EV}) - \mu = \frac{\epsilon v_1}{\epsilon^2 + v_1 + v_2}, \tag{3.3}$$

and variance given by

$$\text{Var}(\hat{\mu}_{EV}) = \frac{v_1 \{(\epsilon^2 + v_2)^2 + v_1 v_2\}}{(\epsilon^2 + v_1 + v_2)^2}. \tag{3.4}$$

In the above equations, we use the subscript "EV" after the authors' names, for simplified notation. Therefore, from (3.3) and (3.4), MSE of $\hat{\mu}_{EV}$ is given by

$$\text{MSE}(\hat{\mu}_{EV}) \equiv b_{EV}^2 + \text{Var}(\hat{\mu}_{EV}) = \frac{v_1(\epsilon^2 + v_2)}{\epsilon^2 + v_1 + v_2}. \tag{3.5}$$

$\hat{\mu}_{EV}$, as defined in (3.2), is the most efficient estimator in the sense of minimizing MSE in a class of biased estimators.

We propose an unbiased composite estimator, which is a weighted combination of mean estimates from ps, given by $\hat{\mu}_{cal}$, and "bias-corrected" estimates from nps, given by $\hat{\mu}_{bc;CLW} \equiv \hat{\mu}_{CLW} - \epsilon$, where subscript "bc" stands for bias-corrected. This composite estimator, denoted by $\hat{\mu}_{comb}$ with subscript standing for "combined", is defined as follows

$$\begin{aligned} \hat{\mu}_{comb}(w) &\equiv w\hat{\mu}_{cal} + (1-w)\hat{\mu}_{bc;CLW} \\ &= w\bar{y}_1 + (1-w)(\bar{y}_2 - \epsilon), \end{aligned} \tag{3.6}$$

where the weight $w \in (0,1)$ is determined by minimizing the MSE of $\hat{\mu}_{comb}(w)$ given by

$$\begin{aligned} \text{MSE}(\hat{\mu}_{comb}(w)) &\equiv E[w\bar{y}_1 + (1-w)(\bar{y}_2 - \epsilon) - \mu]^2 \\ &= w^2 v_1 + (1-w)^2 v_2 . \end{aligned} \tag{3.7}$$

Taking derivative of (3.7) with respect to w and equating to 0 we get $w_1^* = v_2/(v_1 + v_2)$. Finally, we obtain the expression

$$\hat{\mu}_{comb} \equiv \left(\frac{v_2}{v_1 + v_2}\right)\bar{y}_1 + \left(\frac{v_1}{v_1 + v_2}\right)(\bar{y}_2 - \epsilon). \tag{3.8}$$

The proposed estimator in (3.8) is unbiased for μ, as it combines two unbiased estimators: $\hat{\mu}_{cal}$ and $\hat{\mu}_{bc;CLW}$. MSE of $\hat{\mu}_{comb}$ is given by

$$\text{MSE}(\hat{\mu}_{comb}) = \frac{v_1 v_2}{v_1 + v_2}. \tag{3.9}$$

Thus, $\hat{\mu}_{EV}$ and $\hat{\mu}_{comb}$ both minimize MSE, but in disjoint classes of estimators. To compare $\hat{\mu}_{EV}$ with $\hat{\mu}_{comb}$, we compare the respective MSE's given in (3.5) and (3.9). Note that since $\epsilon^2 > 0$, we have the following

$$v_2(\epsilon^2 + v_1 + v_2) = (\epsilon^2 + v_2)v_2 + v_1 v_2 < (\epsilon^2 + v_2)v_2 + v_1(\epsilon^2 + v_2) = (\epsilon^2 + v_2)(v_1 + v_2).$$

From above, since

$$\frac{v_1 v_2}{v_1 + v_2} < \frac{v_1(\epsilon^2 + v_2)}{\epsilon^2 + v_1 + v_2}, \tag{3.10}$$

we have $\text{MSE}(\hat{\mu}_{comb}) < \text{MSE}(\hat{\mu}_{EV})$.

We note that in the previous discussion, the bias term ϵ needs to be known possibly from alternative sources. Assume that $\hat{\epsilon}$, an estimator of ϵ, is constructed from an auxiliary survey that is independent of $\bar{y}_2$. Moreover, let $E(\hat{\epsilon}) = \epsilon$, $\text{Var}(\hat{\epsilon}) = v_b$. Under these assumptions, the combined plug-in estimator can be written as

$$\tilde{\mu}_{comb}(w) \equiv w\bar{y}_1 + (1-w)(\bar{y}_2 - \hat{\epsilon}), \tag{3.11}$$

where the weight $w \in (0,1)$. Using the unbiasedness of $\hat{\epsilon}$ it follows that $E(\tilde{\mu}_{comb}(w)) = \mu$, i.e., $\tilde{\mu}_{comb}(w)$ is an unbiased estimator of μ for all values of $w \in (0,1)$. The corresponding MSE of $\tilde{\mu}_{comb}(w)$ is

$$\text{MSE}(\tilde{\mu}_{comb}(w)) = w^2 v_1 + (1-w)^2 (v_2 + v_b). \tag{3.12}$$

Minimizing the MSE in (3.12) with respect to w yields the optimal weight $w_2^* = (v_2 + v_b)/(v_1 + v_2 + v_b)$. Therefore, substituting this optimal weight into (3.11) yields the plug-in estimator $\tilde{\mu}_{comb}$ given by

$$\tilde{\mu}_{comb} \equiv \left(\frac{v_2 + v_b}{v_1 + v_2 + v_b}\right)\bar{y}_1 + \left(\frac{v_1}{v_1 + v_2 + v_b}\right)(\bar{y}_2 - \hat{\epsilon}) \tag{3.13}$$

and the MSE of $\tilde{\mu}_{comb}$ is given by $\text{MSE}(\tilde{\mu}_{comb}) = v_1(v_2 + v_b)/(v_1 + v_2 + v_b)$. It is worth noting that $\hat{\mu}_{comb}$ is a special case of $\tilde{\mu}_{comb}$ with $v_b = 0$, corresponding to the degenerate case where the bias estimator satisfies $\hat{\epsilon} = \epsilon$ with probability 1. When compared with the estimator proposed in Elliott and Haviland (2007), similar algebraic manipulation as in (3.10) shows that $\text{MSE}(\tilde{\mu}_{comb}) \leq \text{MSE}(\hat{\mu}_{EV})$ provided $v_b \leq \epsilon^2$. This suggests that the plug-in estimator performs better than Elliott and Haviland (2007) when the variance of the bias estimator is relatively small compared to the square of the bias. A specific construction of the bias estimator $\hat{\epsilon}$ is presented in the real data application in Section 5.2.

4. Data description

For this analysis, we use the six sample surveys sourced from Pew's 2021 Benchmarking Study. The data and detailed report are available on the Pew website (weblink). The surveys in this study were administered between June 14 and July 21, 2021. They included interviews with a total of 29,937 U.S. adults and approximately 5,000 in each sample. Three of the surveys were sourced from different probability-based online panels, one of which was Pew's American Trends Panel (ATP). The remaining three surveys came from three different online opt-in sample providers. These surveys were conducted through online platforms.

