Analysis of combined probability and nonprobability samples: a simulation evaluation and application to a teen smoking behavior survey

Article *in* Communication in Statistics- Simulation and Computation · July 2022

DOI: 10.1080/03610918.2022.2102181

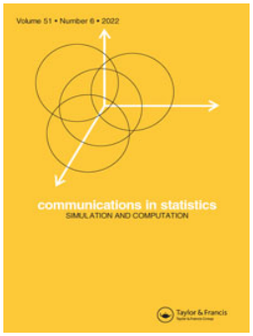
CITATIONS
0

READS
79

7 authors, including:

Wenna Xi
Weill Cornell Medical College
20 PUBLICATIONS **142** CITATIONS
SEE PROFILE

Alice Hinton
The Ohio State University
260 PUBLICATIONS **1,370** CITATIONS
SEE PROFILE



Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Analysis of combined probability and nonprobability samples: a simulation evaluation and application to a teen smoking behavior survey

Wenna Xi, Alice Hinton, Bo Lu, Karol Krotki, Brittney Keller-Hamilton, Amy Ferketich & Amang Sukasih

To cite this article: Wenna Xi, Alice Hinton, Bo Lu, Karol Krotki, Brittney Keller-Hamilton, Amy Ferketich & Amang Sukasih (2022): Analysis of combined probability and nonprobability samples: a simulation evaluation and application to a teen smoking behavior survey, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2022.2102181

To link to this article: https://doi.org/10.1080/03610918.2022.2102181

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2022.2102181

Analysis of combined probability and nonprobability samples: a simulation evaluation and application to a teen smoking behavior survey

Wenna Xi[a], Alice Hinton[b], Bo Lu[b], Karol Krotki[c], Brittney Keller-Hamilton[b], Amy Ferketich[b], and Amang Sukasih[c]

[a]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA; [b]College of Public Health, The Ohio State University, Columbus, OH, USA; [c]RTI International, Rockville, MD, USA

ABSTRACT

In scientific studies with low-prevalence outcomes, probability sampling may be supplemented by nonprobability sampling to boost the sample size of desired subpopulation while remaining representative to the entire study population. To utilize both probability and nonprobability samples appropriately, several methods have been proposed in the literature to generate pseudo-weights, including ad-hoc weights, inclusion probability adjusted weights, and propensity score adjusted weights. We empirically compare various weighting strategies via an extensive simulation study, where probability and nonprobability samples are combined. Weight normalization and raking adjustment are also considered. Our simulation results suggest that the unity weight method (with weight normalization) and the inclusion probability adjusted weight method yield very good overall performance. This work is motivated by the Buckeye Teen Health Study, which examines risk factors for the initiation of smoking among teenage males in Ohio. To address the low response rate in the initial probability sample and low prevalence of smokers in the target population, a small convenience sample was collected as a supplement. Our proposed method yields estimates very close to the ones from the analysis using only the probability sample and enjoys the additional benefit of being able to track more teens with risky behaviors through follow-ups.

ARTICLE HISTORY
Received 22 October 2021
Accepted 8 July 2022

KEYWORDS
Buckeye Teen Health Study; Low-prevalence outcomes; Nonprobability sampling; Probability sampling; Pseudo-weight; Propensity score

1. Introduction

Probability sampling has been the standard practice in survey research for decades. Analysis following the sampling design provides unbiased estimates of the population quantities under the sampling distribution. Well executed probability sampling with good cooperation can guard against selection bias, which is a major threat to the generalizability of sample data. Moreover, probability sampling provides a mathematically sound framework to assess the precision of the estimates, as we know the selection probability for each unit (Lohr 2019). In recent years, however, survey researchers and practitioners have experienced substantially declining response rates in probability surveys. For example, seven nationwide large-scale surveys sponsored by the Department of Health and Human Services showed declining response rates from 1995 to 2015, with some surveys showing a decreasing trend over the first half of the period, while others

CONTACT Wenna Xi wex4002@med.cornell.edu Department of Population Health Sciences, Weill Cornell Medicine, DV-306A, 425 E 61st St New York, NY 10065.

Supplemental data for this article is available online at https://doi.org/10.1080/03610918.2022.2102181.

showing a steady decrease in recent years (Czajka and Beyler 2016). Kohut et al. 2012 reported that the response rate for a typical telephone survey at Pew Research dropped from 36% in 2007 to 9% in 2012. Such low response rate makes the sample more vulnerable to nonresponse bias. Therefore, the validity of using probability samples to make inference about the general population may be in question (Groves 2006; Brick 2011). Meanwhile, the cost and effort required to obtain acceptable levels of cooperation have increased substantially.

Over the past two decades, survey researchers have witnessed the fast growth of nontraditional data collection methods. Aided by advances in technology, a vast amount of data can be gathered in more efficient ways, e.g., via mobile devices, online surveys, and volunteer community samples (Link 2018). These data, however, are collected without a clearly defined sampling framework or a probability-based selection rule. Usually, they are referred to as nonprobability samples (Vehovar, Toepoel, and Steinmetz 2016). Compared with probability samples, nonprobability samples may not have full coverage of the target population due to heterogeneous access to survey platforms (Cornesse et al. 2020) or may have representativeness issue such as in volunteer community samples or general convenient samples. Without a probability sampling foundation, strong modeling or calibration assumptions are usually required to infer population estimates. Though nonprobability samples have performed well in certain areas, such as electoral polling, evidence of accuracy is insufficient in other domains (Dutwin and Buskirk 2017).

When the response rate turns out to be lower than expected or required for key analyses, researchers might need to supplement the probability sample with nonprobability sample, either for the study overall or for only some specific domains of interest (Berzofsky, Williams, and Biemer 2009). This strategy, if implemented appropriately, may help to improve the representativeness and unbiasedness of a probability sample in practice. Following the works of Schonlau et al. 2004, Lee and Valliant 2009, Valliant and Dever 2011, and Elliott and Valliant 2017, our approach is to combine probability and nonprobability samples by calculating the pseudo-weights for the nonprobability sample, and then using combined sample (adjusted probability weights for the probability sample and pseudo-weights for the non-probability sample) for design-based estimation.

From a methodological perspective, there are generally two types of approaches to making inference from nonprobability samples—quasi-randomization and superpopulation models (Elliott and Valliant 2017). The quasi-randomization approach tries to mimic a probability sampling framework by calculating the inclusion probabilities of the nonprobability sample. Pseudo-weights are generated based on inclusion probabilities and are calibrated to a reference probability sample (Schonlau, Van Soest, and Kapteyn 2007). The superpopulation model approach focuses on modeling the relationship between an outcome and relevant covariates, then projects the sample to the target population based on the model. The key difference between the two approaches is that the former uses design-based inference and the latter uses model-based inference conditioning on the collected sample. Recent methodological development also suggests a way of combining the two, namely doubly robust approach (Valliant 2020). Model-assisted weights are calculated based on quasi-randomization weights to construct approximately consistent estimators if the pseudo-inclusion probability distribution or the superpopulation model is correctly specified. Robbins, Ghosh-Dastidar, and Ramchand 2021 further discussed the blending of probability and nonprobability samples using propensity scores by considering two strategies—disjoint blending and simultaneous blending.