Table 4.1, sourced from Kennedy et al. (2024), provides detailed information on the sample sizes and field dates. A common questionnaire was administered in English or Spanish for all six surveys. Information on demographics such as age, race, gender, education, marital status were obtained along with answers on other questions related to insurance coverage, smoking status, blood pressure level, U.S. citizenship, number of kids/adults in households, etc. These six survey datasets can be categorized into two types – *nps* and *ps* as follows:

i. Nonprobability surveys (nps): Three of the six surveys are commercial opt-in surveys, denoted as Opt-in 1, 2, 3. These were obtained from vendors who used quota sampling. Pew provided vendors with quota targets for age × gender, race × Hispanic ethnicity, education from the 2019 ACS. Ipsos was the data collection firm in this case. Since these are online nonprobability panels, selection probabilities assigned to the individuals are unknown.

ii. Probability surveys (ps): The other three are probability surveys. Although the panels are online, panelists are recruited offline, e.g. using probability sampling. To be specific, they are recruited using address-based sampling (ABS) from the U.S. Postal Service Computerized Delivery Sequence File. Hence, following the nomenclature of Kennedy et al. (2024), we refer to these surveys as ABS 1, 2, 3. The study-specific response rates of ABS 1, 2, 3 were 61%, 90%, and 71%, respectively.

Table 4.1
Sample sizes and field dates of six Pew surveys

Sample type	Sample name	ID	Sample size	Field dates
Probability	ABS panel 1	P_1	5,027	June 14 – 28, 2021
	ABS panel 2	P_2	5,147	June 14 – 27, 2021
	ABS panel 3	P_3	4,965	June 29 – July 21, 2021
Nonprobability	Opt-in panel 1	O_1	4,912	June 15 – 25, 2021
	Opt-in panel 2	O_2	4,931	June 11 – 27, 2021
	Opt-in panel 3	O_3	4,955	June 11 – 26, 2021

Note: Address-based sampling (ABS).
Source: Kennedy, Mercer and Lau (2024).

Weighting: Kennedy et al. (2024) used a standard weighting approach of calibration for all six surveys. For the *nps* since there was no design weight, a starting weight of 1 was assigned to every individual. For the *ps*, panel base weights were adjusted for differential probabilities of selection and then used in calibration. Multiple population targets (from ACS, CPS and Pew's National Public Opinion Reference Survey) such as age × sex, race/ethnicity × education, etc., were used to calibrate the starting weights for each sample to a common set of population control totals.

Benchmark: Some of the survey questions were considered for benchmarking analysis, where the benchmarks were obtained from sources such as National Health and Nutrition Examination Survey (NHANES), National Health Interview Survey (NHIS), ACS, CPS, Pew Survey on COVID data tracker and Election projection data. These values are provided at national level as well as by categories of demographic variables such as age (3 categories viz. 18-29, 30-64, 65+ years), race (3 categories viz. White non-Hispanic, Black non-Hispanic and Hispanic) and education (3 categories viz. Higher Secondary (HS) or less, Some college and College graduate). For the analysis of measurement error, these benchmark values are compared with survey estimates (calculated from the six surveys) at these subgroup levels of demographic variables or overall survey level, for response variables of interest. We elaborate this in the following Evaluation Section.

4.1 Evaluation

To evaluate the estimators in terms of both representativeness bias and measurement differences, we use the Mean Squared Deviation/Difference (MSD) as the evaluation criterion – lower MSD implying better performance of an estimator. We define this error in terms of deviation from benchmark values as follows. Let m be the number of benchmark variables for a survey (ps or nps) and k be the number of subgroups based on factors of interest (viz. demographic variables such as age, race, etc). If $\bar{Y}_{jc}$ is the value for question j and subgroup c from benchmark source (considered as true value) and $\hat{\bar{Y}}_{jc}$ is the respective survey estimate (where $j = 1,\ldots,m;\ c = 1,\ldots,k$), then MSD for the estimator $\hat{\bar{Y}}$ is defined as

$$\text{MSD}(\hat{\bar{Y}}) \equiv m^{-1}\sum_{j=1}^{m}\left[k^{-1}\sum_{c=1}^{k}\left(\hat{\bar{Y}}_{jc} - \bar{Y}_{jc}\right)^2\right]. \tag{4.1}$$

As an example, we mention that to evaluate the impact of selection bias improvement we compute $\hat{\mu}_{CLW}$ from nps and calculate $\text{MSD}(\hat{\mu}_{CLW})$. We compare $\text{MSD}(\hat{\mu}_{CLW})$ with $\text{MSD}(\hat{\mu}_{cal})$ from ps. Evaluations can be done at an overall survey level like in (4.1) or at a subgroup level of any factor of interest or at question level or both, removing the averaging factor accordingly. In this data analysis, we only consider binary questions and do not consider the level of a question as an additional component in MSD.

5. Data analysis

Before proceeding into our empirical investigation we take into account the key findings in Kennedy et al. (2024). The authors use the measure Mean Absolute Error (MAE) and observe that MAE values are higher in opt-in surveys as compared to ABS. At the subgroup level, MAE values show a distinct pattern in opt-in surveys that is absent in ABS surveys. For example, among 18-29 year-old's and Hispanic adults, MAE values are higher in opt-in surveys, likely due to "bogus-responding". Bogus responding is more prominent in questions regarding government benefits such as social security, food stamps, unemployment compensation or workers' compensation. Large shares of 18-29 year-old's and Hispanic adults report having received at least three of four such benefits, which is very rare in the true population. According to Kennedy et al. (2024), for opt-in surveys, this ranges between 6% to 11%, while the true population incidence is 0.1%. These insightful findings motivate us to use advanced methods for weighting in opt-in samples in conjunction with bias-correction due to measurement differences. In the following sections, we describe our data analysis methods. All of our codes are available at this website (weblink).

5.1 IPW estimators from opt-in samples

We use the method discussed in Section 2 leveraging `R` package `nonprobsvy` to calculate $\hat{\mu}_{CLW}$ for 3 opt-in samples O_1, O_2, O_3, with P_1, P_2, P_3 respectively as reference. We emphasize that there is no inherent pairing between the surveys O_j and P_j; i.e, any combination of O_j and P_k, for $j, k = 1, 2, 3$, can be used to improve representativeness. In our analysis, we use the matched indices O_j and P_j for $j = 1, 2, 3$, but we expect similar results with other combinations. This assignment also ensures independence across the

pairs, which is critical for estimating $\hat{\epsilon}$ in the composite estimator (3.13) using external data sources independently. In the propensity score logistic model we consider age, gender, race, education and region as auxiliary variables. For opt-in samples, we compare two estimators in terms of MSD – one is survey weighted mean using calibrated weights from Pew ($\hat{\mu}_{cal}$) and the other is IPW estimator ($\hat{\mu}_{CLW}$). For ABS samples, only the first one is applicable ($\hat{\mu}_{cal}$). A list of the estimators that we compare in this data analysis, with detailed formula, is provided in Table 5.1. For these estimators, MSD values are calculated using (4.1) with $m = 12$ benchmark questions having only "Yes"/"No" answer. Details of these questions are provided in Table 5.2.

Table 5.1
List of finite population mean estimators, with equation number for detailed formula and types of surveys

No.	Estimator	Details of the estimation method	Formula	Survey type
1	$\hat{\mu}_{cal}$	Weighted mean using design or calibrated weights	(2.2)	*ps* or *nps*
2	$\hat{\mu}_{CLW}$	IPW estimator of Chen, Li and Wu (2020)	(2.1)	*nps*
3	$\hat{\mu}_{bc;CLW}$	Measurement difference bias-corrected version of $\hat{\mu}_{CLW}$, where bias is calculated from (5.1)	(5.2)	*nps*
4	$\hat{\mu}_{EV}$	Composite estimator of Elliott and Haviland (2007), where bias is calculated from (5.1)	(3.2)	*ps* & *nps*
5	$\tilde{\mu}_{comb}$	Proposed composite estimator combining $\hat{\mu}_{cal}$ and $\hat{\mu}_{bc;CLW}$, where bias is calculated from (5.1)	(3.13)	*ps* & *nps*
6	$\hat{\mu}_{ML;CLW}$	IPW estimator of Chen et al. (2020) using predicted responses	(5.4)	*ps* & *nps*
7	$\hat{\mu}_{ML;comb}$	Proposed composite estimator combining $\hat{\mu}_{cal}$ and $\hat{\mu}_{ML;CLW}$	(5.5)	*ps* & *nps*
8	$\hat{\mu}_{ML;EV}$	Version of $\hat{\mu}_{EV}$, where bias is difference of $\hat{\mu}_{cal}$ and $\hat{\mu}_{CLW}$	(5.6)	*ps* & *nps*

Notes: - In column 5, "or" implies that an estimator is defined for either *ps* or *nps* and "&" implies that an estimator is defined using both *ps* and *nps*.
- Chen, Li and Wu (CLW); Elliott and Haviland (EV); inverse propensity score weighting (IPW); machine learning (ML).

Table 5.2
Details of 12 questions (with only "Yes"/"No" answers) used as benchmark

Question No.		Question Wording
1.	Insurance	Are you currently covered by any form of health insurance or health plan?
2.	Blood pressure	Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?
3.	Parent	Are you the parent or guardian of any children under age 18?
4.	Food allergy	Do you have any food allergies?
5.	Job last year	Did you work at a job or business at any time during 2020?
6.	Retirement account	At any time during 2020 did you have any retirement accounts such as a 401(k), 403(b), IRA (Individual retirement arrangement), or other account designed specifically for retirement savings?
7.	Unemployment compensation	At any time during 2020, did you receive any State or Federal unemployment compensation?
8.	Workers' compensation	During 2020 did you receive any Worker's Compensation payments or other payments as a result of a job-related injury or illness?
9.	Food stamps	At any time during 2020, did you or anyone in your household receive benefits from SNAP (the Supplemental Nutritional Assistance Program) or the Food Stamp program, or use a SNAP or food stamp benefit card?
10.	Social Security	During 2020 did you receive any Social Security payments from the U.S. Government?
11.	Union membership	Are you a member of a labor union or of an employee association similar to a union?
12.	U.S. citizenship	Are you a citizen of the United States?

Source: Sourced from Questionnaire on the Pew Research Centre website https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/(weblink).

In Table 5.3, we compare $\hat{\mu}_{cal}$ and $\hat{\mu}_{CLW}$, in terms of MSD, provided in 10^2 scale, for ease of representation. We see that for opt-in surveys, MSD ($\hat{\mu}_{CLW}$) is greater than MSD ($\hat{\mu}_{cal}$). For example, for O_1, MSD ($\hat{\mu}_{CLW}$) is 1.168, whereas MSD ($\hat{\mu}_{cal}$) is 1.039. Hence, no improvement is observed at an overall survey-level in opt-in surveys, using IPW as compared to calibrated weights. As a result, in comparison to ABS surveys, MSD values for opt-in surveys are still quite high. For example, for O_1, MSD ($\hat{\mu}_{CLW}$) is 1.168 whereas for P_1, MSD ($\hat{\mu}_{cal}$) is 0.080. Next, in Table 5.4, we compare MSD values of the same estimators at subgroup levels of age and race. Herein we observe that excepting a few cases, MSD ($\hat{\mu}_{CLW}$) is greater than MSD ($\hat{\mu}_{cal}$). The exceptions include 6 cases out of 18 as follows – age group 18-29 years for O_1, age group 65+ years for O_2 and O_3, race White for O_1, race Black for O_2 and O_3. Finally, a similar conclusion as in Table 5.3 can be drawn that IPW does not improve upon the estimator using calibrated weights. Again, comparing opt-in surveys to ABS, we observe that bias pertaining to specific subgroups is present in opt-in samples – as reported by Pew. For example, for O_1 in age group 18-29 years, MSD ($\hat{\mu}_{CLW}$) is 2.312, as compared to the value 0.158 of MSD ($\hat{\mu}_{cal}$) for the same category in P_1.

Table 5.3
MSD values (scaled to 10^2) of two population mean estimators, $\hat{\mu}_{cal}$ and $\hat{\mu}_{CLW}$, calculated using 12 benchmark questions from 6 surveys

Estimator	P_1	P_2	P_3	O_1	O_2	O_3
MSD($\hat{\mu}_{cal}$)	0.080	0.210	0.097	1.039	1.024	0.480
MSD($\hat{\mu}_{CLW}$)	-	-	-	1.168	1.237	0.564

Notes: - Refer to Table 5.1 for definition/formula of estimators and MSD definition in (4.1).
- Chen, Li and Wu (CLW); Mean Squared Deviation/Difference (MSD).

Table 5.4
Subgroup level MSD values (scaled to 10^2) of $\hat{\mu}_{cal}$ and $\hat{\mu}_{CLW}$ at 3 age subgroups (18-29, 30-64, 65+ years) and 3 race subgroups (White, Black and Hispanic)

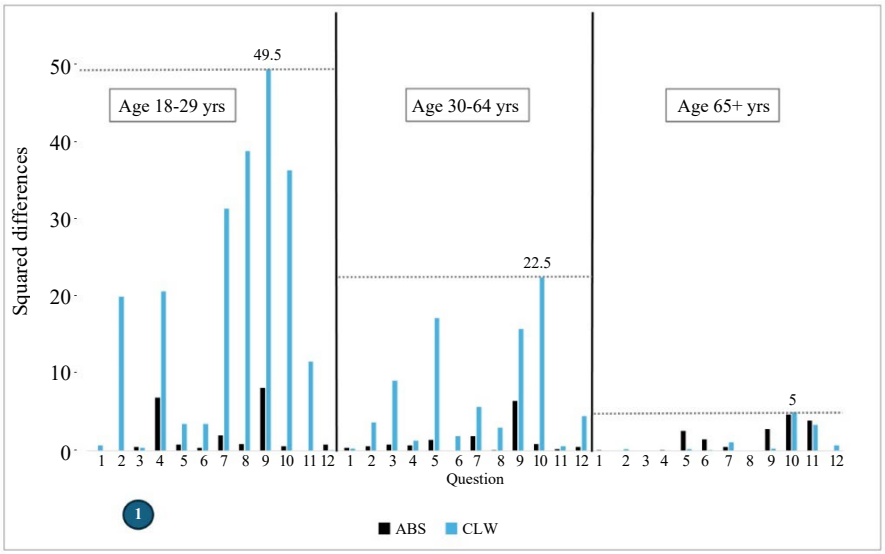
	Factor	Group	P_1	P_2	P_3	O_1	O_2	O_3
MSD($\hat{\mu}_{cal}$)	age	18-29 (yrs)	0.158	0.507	0.177	2.869	3.162	1.790
		30-64 (yrs)	0.104	0.239	0.117	1.345	1.407	0.567
		65+ (yrs)	0.130	0.185	0.135	0.189	0.162	0.102
MSD($\hat{\mu}_{CLW}$)	age	18-29 (yrs)	-	-	-	2.312	3.901	1.930
		30-64 (yrs)	-	-	-	1.497	1.705	0.728
		65+ (yrs)	-	-	-	0.318	0.146	0.097
MSD($\hat{\mu}_{cal}$)	race	White	0.064	0.152	0.090	1.034	0.907	0.390
		Black	0.281	0.502	0.376	1.099	1.121	0.499
		Hispanic	0.102	0.375	0.236	2.625	2.341	2.147
MSD($\hat{\mu}_{CLW}$)	race	White	-	-	-	0.913	1.136	0.471
		Black	-	-	-	1.351	0.877	0.464
		Hispanic	-	-	-	3.441	2.781	2.444

Notes: - Estimates for White and Black adults are based on those who do not identify as Hispanic and values for Asian/other race categories (constituting less than 5% respondents in each survey) are not reported.
- Chen, Li and Wu (CLW); Mean Squared Deviation/Difference (MSD).