In this paper, we focus on producing pseudo-weights by implementing the quasi-randomization method, as it is conceptually closely related to the conventional probability sampling. It only needs one set of pseudo-weight for all outcomes. In contrast, in superpopulation modeling, different models may need to be generated for different outcomes, which leads to varied efficiencies across survey outcomes The quasi-randomization approach is also more robust to the misspecifications of the outcome model.

Our research is motivated by the Buckeye Teen Health Study (BTHS), a population survey for examining tobacco use behaviors among adolescent males in Ohio (Evans et al. 2018; Friedman et al. 2018). The original design was to use a two-stage stratified probability sampling design (stratified by urban/rural status and county) through address-based sampling (ABS). After several months of fielding, the response rate was lower than expected and it was unlikely we could reach the target sample size. The primary outcome was teens' tobacco use behaviors, including the use of cigarettes and/or smokeless tobacco products. Because the prevalence of use was expected to be low, we needed a large enough sample size to track the use of different tobacco products. Therefore we decided to include a nonprobability convenience sample to supplement the probability sample so that we could identify more boys with risky behaviors and track their long-term tobacco use status. Since the probability sample accounts for a large portion of the data, we tried to use quasi-randomization method to take advantage of the probability sampling design.

The major thrust of our paper is to compare several strategies for combining probability and nonprobability samples through a large simulation study, using a sampling design motivated by our BTHS data. Several weighting strategies are considered, from the naive weighting for the nonprobability sample to the more theory-based pseudo-weight modeling method. We also modify the propensity score approach to obtain weights that balance the distribution between the probability and nonprobability samples. To improve the estimation, weight normalization (Hahs-Vaughn 2005) is usually needed so that the sum of the normalized combined weights is equal to the sum of the original probability sample weights, and we propose a strategy that is consistent with the original survey design. Many of these methods are used in practice without a thorough comparison of their statistical performance. We attempt to fill the gap by providing insights regarding their practical utility through simulation studies. The rest of the paper is organized as follows: Sec. 2 describes different estimation strategies being compared; Sec. 3 introduces our motivating example, the BTHS; Sec. 4 presents the simulation study design and results; Sec. 5 shows the data analysis results; and Sec. 6 concludes with a discussion on practical implications and limitations.

2. Methodology

Design-based estimation, which includes point and variance estimation, under probability sampling has been established for many "classical" sampling designs (Cochran 1977), as well as for more complex designs (Lohr 2019). Within the framework of total survey errors, the variance of the estimate has been accepted as a quality measure for the sampling error. Under the design-based estimation approach, weighted estimation is commonly used to produce design-unbiased estimate of the population parameter. The sampling weight computed as the inverse of selection probability is one of the most important components for the design-based estimation. In practice, this weight is adjusted to account for survey nonresponse and then may be subjected to calibration to produce the final analysis weight. Unfortunately, with the nonprobability sample, the sampling weight cannot be calculated without strong assumptions, which results in some researchers ignore weighting when analyzing the nonprobability sample.

Under the quasi-randomization approach, pseudo inclusion probabilities are estimated for the nonprobability sample. In our application, the pseudo sampling weights (hereafter the pseudo-weights) are generated for the nonprobability sample, which is then pooled with the probability sample to create a combined sample. Design-based formulas are used for inference.

We first take a stratified probability simple random sample, and then a convenience sample is collected from the remaining population. For a stratified random sampling with proportional allocation procedure, the sampling weights for the probability sample are equal within each stratum. When adding the nonprobability sample, there are generally three steps to prepare weights for the final inference: (1) pseudo-weights estimation for the nonprobability sample, (2) weight

Table 1. Summary of approaches implemented in weighting comparisons.

Method and notation	Description
Pseudo-weight estimation:	
PW(1)	Treats nonprobability sample units as if they were selected the same way as probability sample units, assuming the nonprobability sample is independent of the probability sample.
PW(2)	Sets the pseudo-weight to 1 (nonprobability sample unit only represents itself).
PW(3)	Estimates the pseudo-weight using a model which relies on a probability sample with commonly observed covariates in probability and non-probability samples.
PW(4)	Estimates the pseudo-weight using a model based on the combined probability and non-probability samples, and balances the propensity scores by re-weighting the nonprobability sample to make it look like the probability sample.
Weight normalization:	
WN(1)	Treats the nonprobability sample as a separate stratum and normalizing the pseudo-weights by their sample sizes to preserve the proportion of each sample in the combined sample.
WN(2)	Treats both nonprobability and probability sample units as coming from the sampling strata defined by the design variables, and normalizes the pseudo-weights to preserve the total weight of each stratum as the population size in the combined sample.
Raking:	
RK(1)	No raking.
RK(2)	Rakes to the known marginal population distribution.
RK(3)	Rakes to the marginal distribution in the probability sample.

normalization, and (3) raking. Our research considered four different pseudo-weight estimations, two weight normalizations, and three raking approaches, which are summarized in Table 1 and described in more detail in the following subsections.

2.1. Pseudo-weights estimation

We consider four methods to estimate pseudo-weights for the nonprobability sample: naïve weight (Naïve), unity weight (Unity), inclusion probability adjusted weight (IPA), and propensity score adjusted weight (PSA).

Pseudo-weight 1 or PW(1):

The *naïve weight* method treats nonprobability sample units as if they were selected the same way as probability sample units. The nonprobability sample units can be merged into their corresponding strata based on their covariates, and the sampling weight can be calculated directly. For both probability and nonprobability sample units, stratified simple random sampling weight can be calculated as the population size over the total sample size (probability plus nonprobability samples).

Pseudo-weight 2 or PW(2):

The *unity weight* method assumes that each nonprobability sample unit only represents itself, so its pseudo-weight is set to 1. Nonprobability sample units are merged into their corresponding strata based on their covariates. For the probability sample, weights are unchanged and sum to the population size. As a result, in the combined sample, the sum of weights exceeds the population size, and thus requires normalization (methods described in the next section). When the size of the probability sample is large enough and the probability sample is assumed to produce

design-unbiased estimates, a small size of the nonprobability sample may not have significant impact on the accuracy or the precision of the estimates.

Pseudo-weight 3 or PW(3):

The *inclusion probability adjusted weight* method is based on Elliott 2009. It assumes that the nonprobability sample also has some unknown selection probability, which relies on a probability sample with commonly observed covariates. The nonprobability sample is assumed to be independent of the probability sample, and the inclusion into the study may be assumed to be independent of outcome given sample design, though in practice the probability selection may correlate with the exposure of interest.