To understand the impact of IPW method on selection bias improvement at a more granular level, we calculate the squared difference of estimators ($\hat{\mu}_{CLW}$ and $\hat{\mu}_{cal}$) from benchmark values at 3 subgroup levels of age for 12 questions. In this section, we discuss the squared difference results for surveys O_3 and P_3 only, as in the next section we will use the other surveys (O_1, O_2 and P_1, P_2) to construct an estimator of bias. Hereon, we use superscripts on estimators to denote the source sample, such as P_1, P_2, P_3 or O_1,

O_2, O_3. In Figure 5.1, light blue bars denote squared differences for $\hat{\mu}_{CLW}^{O_3}$ and black bars denote the same for $\hat{\mu}_{cal}^{P_3}$. We mark the highest value of blue bars for each age group in dotted line. The squared difference values in this figure and the rest of the figures in Section 5 are scaled to 10^3, for ease of visualization. Similar conclusions as before can be drawn here – among the three age groups, group 18-29 years displays the highest difference between heights of the two types of bars. We observe from the figure that the highest value of squared difference for $\hat{\mu}_{CLW}^{O_3}$ is 49.473 in the age group 18-29 years. This value comes from Question 9 regarding Food Stamps. Bias is often significant in questions that are sensitive, subjective, or challenging to answer due to ambiguous definitions of concepts or reliance on the respondent's memory. Based on these criteria, estimates for Question 9 are more likely to have high squared differences from benchmark values. Overall, we see that the blue bars are quite high, compared to the black bars. Thus, we conclude that improving representativeness alone does not improve estimates from opt-in surveys.

Figure 5.1 Plot of squared differences of two population mean estimates with benchmark $\bar{Y}$, i.e., $(\bar{Y} - \hat{\mu}_{cal}^{P_3})^2$ in black and $(\bar{Y} - \hat{\mu}_{CLW}^{O_3})^2$ in blue, for 12 questions and 3 age groups 18-29, 30-64, 65+ years



Notes: - Highest values of blue bars in each age group are denoted in dotted lines. Y -axis is scaled to 10^3.
- Address-based sampling (ABS); Chen, Li and Wu (CLW).

5.2 Measurement difference bias-corrected IPW estimators

In this section, we address bias due to measurement differences. The findings in Kennedy et al. (2024) suggest that all the opt-in samples suffer from a similar bias. The authors confirm that this measurement error bias is due to the effect of age and race, which we have also statistically tested using ANOVA models. Hence, following the empirical investigation of Kennedy et al. (2024), we assume that the distribution of bias is similar between the opt-in samples. With this prior knowledge of factors causing bias and the assumption of similar bias between opt-in samples, we compute estimate of bias using O_1, O_2 and P_1, P_2, keeping O_3 for bias correction and P_3 for validation. Recall that from the model defined in (3.1), $\hat{\mu}_{CLW} - \hat{\mu}_{cal}$ yields an unbiased estimator of ϵ. Thus, we define an estimate of the bias at question × age group level as the mean of pair-wise differences between nps and ps. For question j and age group level c, this is given by

$$\begin{aligned}\hat{\epsilon}^{jc} &\equiv 4^{-1}\{(\hat{\mu}_{CLW}^{jc;O_1}-\hat{\mu}_{cal}^{jc;P_1})+(\hat{\mu}_{CLW}^{jc;O_1}-\hat{\mu}_{cal}^{jc;P_2})+(\hat{\mu}_{CLW}^{jc;O_2}-\hat{\mu}_{cal}^{jc;P_1})+(\hat{\mu}_{CLW}^{jc;O_2}-\hat{\mu}_{cal}^{jc;P_2})\}\\ &= 2^{-1}\{(\hat{\mu}_{CLW}^{jc;O_1}-\hat{\mu}_{cal}^{jc;P_1})+(\hat{\mu}_{CLW}^{jc;O_2}-\hat{\mu}_{cal}^{jc;P_2})\},\end{aligned} \tag{5.1}$$

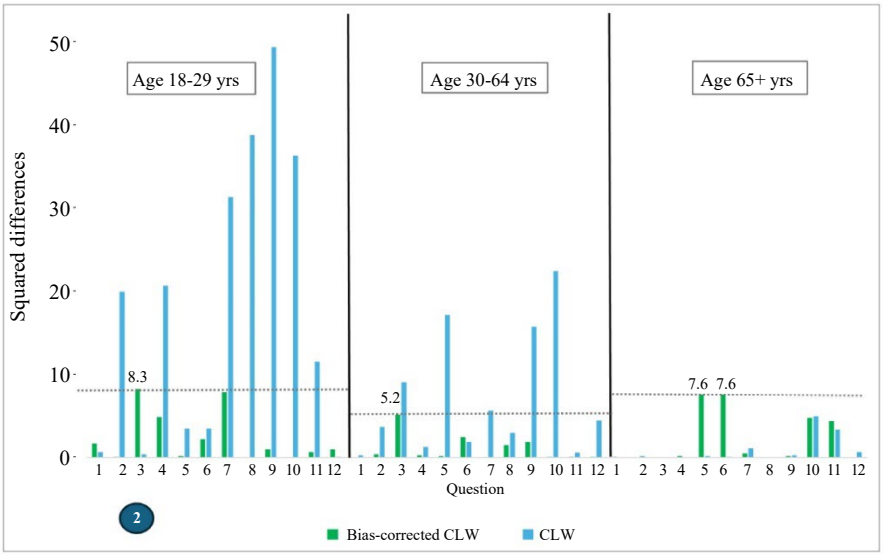
where $j=1,\ldots,12;\quad c=1,\ 2,\ 3$. Due to the additive bias structure in the model defined in (3.1), we calculate the bias-corrected estimates at question j and age group level c in O_3 as

$$\hat{\mu}_{bc;CLW}^{jc;O_3} \equiv \hat{\mu}_{CLW}^{jc;O_3}-\hat{\epsilon}^{jc},\quad j=1,\ldots,12;\quad c=1,\ 2,\ 3. \tag{5.2}$$

In this section, we describe the bias correction method focusing on the factor age, but similar correction can be done using the factor race. We do not consider both factors together (age × race) for the bias-correction, due to unavailability of benchmarks at such granular levels.

We compare $\hat{\mu}_{bc;CLW}^{jc;O_3}$ with $\hat{\mu}_{CLW}^{jc;O_3}$ in terms of squared differences with benchmarks to see the effect of measurement bias correction due to age. In Figure 5.2, maintaining color consistency with the previous figure, we plot the squared differences of $\hat{\mu}_{CLW}^{jc;O_3}$ in blue bars and that of bias-corrected estimates $\hat{\mu}_{bc;CLW}^{jc;O_3}$ in green bars. The highest value of green bars in every age group is marked in dotted line. We see that in most cases, i.e. question × age group cell, the green bars are lower in height than blue bars. The highest value of green bar in age group 18-29 years is 8.315. This is much lower as compared to the highest value of blue bar in the same group (49.472 as noted in Figure 5.1). Thus, from the empirical investigation we conclude that improving measurement differences after improving representativeness produces better estimates.

Figure 5.2 Plot of squared differences of $\hat{\mu}_{CLW}^{O_3}$ and its bias-corrected version $\hat{\mu}_{bc;CLW}^{O_3}$ with benchmark, i.e., $(\bar{Y}-\hat{\mu}_{CLW}^{O_3})^2$ in blue and $(\bar{Y}-\hat{\mu}_{bc;CLW}^{O_3})^2$ in green, for 12 questions and 3 age groups 18-29, 30-64, 65+ years



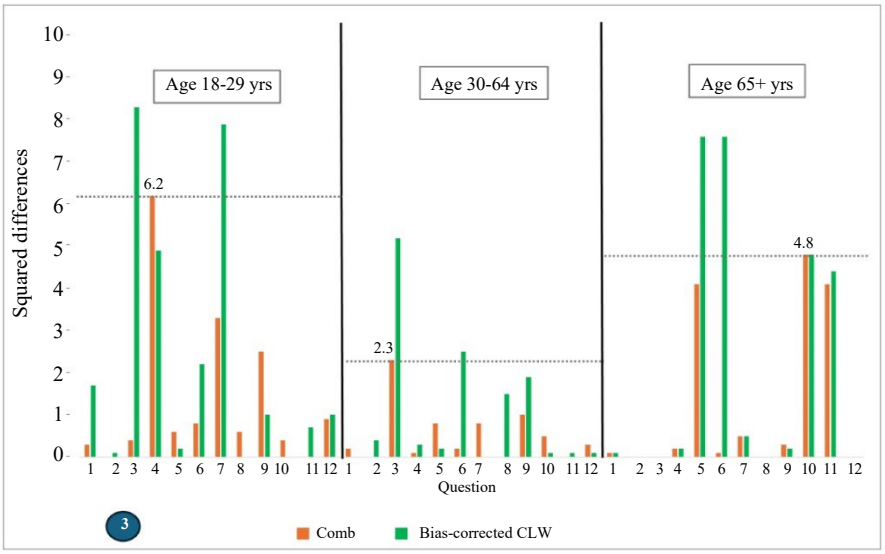
Notes: - Highest values of green bars in each age group are denoted in dotted lines. Y -axis is scaled to 10^3.
- Chen, Li and Wu (CLW).

5.3 Composite estimator from ABS and opt-in samples

We calculate the composite estimators $\hat{\mu}_{EV}$ and $\tilde{\mu}_{comb}$ described in Section 3. For a fair comparison, for both the estimators we take the estimate of bias $\hat{\epsilon}$ as defined in (5.1) in Section 5.2, and the values of v_1 , v_2 as the estimated variances of $\hat{\mu}_{CLW}$ and $\hat{\mu}_{cal}$ (from R packages `nonprobsvy` and `survey` respectively). We show two figures in this section – Figure 5.3 and 5.4. In the first one we compare proposed composite estimator ($\tilde{\mu}_{comb}$) with the bias-corrected estimator from the previous section ($\hat{\mu}^{O_3}_{bc:\,CLW}$). In the second one we compare our proposed composite estimator ($\tilde{\mu}_{comb}$) with the existing composite estimator $\hat{\mu}_{EV}$. In Figure 5.3, we plot the squared differences of $\tilde{\mu}_{comb}$ from benchmark in orange bars and that of $\hat{\mu}^{O_3}_{bc:\,CLW}$ in green bars (as in Figure 5.2). In this plot we see that on an average the orange bars are lower in height than the green bars, implying that the proposed composite estimator is even better than bias-corrected IPW estimator. The highest value of orange bar in age group 18-29 years is 6.182. This is lower in comparison to the same for green bar, which is 8.315 (as shown in Figure 5.2).

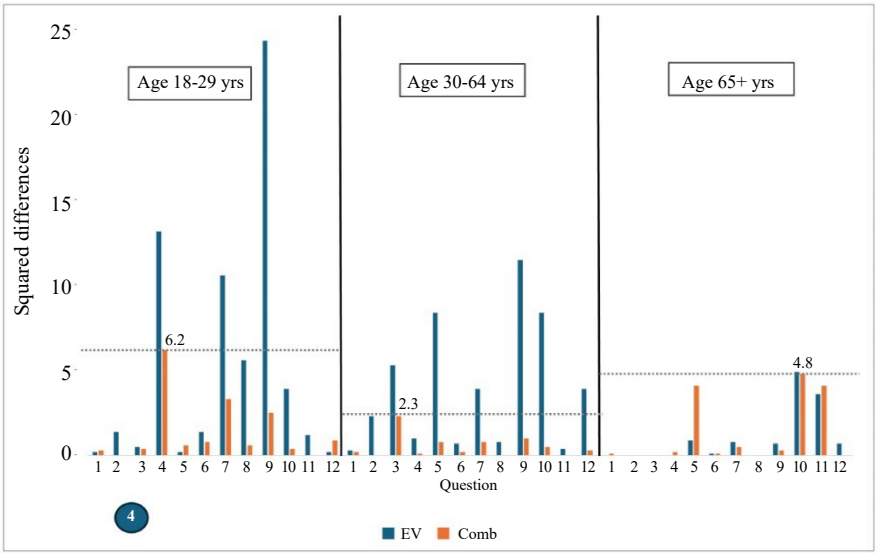
In Figure 5.4, we similarly compare $\tilde{\mu}_{comb}$ and $\hat{\mu}_{EV}$ in terms of their squared differences with benchmarks. Again, on an average, the values for $\tilde{\mu}_{comb}$ (denoted in orange bars) are lower than that of $\hat{\mu}_{EV}$ (denoted in dark blue bars). We observe from the dotted lines that the highest squared difference with benchmark is also lower for $\tilde{\mu}_{comb}$ than $\hat{\mu}_{EV}$ in age groups 18-29 and 30-64 years. However, in the 65+ age group, the performance of $\tilde{\mu}_{comb}$ and $\hat{\mu}_{EV}$ is similar. Hence, we conclude that our proposed composite estimator performs better than the earlier one in the literature.

Figure 5.3 Plot of squared differences with benchmark of $\hat{\mu}^{O_3}_{bc:CLW}$ (using O_3) and $\tilde{\mu}_{comb}$ (using O_3 and P_3), i.e., $(\bar{Y} - \hat{\mu}^{O_3}_{bc:CLW})^2$ in green and $(\bar{Y} - \tilde{\mu}_{comb})^2$ in orange, for 12 questions and 3 age groups 18-29, 30-64, 65+ years



Notes: - Highest values of orange bars in each age group are denoted in dotted lines. Y -axis is scaled to 10^3.
- Chen, Li and Wu (CLW).

Figure 5.4 Plot of squared differences with benchmark of composite estimators $\hat{\mu}_{EV}$ and $\tilde{\mu}_{comb}$, i.e., $(\bar{Y} - \hat{\mu}_{EV})^2$ in dark blue and $(\bar{Y} - \tilde{\mu}_{comb})^2$ in orange, for 12 questions and 3 age groups 18-29, 30-64, 65+ years



Notes: - Highest values of orange bars in each age group are denoted in dotted lines. Y -axis is scaled to 10^3.
- Elliott and Haviland (EV).