We let $Z_i = 0$ if unit i is in the probability sample and $Z_i = 1$ if unit i is in the nonprobability sample. We let S_i be the indicator that unit i is sampled for the probability sample, and $\mathbf{x}_i$ be the vector of commonly observed covariates that are predictive of participation. Elliott and Valliant 2017 showed that pseudo-weight w_i for the nonprobability sample unit i is given by:

$$w_i \propto \tilde{w}_i \frac{\hat{P}(Z_i = 0 \mid \mathbf{x}_i = \mathbf{x}_0)}{\hat{P}(Z_i = 1 \mid \mathbf{x}_i = \mathbf{x}_0)}, \tag{1}$$

where $\tilde{w}_i = 1/\hat{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_0)$ is the weight associated with selecting a sample unit i in the probability sampling frame; that is, the inverse of the probability of selecting a sample unit with covariates $\mathbf{x}_i = \mathbf{x}_0$ in the probability sampling frame. If the probability sample is drawn via stratified simple random sampling with proportional allocation, then $\tilde{w}_i$ is the same for all i, hence can be dropped in (1). In situations where $\mathbf{x}_i$ do not correspond exactly with the probability sampling design variables, beta regression can be used to predict the selection probability for the probability sample $\hat{P}(S = 1|\mathbf{x}_i = \mathbf{x}_0)$ (Ferrari and Cribari-Neto 2004). The probability $P(Z_i = z|\mathbf{x}_i = \mathbf{x}_0)$ can be estimated via logistic regression in the combined sample. For the probability sample, weights are calculated following the sampling design.

Pseudo-weight 4 or PW(4):

The *propensity score adjusted weight* method has been used to calculate pseudo-weights for nonprobability web surveys (Schonlau et al. 2004; Lee and Valliant 2009). Valliant and Dever 2011 used the propensity score stratification to calibrate weights for an online convenience sample based on a reference probability sample. We borrow the idea of propensity grouping but modify the calculation of propensity score weights by taking advantage of the balancing property of propensity scores. That is, we re-weight the combined sample to make it look like the probability sample in terms of covariate distributions (Austin 2011). Since the target population distribution is based on the probability sample, which is supposed to be representative, we will model the propensity score as the probability of being in the probability sample. The specific steps are:

Step 1: Combine the probability and nonprobability samples. Use weighted logistic regression to estimate the propensity score, $e_i = P(Z_i = 0 \mid X_i)$, of belonging to the probability sample. The weight used in the model was defined as the initial sampling weights for probability sample units and unity weights for nonprobability sample units. Include all relevant covariates in the propensity score model.

Step 2: Based on the propensity score, classify the data into a fixed number of groups. The number of groups depends on the total sample size and the distribution of propensity scores. The number cannot be too small; otherwise, the propensity score adjustment is not effective. It also cannot be too large, because we need a reasonable amount of probability

sample units in each group. In our simulation, we used 20 equal-sized groups. For group j, the average propensity score is calculated and denoted as e_j.

Step 3: Within each group, calculate the propensity score weight for each unit. To ensure the covariate distribution is balanced, for units from the probability sample, propensity score weights are set to 1 ($v_i = 1$); for units from the nonprobability sample, propensity score weights are calculated as for unit i in group j:

$$v_i = \frac{e_j}{1 - e_j}, \tag{2}$$

Step 4: Within each group, calculate pseudo-weights for nonprobability sample units (while adjusting the initial sampling weights for probability sample units). To ensure the combined sample represents the target population as the probability sample, we re-distribute the sum of probability sample weights to all units. The pseudo-weight for unit i in group s is:

$$w_i = \frac{\sum_{j \in s} I\left(Z_j = 0\right) \tilde{w}_j}{\sum_{j \in s} v_j} \times v_i \tag{3}$$

where $I(\cdot)$ denotes the indicator function with value equal to 1 if $Z_j = 0$ and 0 otherwise, v_j is calculated in Step 3, and $\tilde{w}_j$ is the probability sample weight if the jth unit is from the probability sample and 0 for the nonprobability sample. Note that weights for the probability sample are also adjusted after this step.

2.2. Weight normalization

Since nonprobability and probability samples are combined for final inference, weights need to be normalized (Korn and Graubard 1999). We consider two approaches:

Weight normalization 1 or WN(1):

Elliott 2009 proposed treating the nonprobability sample as a separate stratum and normalizing the pseudo-weights to preserve the proportion of each sample in the combined sample. For nonprobability sample unit i, the normalized pseudo-weight is (the superscript "ss" stands for "separate stratum"):

$$\hat{w}_i^{ss} = \frac{n_{np}}{n_p + n_{np}} \times \frac{\sum_{j=1}^{n_p + n_{np}} I\left(Z_j = 0\right) \tilde{w}_j}{\sum_{j=1}^{n_p + n_{np}} I\left(Z_j = 1\right) w_j} \times w_i, \tag{4}$$

where n_{np} is the size of the nonprobability sample, and n_p is the size of the probability sample. For probability sample unit i, the normalized weight is:

$$\hat{w}_i^{ss} = \frac{n_p}{n_p + n_{np}} \times \tilde{w}_i. \tag{5}$$

Weight normalization 2 or WN(2):

Though Elliot's method stabilizes the total weight, it represents a deviation from the original survey design by introducing an additional stratum. We propose a normalization strategy that is consistent with the original probability sample design by treating both nonprobability and

probability sample units as coming from the sampling strata defined by the design variables, and normalizing the pseudo-weights to preserve the total weight of each stratum as the population size in the combined sample. In the combined sample, for unit i of stratum s, the normalized pseudo-weight is (the superscript "dc" stands for probability sample "design consistent"):

$$\hat{w}_i^{dc} = \frac{N_s}{\sum_{j \in s} w_j} \times w_i, \tag{6}$$

where N_s is the population size in stratum s. Practically, the weight normalization step can be combined with the pseudo-weight step to simplify the coding.

2.3. Raking

Following standard survey practice, we implement raking to ensure that the sample distributions match those of the population. Depending on whether the population-level information is available, there are two ways to rake pseudo-weights. When population-level information is available, the raking procedure can be applied to match covariates with their marginal distributions in the population. When population-level information is unavailable, estimates based on the probability sample can be used as proxies to population quantities and the raking procedure can be applied to match up covariates with their marginal distributions in the probability sample. To represent the entire population, raked pseudo-weights are then proportionally inflated so that they sum up to the target population size. In our simulation, we consider the following three strategies:

Raking 1 or RK(1): without raking;
Raking 2 or RK(2): raking to the known marginal population distribution;
Raking 3 or RK(3): raking to the marginal distribution in the probability sample.