5.4 Comparison of composite estimator with ABS

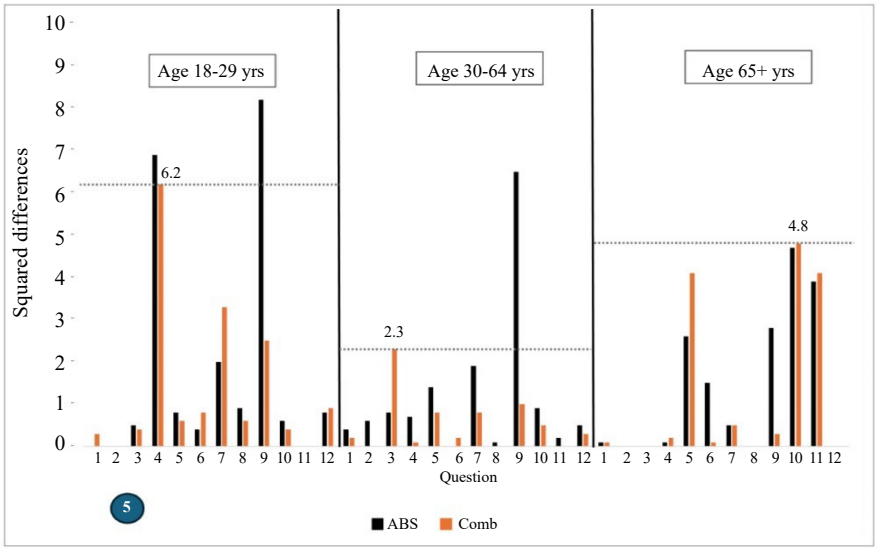
After the discussion about composite estimators in Section 5.3, one might be tempted to ask the question whether the proposed composite estimator (from both *ps* and *nps*) outperforms the estimator from *ps* alone. In this section, we want to investigate if the performance of composite estimator can be further improved. In Figure 5.5, we provide a comparison of squared differences of $\tilde{\mu}_{comb}$ (denoted in orange bars) and $\hat{\mu}_{cal}^{P_3}$ (denoted in black bars), maintaining the color consistency from previous figures. For the age group 18-29 years, we see that the estimator from *ps* alone is better than our proposed composite estimator for some questions (viz. 1, 6, 7, 12). We also observe that for Question 10 (concerning social security) in the 65+ age group, all estimators discussed in Section 5 exhibit similar values for squared differences from the benchmark (Figures 5.1-5.5). This can be attributed to the fact that older or retired people are more likely to receive such benefits and are less susceptible to bogus responding. As a result, estimates from *nps* are already close to *ps*.

In our case study, *ps* already has a large sample size (close to five thousand). As a result *ps* itself is good enough (in most cases) for estimation of finite population characteristics. Thus in order to explain situations where performance of the proposed composite estimator is significantly better in terms of MSD, we generate *ps* of smaller sample sizes. To this end, we draw samples from P_3 using a stratified sampling procedure with proportional allocation, where the strata are created by segmentation based on survey weights in P_3, viz., (0, 0.5), [0.5, 1) and so on. We then calculate MSD values of $\hat{\mu}_{EV}$, $\tilde{\mu}_{comb}$ and $\hat{\mu}_{cal}$ for the smaller samples from P_3, keeping the original *nps* O_3 intact. We note here that the estimators $\hat{\mu}_{CLW}$ and $\hat{\mu}_{bc;\,CLW}$ (computed using only O_3) remain the same, as there is no change in sample size of O_3. Their respective MSD values are 0.564 and 0.183. In Table 5.5, we provide MSD values (scaled to 10^2, comparable to Table 5.3), along with relative change in MSD (in % inside parenthesis) of estimators with respect to $\hat{\mu}_{cal}$, defined as

$$\frac{\text{MSD}\,(\hat{\mu}_{cal}) - \text{MSD}\,(\hat{\mu})}{\text{MSD}\,(\hat{\mu})} \times 100, \tag{5.3}$$

where $\hat{\mu} \in \{\hat{\mu}_{EV}, \tilde{\mu}_{comb}\}$.

Figure 5.5 Plot of squared differences of $\hat{\mu}_{cal}^{P_3}$ (using P_3) and $\tilde{\mu}_{comb}$ (using P_3 and O_3) with benchmark, i.e., $(\bar{Y} - \hat{\mu}_{cal}^{P_3})^2$ in black and $(\bar{Y} - \tilde{\mu}_{comb})^2$ in orange, for 12 questions and 3 age groups 18-29, 30-64, 65+ years



Notes: - Highest values of orange bars in each age group are denoted in dotted lines. Y-axis is scaled to 10^3.
- Address-based sampling (ABS).

Table 5.5
MSD values (in 10^2 scale) of estimator $\hat{\mu}_{cal}$ (from P_3 and its samples), $\tilde{\mu}_{comb}$ and $\hat{\mu}_{EV}$ (from P_3 and its samples, combined with O_3), starting from the original sample size (4,912) decreasing up to 100

ps sample size	$\hat{\mu}_{cal}$	$\hat{\mu}_{EV}$	$\tilde{\mu}_{comb}$
4,912	0.097	0.337 (-71.284)	0.101 (-3.767)
1,000	0.231	0.295 (-21.718)	0.158 (45.896)
500	0.399	0.385 (3.642)	0.289 (38.293)
100	1.073	0.748 (43.476)	0.626 (71.399)

Notes: - Numbers presented in $(\cdot)$ correspond to the relative change in Mean Squared Deviation/Difference (MSD) (in %) of the estimators compared to $\hat{\mu}_{cal}$, defined in (5.3).
- Elliott and Haviland (EV).

We observe that as sample size of *ps* decreases, the proposed composite estimator $\tilde{\mu}_{comb}$ has lower MSD values in comparison to $\hat{\mu}_{cal}$. For example, for *ps* sample size 100, MSD $(\tilde{\mu}_{comb})$ is 0.626, which is much lower in comparison to the value of 1.073 of MSD $(\hat{\mu}_{cal})$. This is also reflected in the relative change in MSD value between these two estimators (71.399%). However, for the original sample P_3, we see that MSD $(\tilde{\mu}_{comb})$ > MSD $(\hat{\mu}_{cal})$ and relative change in MSD is -3.767%. Similar phenomenon is observed for

$\hat{\mu}_{EV}$, where MSD ($\hat{\mu}_{EV}$) < MSD ($\hat{\mu}_{cal}$) for ps sample sizes 500 and 100. It is interesting to note that the MSD of $\hat{\mu}_{EV}$ is always more than that of $\tilde{\mu}_{comb}$. Finally, we conclude that the proposed composite estimator is not better than the estimator from ps alone, when ps has an adequate sample size. But as the sample size of ps decreases the variance of $\hat{\mu}_{cal}$ increases, as a result the composite estimator $\tilde{\mu}_{comb}$ outperforms $\hat{\mu}_{cal}$.

5.5 Predictive model for bias correction

In Section 3, we mention that in order to compute composite estimators the bias ϵ needs to be known from alternate sources or historical data, which is why in Section 5.2 we estimate bias from O_1, O_2, P_1, P_2 and apply the bias correction in O_3. It is important to note that the estimate of bias proposed by Elliott and Haviland (2007) cannot be applied to our case, as then our composite estimator ends up being the estimator from ps itself, i.e., in (3.13), using $\hat{\epsilon} = \bar{y}_2 - \bar{y}_1$, we get $\tilde{\mu}_{comb} = \bar{y}_1$. When other sources of information about bias are not available to obtain $\hat{\epsilon}$, we propose an alternative solution in this section. To this end, we fit a model on P_3 (relating responses to auxiliary variables), and using the fitted parameter estimates from this model, we predict the unit level responses in O_3 (or equivalently, the probabilities of responding "Yes"). This data integration approach can be related to that proposed in Sen and Lahiri (2025) in the context of finite population proportion estimation. The difference being that in this case the prediction is done on the nonprobability survey, whereas in Sen and Lahiri (2025), the responses are predicted for a bigger probability survey.

In this article, for model building we use the `R` package `caret` (Kuhn, Wing, Weston, Williams, Keefer, Engelhardt, Cooper, Mayer, Kenkel, R Core Team, Benesty, Lescarbeau, Ziem, Scrucca, Tang, Candan and Hunt, 2020). With the help of `caret`, we fit machine learning (ML) models – Random Forest (RF) and Gradient Boosting Model (GBM), using cross validation (CV) as the re-sampling method and tuning hyper-parameters such as number of boosting iterations, maximum tree depth, shrinkage, etc. In Table 5.6 we provide necessary details on modeling building. We observe from Table 5.6 that both RF and GBM perform quite similarly in terms of accuracy (0.85) on the test data, but GBM has higher Area Under the Curve (AUC) value (0.92). In order to choose the final model, in addition to AUC and accuracy, we also leverage the MSD criteria considered in earlier sections. Finally, as GBM has lower MSD and higher AUC than RF, we choose this as our predictive model. The final values of tuning parameters used for the model are `n.trees` = 1,000, `interaction.depth` = 3, `shrinkage` = 0.05 and `n.minobsinnode` = 10 and `logLoss` was used to select the optimal model using the smallest value. Question levels are among the topmost important variables responsible for classification. In Figure 5.6, we show the `logLoss` with respect to different tuning parameters and also the Receiver Operating Characteristic (ROC) curves of the two binary classifier models (RF and GBM).