3. Real data example: The Buckeye Teen Health Study

The research presented in this paper was motivated by the BTHS. The goal of the study was to determine factors associated with the initiation of tobacco use. Participants were recruited through the use of both probability and nonprobability methods, specifically, ABS (Iannacchione 2011) and community recruitment, in ten Ohio counties, which consisted of one urban county (Franklin) and nine rural Appalachian counties (Brown, Guernsey, Lawrence, Muskingum, Scioto, Washington, Clermont, Morgan, and Noble). Of the 1,220 participants enrolled in the study, 991 (81.2%) were recruited through ABS and 229 (18.8%) were from community recruitment. All procedures were approved by the Institutional Review Board at The Ohio State University.

The ABS was implemented with stratification by counties. The addresses were taken from the US Postal Service computerized delivery sequence file. Within each stratum, the probability of selecting each address is equal. Initially, a postcard was sent stating that a letter would arrive with a study opportunity. Selected households were then sent a letter describing the study along with a $2 bill incentive and a short questionnaire, in which the adult members of the household were to be listed along with the gender and age of any children living in the household. If no response was received, a second letter and questionnaire, without the $2 incentive, were sent, approximately three weeks after the initial mailing with incentive and questionnaire. Households with an adolescent male between 11 and 16 who indicated interest in the study were contacted by a trained interviewer to ensure eligibility and then, if eligible, to schedule a baseline interview. We obtained both consent from the parent and assent from the youth. Only a single youth was interviewed per household. If multiple boys were eligible we interviewed the youth with the most recent birthday relative to the screening date.

The ABS recruitment took longer than initially anticipated and was thus supplemented with the community recruitment to obtain the target number of participants. During the fielding, the

Table 2. Hypothetic population for simulation.

	Description	Model
Stratification Variables		
S_1	Binary: 0, 1	
S_2	Categorical: $1-10$	
Covariates		
X_1	Continuous	Normal (Mean $= 20$, SD $= 3$)
X_2	Binary: 0, 1	$P = 0.9$ if $S_1 = 0$; and $P = 0.65$ if $S_1 = 1$
X_3	Categorical: 1, 2, 3	$P = (0.45, 0.3, 0.25)$ if $S_1 = 0$; and $P = (0.25, 0.3, 0.45)$ if $S_1 = 1$
X_4	Categorical: 1, 2, 3	$P = (0.1,\ 0.3,\ 0.6)$
Outcome Variables		
Y_1	Continuous; strongly affected by X_4	Normal (Mean $= 1 + 2S_1 + 3X_1 + 4X_2 + 5X_{4.2} + 6X_{4.3}$, SD $= 10$)
Y_2	Continuous; weakly affected by X_4	Normal (Mean $= 1 + 2S_1 + 3X_1 + 4X_2 + 0.5X_{4.2} + 0.6X_{4.3}$, SD $= 10$)
Y_3	Binary; strongly affected by X_4	$logit(p) = -2 + 0.2S_1 + 0.3X_1 - 0.4X_2 - 0.5X_{3.2} - 0.6X_{3.3} - 3X_{4.2} - 4X_{4.3}$
Y_4	Binary; weakly affected by X_4	$logit(p) = -5 + 0.2S_1 + 0.3X_1 - 0.4X_2 - 0.5X_{3.2} - 0.6X_{3.3} - 0.3X_{4.2} - 0.4X_{4.3}$

where $X_{i.j}$ is dummy/binary variable derived from variable X_i with reference value $X_i = 0$, and value $X_{i.j} = 1$ if $X_i = j$; otherwise, $X_{i.j} = 0$.

data collection through ABS resulted in a total of 991 respondents (20.3% unweighted response rate). The potential bias due to survey nonresponses was addressed through weighting. However, because the target precision based on designed sample size might not be achieved, we decided to add the nonprobability sample through community recruitment. Any household not sampled by the probability survey was eligible for the nonprobability survey. This supplementation was particularly important for the BTHS because of the low prevalence of tobacco use in the target population.

The study was advertised both on the radio and in local newspapers. The study staff also attended community-based events, including county fairs, farmer's markets, and other local events, to recruit participants. Investigators also spoke on the radio and on television to promote the study. Households recruited directly from the community were subsequently screened to determine eligibility and, if eligible, a baseline interview was scheduled. A total of 229 participants were recruited using these methods. Potential selection bias was addressed through calibration weighting based on the assumption of ignorability in selection.

In this paper, we focus on estimating the prevalence of tobacco use at baseline and mean body mass index (BMI), to illustrate various strategies for combining probability and nonprobability samples using both binary and continuous outcomes.

4. Simulation studies

We first generated a fixed finite hypothetic population, then applied different sampling designs and obtained repeated samples. Within each sample, different pseudo-weighting strategies were implemented, and survey weight-adjusted inferences were conducted. The results are summarized over repeated samples to evaluate the performance.

4.1. Hypothetic population

A finite population of 50,000 ($N = 50,000$) individuals, with two stratification variables, four covariates, and four outcome variables, was generated as the underlying true population. Table 2 summarizes the characteristics of each variable and how they were generated. Specifically, to

mimic the BTHS, two stratification variables, S_1 and S_2, were used. Variable S_1 had two strata (1 and 0), indicating the urbanicity status (urban or rural) of each county. Variable S_2 was nested within S_1 and had ten strata in total, representing counties. The first eight counties ($S_2 = 1, ..., 8$) were nested within $S_1 = 0$ and had population sizes of $N_1 = 1,000$, $N_2 = 1,400$, $N_3 = 1,800$, $N_4 = 2,200$, $N_5 = 2,600$, $N_6 = 3,000$, $N_7 = 5,000$, and $N_8 = 8,000$. The last two counties ($S_2 = 9$ and 10) were nested within $S_1 = 1$ and had population sizes of $N_9 = 10,000$ and $N_{10} = 15,000$. Four covariates were considered: X_1 was continuous (normally distributed), X_2 was binary, X_3 and X_4 were categorical with three categories each. To thoroughly examine the pseudo-weighting methods, two types of outcome variables were considered: variables Y_1 and Y_2 were continuous (normally distributed), with X_4 strongly associated with Y_1 and weakly associated with Y_2; variables Y_3 and Y_4 were binary, with X_4 strongly associated with Y_3 and weakly associated with Y_4. Variables Y_1 and Y_2 were generated based on S_1, X_1, X_2, and X_4 through linear models. Variables Y_3 and Y_4 were generated based on S_1, X_1, X_2, X_3, and X_4 through logistic regression models.

4.2. Sampling design

The simulation process repeatedly drew a 10% combined sample (5,000 individuals; $n = n_p + n_{np} = 5,000$) from the hypothetical population 5,000 times ($m = 5,000$ replications), where n_p indicates the sample size for probability sample and n_{np} indicates the sample size for nonprobability sample. To evaluate the practical performance of each method, five different combinations of probability and nonprobability sample proportions were considered: the proportions of probability sample were set at 10%, 25%, 50%, 75%, and 90%. For each combination, the probability sample was first selected using proportional allocation across strata and the nonprobability sample was then selected based on a logistic regression model with one outcome of interest as a covariate to mimic the nonprobability sampling mechanism.

Following Valliant and Dever 2011, we used a complex underlying probability sampling mechanism to generate the nonprobability sample. Though the assumption used in PW(3) and PW(4) is that outcome is independent of selection probability, to represent the practical mechanism, we include a situation where an outcome variable was included to reflect the fact that the underlying probability sampling mechanism was unknown and therefore could not be modeled correctly. The following mechanism was considered, where probabilities of being selected p depended on covariates (X_1, X_2, X_3, and X_4) and one outcome variable collected in the survey (Y_4). The nonprobability sample was selected based on the estimated probability from the following model:

$$logit(p) = -10 + X_1 - 2X_2 + 3X_{3.2} + 4X_{3.3} - 5X_{4.2} - 6X_{4.3} + 7Y_4 \tag{7}$$

Note that the selection of nonprobability sample depends on X_2 and X_3, both of which depend on the stratification variable S_1. So, this model allows the nonprobability samples to be "imbalanced" with regards to sampling stratification characteristics.

4.3. Metrics for evaluation

After the probability sample was drawn, sampling weights were first calculated based on the selection probabilities. Then the nonprobability sample was pooled with the probability sample to create pseudo-weights (PW). All four PW methods (i.e., PW(1), PW(2), PW(3), and PW(4)) discussed in Sec. 3 were applied.

In the IPA procedure, the logistic regression model estimating probabilities $P(Z_i = z \mid \mathbf{x}_i = \mathbf{x}_0)$, $z = 0, 1$, included all covariates, i.e., X_1, X_2, X_3, and X_4. In the PSA procedure, the weighted logistic regression model also included all covariates, X_1, X_2, X_3, and X_4,

Table 3. Summary of simulation methods.

Naïve	Unity	IPA	PSA	Probability Sample Only
1: PW(1) + RK(1)	4: PW(2) + WN(1) + RK(1)	10: PW(3) + WN(1) + RK(1)	16: PW(4) + WN(1) + RK(1)	22: RK(1)
2: PW(1) + RK(2)	5: PW(2) + WN(1) + RK(2)	11: PW(3) + WN(1) + RK(2)	17: PW(4) + WN(1) + RK(2)	23: RK(2)
3: PW(1) + RK(3)	6: PW(2) + WN(1) + RK(3)	12: PW(3) + WN(1) + RK(3)	18: PW(4) + WN(1) + RK(3)	24: RK(3)
	7: PW(2) + WN(2) + RK(1)	13: PW(3) + WN(2) + RK(1)	19: PW(4) + WN(2) + RK(1)	
	8: PW(2) + WN(2) + RK(2)	14: PW(3) + WN(2) + RK(2)	20: PW(4) + WN(2) + RK(2)	
	9: PW(2) + WN(2) + RK(3)	15: PW(3) + WN(2) + RK(3)	21: PW(4) + WN(2) + RK(3)	

Abbreviations: Pseudo-weight methods: PW(1): naïve method; PW(2): unity method; PW(3): inclusion probability adjusted method; PW(4): propensity score adjusted method. Weight normalization methods: WN(1): nonprobability sample as a separate stratum; WN(2): nonprobability sample strata consistent with probability sample. Raking methods: RK(1): No raking; RK(2): use external information to rake; RK(3): use internal information to rake.

to estimate the propensity score of belonging to the probability sample (Step 1). Based on these scores, data were divided into 20 equal-sized subgroups (Step 2).

After calculating pseudo-weights for the nonprobability sample (and updating weights for the probability sample in PW(1) and PW(4)), every unit in the combined sample has either a weight or a pseudo-weight. The combined sample can be treated as a stratified sample with different weights (except for the naïve weight method) within each stratum, and weight normalization (WN) and raking (RK) may be applied. The post-stratification (raking) step only involves categorical covariates, i.e., X_2, X_3, and X_4 using their marginal distributions.

Several statistical measures were reported in the simulation. To define them, let us first introduce the following notations:

N : Grand population size.

N_h : Population size in Stratum h.

n_j^h : Sample size in Stratum h in the j-th simulated combined sample; $j = 1, ..., m$.

m : Number of repeated simulation runs; $m = 5,000$.

$Y_{i,pop}$: Population mean of the i-th variable Y_i, $i = 1,\ 2,\ 3,\ 4$.

y_{ij}^{hk} : Value of k-th unit Y_i in Stratum h in the j-th simulated combined sample; $i = 1,\ 2,\ 3,\ 4,\ j = 1, ..., m,\ k = 1, ..., n_j^h,\ h = 1, ..., 10$.

w_j^{hk} : Weight/Pseudo-weight of k-th unit in Stratum h in the j-th simulated combined sample; $j = 1, ..., m,\ k = 1, ..., n_j^h,\ h = 1, ..., 10$.

$\overline{y}_{ij}$: Estimated mean of Y_i in the j-th simulated combined sample; $\overline{y}_{ij} = \frac{1}{N}\sum_{h=1}^{10}\sum_{k=1}^{n_j^h} w_j^{hk} y_{ij}^{hk}$; $i = 1,\ 2,\ 3,\ 4,\ j = 1, ..., m$.

$SE(\overline{y}_{ij})$: Standard error of $\overline{y}_{ij}$; which was calculated using the Taylor Series method (calculated using R survey package).

The reported statistical measures are summarized below:

- Percent Bias (% Bias)
  The difference between the mean of point estimates and the population truth $Y_{i,pop}$ as the percentage of $Y_{i,pop}$, $\left[\left(\frac{1}{m}\sum_{j=1}^{m}\overline{y}_{ij} - Y_{i,pop}\right)/Y_{i,pop}\right] \times 100$.
- 95% Confidence Interval Coverage (95% CI CVGE.)
  Percentage of 95% confidence intervals of $\overline{y}_{ij}$ that cover the population truth $Y_{i,pop}$. The 95% confidence interval of $\overline{y}_{ij}$ is defined as $(\overline{y}_{ij} - 1.96 \times SE(\overline{y}_{ij}),\ \overline{y}_{ij} + 1.96 \times SE(\overline{y}_{ij}))$.
- Monte Carlo Standard Error (SE.MC)
  Standard error of the point estimates, $\sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(\overline{y}_{ij} - \frac{1}{m}\sum_{j=1}^{m}\overline{y}_{ij}\right)^2}$.