We then use the same IPW method (described in Section 2) on the predicted probabilities from the chosen model (GBM) to produce a new population mean estimator of nps. This is given by:

$$\hat{\mu}_{ML;\,CLW} \equiv (\hat{N}^A)^{-1} \sum_{i \in S_A} (\hat{p}_i^y / \hat{\pi}_i^A), \tag{5.4}$$

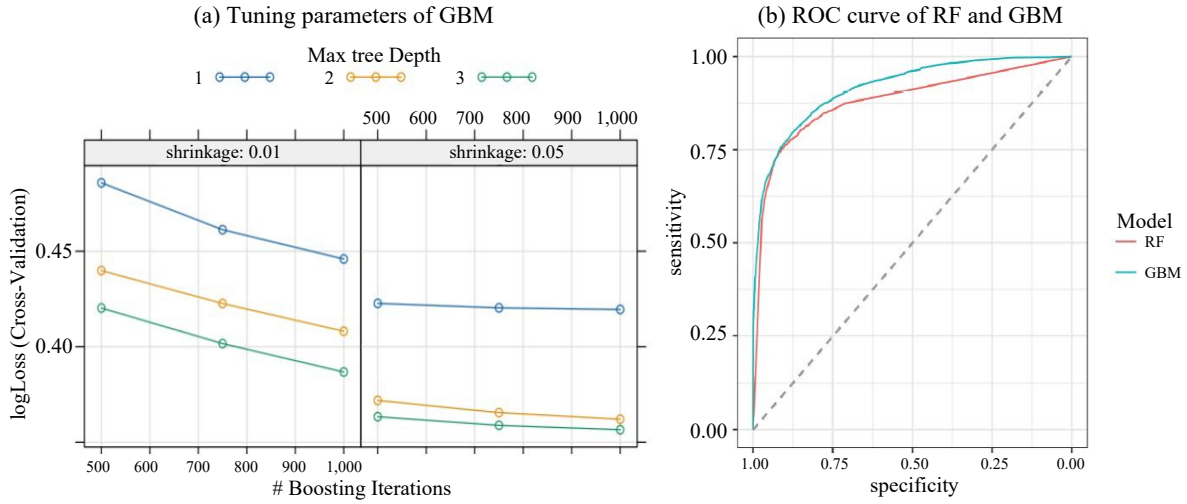
where $\hat{p}_i^y$ is the estimated likelihood that the individual i responds "Yes" to variable y given the auxiliary characteristics of i, $\hat{\pi}_i^A$ is obtained from the IPW method, and $\hat{N}^A = \sum_{i \in S_A} (\hat{\pi}_i^A)^{-1}$. The difference of $\hat{\mu}_{ML;\,CLW}$ defined in (5.4) with $\hat{\mu}_{CLW}$ defined in (2.1) is that in (5.4) we use estimated likelihood values ($\hat{p}^y$) instead of true responses (y).

Table 5.6
Details of the predictive modeling on *ps* P_3 of 4,912 respondents, with the aim of obtaining bias-corrected estimates from *nps* O_3

Data structure	Question × respondent level data
Size (Train + test)	59 thousand observations split into train (80%) and test (20%)
Response variable type	Binary (Refusals are not considered)
Auxiliary variables (no. of levels)	Question (12), age (3), race-ethnicity (5), education (3), gender (3), region (4)
Predictive models (AUC, Accuracy)	Random Forest (0.88, 0.85), Gradient Boosting (0.92, 0.85)

Note: Area under the curve (AUC).

Figure 5.6 Results of predictive modeling on probability sample P_3



Note: Gradient Boosting Model (GBM); Random Forest (RF); Receiver Operating Characteristic (ROC).

As a result, the new composite estimator in this case is constructed as in (3.8), but using (5.4) instead of (2.1). The proposed estimator $\hat{\mu}_{ML;\,comb}$ is defined as

$$\hat{\mu}_{ML;\,comb} \equiv \left(\frac{v_2}{v_1 + v_2}\right) \hat{\mu}_{cal} + \left(\frac{v_1}{v_1 + v_2}\right) \hat{\mu}_{ML;\,CLW}. \tag{5.5}$$

We compare the above estimator with a different version of $\hat{\mu}_{EV}$, which we denote by $\hat{\mu}_{ML;\,EV}$. Although modeling is not used in the construction of this estimator, for ease of comparison we use the same subscript

"ML". In this estimator, the unknown bias is estimated by the difference of weighted means from ps and nps, as suggested by Elliott and Haviland (2007). Hence, $\hat{\mu}_{ML;\,EV}$ is defined as

$$\hat{\mu}_{ML;\,EV} \equiv \frac{\{(\bar{y}_1 - \bar{y}_2)^2 + v_2\}\ \bar{y}_1 + v_1 \bar{y}_2}{(\bar{y}_1 - \bar{y}_2)^2 + v_1 + v_2}. \tag{5.6}$$

In Table 5.7, we show the values of the two composite estimators $\hat{\mu}_{ML;\,comb}$ (proposed) and $\hat{\mu}_{ML;\,EV}$ (existing) and their respective MSD values for 12 questions considered for the analysis. Benchmark numbers are also provided alongside the estimates. We observe that for all questions, $\text{MSD}(\hat{\mu}_{ML;\,comb}) < \text{MSD}(\hat{\mu}_{ML;\,EV})$.

Table 5.7
Composite Estimators $\hat{\mu}_{ML;comb}$ (proposed) and $\hat{\mu}_{ML;EV}$ (existing) and their Mean Squared Deviation/Difference (MSD) values (in 10^3 scale) for 12 binary benchmark questions given in Table 5.2

Question	Benchmark	$\hat{\mu}_{ML;EV}$	MSD ($\hat{\mu}_{ML;EV}$)	$\hat{\mu}_{ML;comb}$	MSD ($\hat{\mu}_{ML;comb}$)
1. Insurance	90.800	88.979	4.450	90.755	4.447
2. Blood Pressure	31.100	37.725	5.271	35.709	5.109
3. Parent	26.000	22.336	10.189	25.977	10.099
4. Food allergy	9.400	14.038	0.984	12.823	0.966
5. Job last year	64.200	58.262	20.105	63.476	19.817
6. Retirement account	49.900	49.866	25.016	51.119	25.015
7. Unemployment compensation	9.300	16.846	0.970	12.189	0.825
8. Workers' compensation	0.400	6.637	0.284	1.144	0.010
9. Food stamps	11.100	21.549	5.328	16.416	5.247
10. Social Security	21.800	29.898	26.574	26.189	26.413
11. Union membership	5.600	10.410	0.657	8.055	0.595
12. U.S. citizenship	92.500	96.299	4.413	96.064	4.332

Note: Elliott and Haviland (EV); machine learning (ML).

Finally, we summarize the two cases of bias-correction –

Case 1: When bias is computed from alternative sources.

Case 2: When bias-corrected estimates are predicted using modeling.