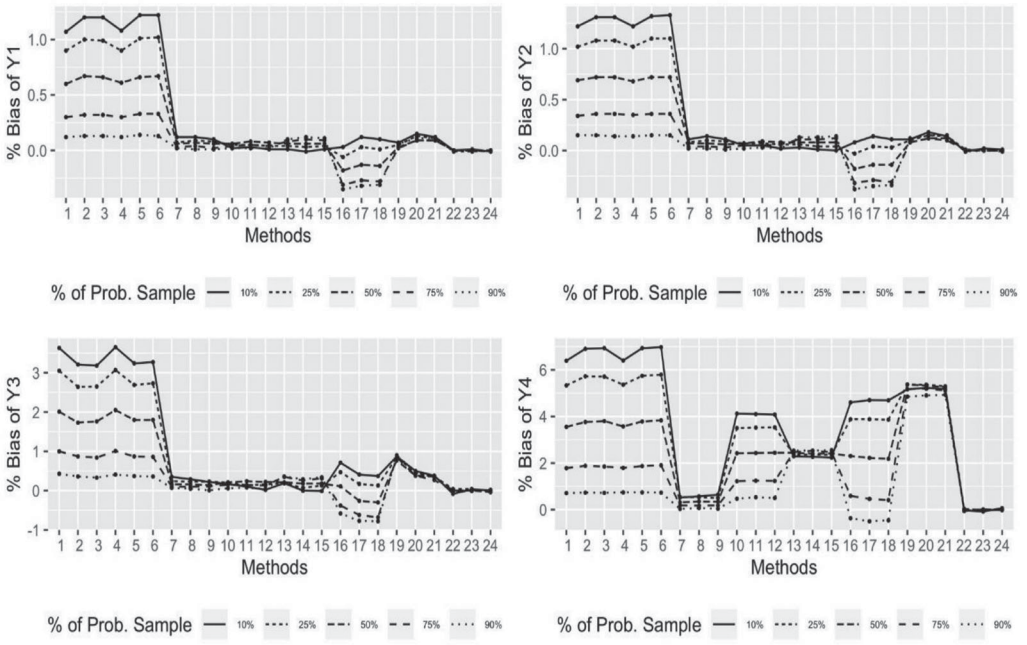


Figure 1. Percent bias of outcome variables.

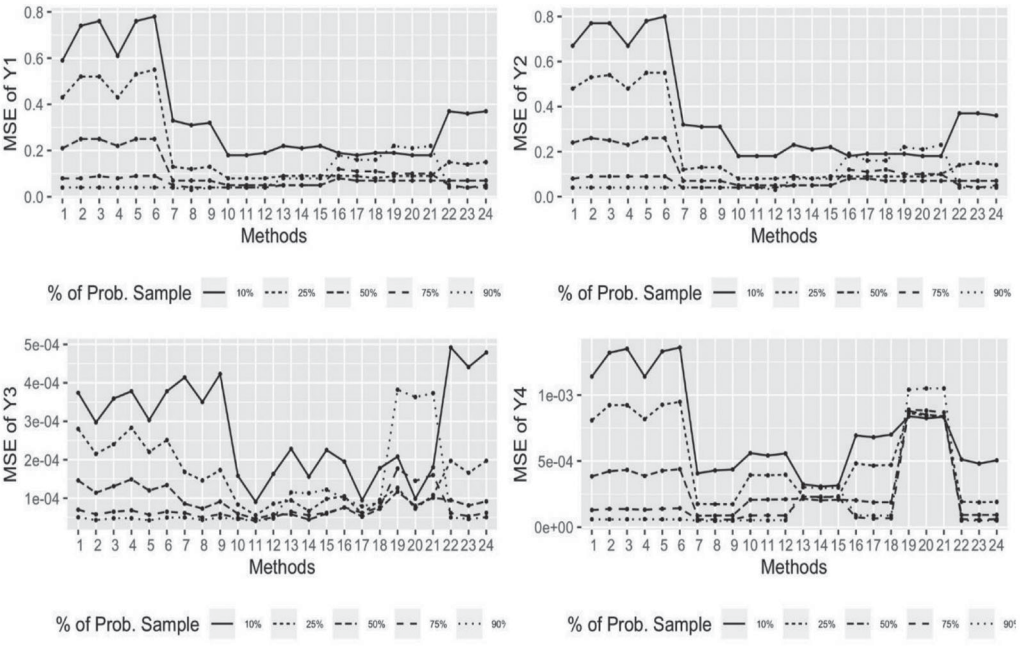


Figure 2. Mean squared error of outcome variables.

- Mean Squared Error (MSE)
  Mean squared error of the point estimates, $\frac{1}{m}\sum_{j=1}^{m}\left(\overline{y}_{ij} - Y_{i,pop}\right)^2$.

4.4. Simulation results

Twenty-four methods were compared (see Table 3) and the simulation results on bias and MSE are summarized in Figures 1 and 2. Additional results on coverage and SE.MC are reported in the supplementary material (Figures A.1 and A.2). Detailed numerical summaries can also be found in the supplementary material (Tables A.1–A.5). Three methods ($22-24$) use only the probability sample and thus serve as a benchmark to assess the performance of pseudo-weights. Estimates are reported for four outcome variables: a continuous variable (Y_1) that is highly affected by X_4, i.e., large X_4 coefficient in the data generating model; a continuous variable (Y_2) that is weakly affected by X_4, i.e., small X_4 coefficient in the data generating model; a binary variable (Y_3) that is highly affected by X_4; and a binary variable (Y_4) that is weakly affected by X_4.

With respect to the simulation scenarios, we found: Continuous outcomes were estimated with less percentage bias than binary outcomes. All bias measures dropped substantially as the proportion of probability sample increased. The 95% CI coverages were generally poor with the low proportion of probability sample, except for methods 7, 8, and 9. But the coverage improved substantially when the probability sample proportion increased. MSEs were not directly comparable between continuous and binary outcomes because of the scale difference, but they were also improved when the probability sample proportion increased, except for methods 16-21, the MSEs increased in Scenarios 4 and 5 (when the proportion of nonprobability sample was 25% and 10%, respectively).

With respect to the pseudo-weight methods being compared, we found that the naïve weight methods 1–3 performed poorly, with large biases and poor CI coverages. This is expected as they simply treat the nonprobability sample the same way as the probability sample.

The unity weight methods 4–9 resulted in improvements, some dramatic. If the nonprobability sample was treated as a separate stratum during weight normalization (i.e., WN(1)), methods 4–6 yielded poor results, similar to the naïve methods. However, if the weight normalization was consistent with the original design strata (i.e., WN(2)), methods 7–9 tended to yield very good overall performance, especially in terms of 95% CI coverages and relative biases. With low probability sample proportion, biases were a bit larger than inclusion probability adjusted weight methods (except for Y_4). The differences diminished when the probability sample proportion increased.

The model-based pseudo-weight methods 10–15 also show very good overall performance with the smallest biases and MSEs (except for Y_4) when the probability sample proportion was low. This is likely due to the use of the model. However, the 95% CI coverages were under the nominal level.

The propensity score adjusted pseudo-weight methods 16–21 showed fairly good results for Y_1, Y_2, and Y_3, when the probability sample proportion was low. But, when the probability sample proportion became high, they showed larger biases and MSEs than both the unity weight and the inclusion probability adjusted weight methods. Their 95% CI coverages were also not ideal, either below or above the nominal level. These methods were not good for Y_4, performing poorly with low probability sample proportion and yielding large biases even with high proportion (methods 19–21). This is likely because we used the average propensity score within each stratum to calculate the weight, which might result in inaccurate weight when the stratum covers a large range of propensity score values.

Overall, we would recommend methods 7–9 for survey practitioners, as they offer the most credible 95% CI coverages across all scenarios with small bias. When the proportion of probability sample is too low, however, alternative methods such as methods 13-15 provide better MSEs. Their results are also fairly consistent regardless of the raking strategy. This is likely due to the

weight normalization, which is based on the probability sample weights. The probability sample weights are usually raked to ensure the sample characteristics match with those from the population. Such effect may be carried over by the nonprobability sample, which makes the additional raking step less critical. When the probability sample proportion is high, inclusion probability adjusted weighting strategies coupled with appropriate weight normalization (methods 10–12) provides slightly more accurate estimates, but they tend to underestimate the variance when the probability sample proportion is low.

Last but not least, analyses using only the probability sample showed very good performance, even when the probability sample proportion was low, i.e., 10%. As expected, point estimates were virtually unbiased and 95% CI coverages were always adequate. But due to the smaller sample size used, the MSE results were not as good as pseudo-weight-based methods for binary outcomes when the proportion of probability sample was low. Overall, our recommended methods 7-9 and probability-sample-only methods had similar outcomes in terms of estimating population quantities, but methods 7-9 enjoy the benefit of including more sample, which is important for studies focusing on low-prevalence outcomes.

5. Results from The Buckeye Teen Health Study

One important focus of the BTHS is on risk factors for smokeless tobacco (ST) initiation and dual use of ST and cigarettes. To track the initiation of ST use or dual use, it is important to get an accurate estimate of any tobacco product use at baseline. Therefore, we applied the series of methods discussed in Sec. 2 to estimate the prevalence of tobacco use at baseline, using the binary variable "ever use of any tobacco products." To illustrate the estimation of a continuous outcome, we considered the body mass index (BMI), as it is an important indicator of adolescents' health development. BMI is calculated as weight (in kilograms) over height (in centimeters) squared. We also considered five covariates that are potentially related to the selection into probability or nonprobability samples: age (continuous), race (White vs. Non-White), annual household income (<$25 K, $25 K – $50 K, >$50 K), parents' highest education (<High School, GED/High School, >=College), and parents' tobacco use (Yes vs. No). Because the level of missing data for these covariates was low, hot deck single imputation was used to create a complete analytic dataset. Two participants had missing BMI, therefore, the analysis of BMI only included 1,218 subjects. The comparison of means and proportions for selected covariates and outcome variables between probability and nonprobability samples are given in Table 4.

To estimate pseudo-weights for the nonprobability sample, all five covariates were used in both the logistic regression model in the inclusion probability adjusted weight method (IPA) and the weighted logistic regression model in Step 1 of the propensity score adjusted weight method (PSA). As in the simulation study, in Step 2 of PSA, data were divided into 20 equal-sized subgroups, based on the estimated propensity scores. Since true covariate distributions in the population are unknown, the second raking method (RK(2)), which relies on external information, cannot be applied. In the third raking method (RK(3)), we used the probability sample component to produce the target marginal distribution quantities for all four discrete covariates (i.e., race, annual household income, parents' highest education, and parents' tobacco use), then raked the combined sample to match them. For population level quantities, the probability sample estimates may serve as good proxies, as its sample size is large and accounts for more than 80% of the total sample (as verified by our simulation studies). For comparison, weighted population mean and standard error estimates, both without raking (method 22) and with raking (method 24), were also reported using only the probability sample.

The results of the BTHS analysis are summarized in Table 5 (Y_1 for BMI and Y_2 for tobacco product use). The reported mean and SE correspond to $\bar{y}_{ij}$ and $SE(\bar{y}_{ij})$, respectively, defined in Sec. 4.3, except that here $j = m = 1$. We only used methods that performed reasonably well in

Table 4. BTHS mean and proportions of selected variables by type of samples.

Outcome	Probability sample, unweighted	Probability sample, weighted by sampling weight	Nonprobability sample, unweighted
Mean of Age	14.118	13.993	13.759
Proportion of White	0.774	0.721	0.690
Proportion of Household Income			
< 25 K	0.150	0.142	0.166
25K $-$ 50 K	0.194	0.211	0.218
> 50 K	0.656	0.648	0.616
Proportion of Parent's Highest Education:			
$<$ High School	0.024	0.023	0.035
GED/High School	0.122	0.106	0.096
$>=$ College	0.854	0.871	0.869
Proportion of Parents' Tobacco Use	0.322	0.307	0.301
BMI	22.860	22.679	22.704
Prevalence of Tobacco Products Use	0.191	0.166	0.157

Table 5. BTHS results. The methods are numbered in consistent with the simulation methods defined in Table 3.

	Y_1		Y_2	
Methods	Mean	SE	Mean	SE
7	22.68	0.32	0.17	0.02
9	22.72	0.33	0.17	0.01
10	22.74	0.28	0.17	0.01
12	22.78	0.28	0.17	0.01
13	22.73	0.30	0.16	0.01
15	22.81	0.31	0.16	0.02
19	22.50	0.42	0.17	0.04
21	22.57	0.44	0.18	0.04
22	22.68	0.32	0.17	0.02
24	22.72	0.33	0.17	0.01

the simulation study. All selected methods produced very similar results, with methods 7 and 9 being closest to the probability-sample-based estimates. These results are expected, since most of the participants (81.2%) were from the probability sample, and methods 7 and 9 performed very well in most of the simulated scenarios in Sec. 4. Inclusion probability adjusted weight methods 10 and 12 were also very good with smaller standard error estimates, but this might be at the cost of inadequate confidence interval coverage as we observed in simulations. Propensity score methods 19 and 21 seemed to underestimate the mean of BMI and overestimate the standard errors.