A comparison between the two cases would not be fair, given the fact that improved results are expected if bias is known from alternative sources. In Table 5.8, we provide the MSD values of comparable estimators in these two scenarios. In Case 1: we compare $\tilde{\mu}_{comb}$, $\hat{\mu}_{EV}$, $\hat{\mu}_{CLW}$ and $\hat{\mu}_{bc;\,CLW}$. Whereas in Case 2: we compare $\hat{\mu}_{ML;\,comb}$, $\hat{\mu}_{ML;\,EV}$ and $\hat{\mu}_{ML;\,CLW}$. We observe that in both cases the proposed estimators ($\tilde{\mu}_{comb}$ and $\hat{\mu}_{ML;\,comb}$) have lower MSD values than existing composite estimators ($\hat{\mu}_{EV}$ and $\hat{\mu}_{ML;\,EV}$) as well as other estimators ($\hat{\mu}_{CLW}$, $\hat{\mu}_{bc;\,CLW}$ and $\hat{\mu}_{ML;\,CLW}$). It is also noteworthy that the bias-correction is effective in both situations by significantly lowering MSD values. Referring back to Table 5.3 in Section 5.1, we see that the MSD value of $\hat{\mu}_{cal}^{P_3}$ is 0.097. Hence, the proposed composite estimators ($\tilde{\mu}_{comb}$, $\hat{\mu}_{ML;\,comb}$) are doing a good job to achieve comparable accuracy for an nps (combined with ps) as would be obtained normally for a ps of similar size.

Table 5.8
Comparison of Mean Squared Deviation/Difference (MSD) values (in 10^2 scale, given inside parenthesis) of proposed and competing estimators calculated using 12 benchmark questions from *nps* O_3 and *ps* P_3 (in case of composite estimators) for two cases of bias correction

	Proposed Composite	Existing Composite	Other
Case 1 :	$\tilde{\mu}_{comb}$ (0.101)	$\hat{\mu}_{EV}$ (0.337)	$\hat{\mu}_{CLW}$ (0.564) $\hat{\mu}_{bc;CLW}$ (0.183)
Case 2 :	$\hat{\mu}_{ML;comb}$ (0.092)	$\hat{\mu}_{ML;EV}$ (0.355)	$\hat{\mu}_{ML;CLW}$ (0.099)

Notes: - Refer to Table 5.1 for details on the estimators.
- Chen, Li and Wu (CLW); Elliott and Haviland (EV); machine learning (ML).

6. Discussion

In this article, we address the problem of improving finite population inference through construction of composite estimators, integrating data from probability and nonprobability surveys. Our goal is to mitigate bias arising from issues of representativeness and measurement error in nonprobability surveys. To this end, we propose two bias correction methods – (i) Measurement difference correction, applicable when additional data sources beyond a probability and a nonprobability survey are available; and (ii) Model-based bias correction using advanced machine learning models, appropriate for settings where only a probability and a nonprobability survey are accessible. We illustrate the effectiveness of these approaches through a case study based on the data from Kennedy et al. (2024), showing that the proposed estimators yield improved accuracy relative to large government surveys regarded as gold standards. Method (i) offers a computationally straightforward solution, while method (ii) is more broadly applicable in practice, especially in scenarios with limited data sources.

We also demonstrate that the proposed composite estimator achieves improved performance (specifically, lower MSD) when the sample size of the probability survey is smaller relative to that of the nonprobability survey. However, reducing the probability sample size can give rise to small area estimation challenges. In our case study, we observe that some subgroups (defined by question × age group combinations) have no observations when the probability sample size is substantially reduced. In such situations, small area modeling techniques, as discussed in Nandram and Rao (2024) and Franco and Bell (2013), become particularly valuable. Addressing this issue presents an interesting direction for future research.

The proposed methodology in this paper is tested on a specific case study involving opt-in surveys in the U.S. The performance of the method may vary across different countries, survey modes, or cultural contexts, where the nature of selection bias and measurement error may differ substantially. Additionally, the proposed methodology relies on certain assumptions, which are discussed in this paper along with references that address potential violations. As a direction for future research, it would be valuable to develop composite estimators that account for measurement errors in both probability and nonprobability surveys. Overall, our dual-focus approach of tackling both selection bias and measurement differences offers a

promising framework for improving survey methodology in the context of growing reliance on non-probability surveys.

Acknowledgements

The authors wish to thank Andrew Mercer, Senior Research Methodologist at the Pew Research Center, for his support on this research. The authors are also thankful to the two anonymous referees and associate editor for giving constructive comments on an earlier version of this article.

References

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of offcial statistics? *Survey Methodology*, 46(1), 1-28. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50(1), 77-106. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00004-eng.pdf.

Biemer, P. (2009). Chapter 12 – Measurement errors in sample surveys. *Handbook of Statistics*, 29, 281-315.

Breidt, F.J. and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.

Buelens, B. and van den Brakel, J.A. (2015). Measurement error calibration in mixed-mode sample surveys. *Sociological Methods & Research*, 44(3), 391-426.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Dagdoug, M., Goga, C. and Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542), 1234-1251.

Elliott, M.N. and Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology*, 33(2), 211-215. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10498-eng.pdf.

Franco, C. and Bell, W.R. (2013). Applying bivariate binomial/logit normal models to small area estimation. JSM 2013 - Survey Research Methods Section.

Groves, R.M. and Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5), 849-879.

Hansen, M., Hurwitz, W. and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute, 32nd Session*, 38, 359-374.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.

Jäckle, A., Roberts, C. and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78(1), 3-20.

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J. and Asare-Marfo, D. (2021). Strategies for detecting insincere respondents in online polling. *Public Opinion Quarterly*, 85(4), 1050-1075.

Kennedy, C., Mercer, A. and Lau, A. (2024). Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith. *Survey Methodology*, 50(1), 3-21. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00013-eng.pdf.

Kern, C., Klausch, T. and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13(1), 73-93.

Kim, J.K. and Morikawa, K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. *Calcutta Statistical Association Bulletin*, 75(1), 8-27.

Kim, J.K., Park, S., Chen, Y. and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3), 941-963.

Kim, J.K. and Tam, S.M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2), 382-401.

Klausch, T., Schouten, B., Buelens, B. and van den Brakel, J. (2017). Adjusting measurement bias in sequential mixed-mode surveys using re-interview data. *Journal of Survey Statistics and Methodology*, 5(4), 409-432.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt, T. (2020). Package "caret". *The R Journal*, 223(7), 48.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3), 319-343.

Little, R.J. and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mercer, A. and Lau, A. (2023). Comparing two types of online survey samples. *Pew Research Center*. Available at: https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/.

Nandram, B. and Rao, J.N.K. (2024). Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. *Statistics and Applications*, ISSN 2454-7395 (online), 22(1), 343-374.

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, 123-150.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, 3(4), 425-483.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83(1), 242-272.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation, 2nd Edition*. New York: John Wiley & Sons, Inc.

Rivers, D. (2007). Sampling for web surveys. *Joint Statistical Meetings*, American Statistical Association, Alexandria, VA.

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

Schaible, W.L. (1978). Choosing weights for composite estimators for small area statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 741, 746.

Schouten, B., van den Brakel, J.A., Buelens, B., van der Laan, J. and Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, 42(6), 1555-1570.

Sen, A. and Lahiri, P. (2025). Estimation of finite population proportions for small areas – A statistical data integration approach. *Journal of Survey Statistics and Methodology*, 13(3), 309-332.

Suzer-Gurtekin, Z.T., Heeringa, S.G. and Vaillant, R. (2012). Investigating the bias of alternative statistical inference methods in sequential mixed-mode surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4711-4725.

Toth, D. and Eltinge, J.L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496), 1626-1636.

Valliant, R. and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1), 105-137.

Vannieuwenhuyze, J., Loosveldt, G. and Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74(5), 1027-1045.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.*, 40(24), 5237-5250.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2), 283-311. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.