6. Discussion

Our research on identifying good practical strategies of combining probability and nonprobability samples is motivated by an adolescent male health survey. A large portion of participants were selected as a probability sample and a convenience sample was added to boost the sample size of potential smokers. Given the high proportion of probability sample, we focused on popular approaches under the pseudo-randomization framework, to take advantage of the probability sampling design. We conducted a simulation study to compare their empirical performance under scenarios with different probability sample proportions. The results suggested that unity weight methods with weight normalization that is consistent with the original sampling strata and the inclusion probability adjusted weight methods yield very good overall performance. This finding

is also consistent with Mercer, Lau, and Kennedy 2018, where they found that complex statistical methods may not yield more accurate results for online opt-in samples and emphasized the choice of covariates over the structure of the model. The real data analysis also revealed that both the unity weight method with normalization and inclusion probability adjusted weight methods performed well, with results close to the analysis only using the probability sample. This implies that using the pseudo-weight-based methods does not sacrifice the estimation accuracy and the additional benefit is to keep a larger dataset, which is critical for studies demanding a sufficient sample size to track individuals with low-prevalence outcomes.

Following the probability sampling design, we calculated the variance using a design-based approach (Taylor series linearization), which is consistent with the implementation in Robbins, Ghosh-Dastidar, and Ramchand (2021). Our simulation results also imply that combining a large portion of probability sample with a small portion of nonprobability sample seems to provide reasonable variance estimates. However, when the proportion of nonprobability sample is large, the estimates may become too unstable, not suitable for publication in practice. To further account for the fact that the pseudo weights are estimated, Valliant (2020) suggested the use of resampling-based technique, such as jackknife and bootstrap, both of which are computationally intensive methods. It will be an interesting future research topic to compare linearization and resampling methods to see if the additional computation cost is justified by the possible gain in variance estimation.

Since our findings are based on simulation studies, they depend on the simulation setup. We set the distribution and parameter values to mimic the BTHS design. There are several limitations: First, we only consider stratified sampling design. As clustered sampling design is also popular in practice, future simulation studies may need to include a clustering component to see how these methods perform differently. Second, technically, it is challenging to generate nonprobability sample in simulation. We followed Valliant and Dever 2011, but their approach is still based on a probability sampling model. The trick is that the true selection model is not known, so it can be viewed as nonprobability to some degree. In practice, the nature of the nonprobability sample might be more extreme than what a simple probability model can predict. It may add another layer of complexity in the estimation process. Unfortunately, there is not much literature regarding how to simulate nonprobability data and it could be an interesting direction to explore further. Third, we compared methods only under the pseudo-randomization framework. This is because our motivating example has a large portion of probability sample and the design-based inference should be more appropriate. When the portion of probability sample is low, the model-based inference may be helpful. A mis-specified model will introduce bias and our simulation study shows that inclusion probability adjusted weights tend to underestimate the variance, thus leading to low coverage of confidence intervals. A more rigorous comparison between pseudo-randomization methods and superpopulation models under low probability sample proportion settings may be needed. Finally, in our simulation we treated the weights for both the probability and nonprobability samples as if they had accounted for nonresponse so that potential nonresponse bias was not included in the simulation.

Acknowledgments

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Food and Drug Administration. The authors thank the reviewer and the editor for insight comments, which leads to a substantially improved version.

Declaration of interest statement

The authors declare no conflict of interest.

Funding

This work was supported by grant P50CA180908 from the National Cancer Institute and Food and Drug Administration's Center for Tobacco Products.

ORCID

Wenna Xi http://orcid.org/0000-0002-5427-5689
Bo Lu http://orcid.org/0000-0002-3807-7869

References

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46 (3):399–424. doi:10.1080/00273171.2011.568786.

Berzofsky, M., R. Williams, and P. Biemer. 2009. Combining probability and non-probability sampling methods: Model-aided sampling and the O* NET data collection program. *Survey Practice* 2 (6):1–6. doi:10.29115/SP-2009-0028.

Brick, J. M. 2011. The future of survey sampling. *Public Opinion Quarterly* 75 (5):872–88. doi:10.1093/poq/nfr045.

Cochran, W. G. 1977. *Sampling techniques*. 3rd ed. New York, NY: John Wiley & Sons.

Cornesse, C., A. G. Blom, D. Dutwin, J. A. Krosnick, E. D. De Leeuw, S. Legleye, J. Pasek, D. Pennay, B. Phillips, J. W. Sakshaug, et al. 2020. A review of conceptual approaches and empirical evidence on probability and non-probability sample survey research. *Journal of Survey Statistics and Methodology* 8 (1):4–36. doi:10.1093/jssam/smz041.

Czajka, J. L, and A. Beyler. 2016. *Declining response rates in federal surveys: Trends and implications (background paper)*. Mathematica Policy Research. https://aspe.hhs.gov/sites/default/files/private/pdf/255531/Decliningresponserates.pdf.

Dutwin, D, and T. D. Buskirk. 2017. Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly* 81 (S1): 213–39. doi:10.1093/poq/nfw061.

Elliott, M. R. 2009. Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* 2 (6):1–7. doi:10.29115/SP-2009-0025.

Elliott, M. R, and R. Valliant. 2017. Inference for nonprobability samples. *Statistical Science* 32 (2):249–64. doi:10.1214/16-STS598.

Evans, A. T., E. Peters, B. Keller-Hamilton, C. Loiewski, M. D. Slater, B. Lu, M. E. Roberts, and A. K. Ferketich. 2018. Warning size affects what adolescents recall from tobacco advertisements. *Tobacco Regulatory Science* 4 (3):79–87. doi:10.18001/TRS.4.3.7.

Ferrari, S, and F. Cribari-Neto. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31 (7):799–815. doi:10.1080/0266476042000214501.

Friedman, K. L., M. E. Roberts, B. Keller-Hamilton, K. A. Yates, E. D. Paskett, M. L. Berman, M. D. Slater, B. Lu, and A. K. Ferketich. 2018. Attitudes toward tobacco, alcohol, and non-alcoholic beverage advertisement themes among adolescent boys. *Substance Use & Misuse* 53 (10):1706–14. doi:10.1080/10826084.2018.1429473.

Groves, R. M. 2006. Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly* 70 (5):646–75. doi:10.1093/poq/nfl033.

Hahs-Vaughn, D. L. 2005. A primer for using and understanding weights with national datasets. *The Journal of Experimental Education* 73 (3):221–48. doi:10.3200/JEXE.73.3.221-248.

Iannacchione, V. G. 2011. The changing role of address-based sampling in survey research. *Public Opinion Quarterly* 75 (3):556–75. doi:10.1093/poq/nfr017.

Kohut, A., S. Keeter, C. Doherty, M. Dimock, and L. Christian. 2012. *Assessing the representativeness of public opinion surveys* Pew Research Center (Washington, DC). https://www.pewresearch.org/politics/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.

Korn, E. L, and B. I. Graubard. 1999. *Analysis of health surveys*. Vol. 323. New York, NY: John Wiley & Sons.

Lee, S, and R. Valliant. 2009. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research* 37 (3):319–43. doi:10.1177/0049124108329643.

Link, M. 2018. New data strategies: Nonprobability sampling, mobile, big data. *Quality Assurance in Education* 26 (2):303–14. doi:10.1108/QAE-06-2017-0029.

Lohr, S. L. 2019. *Sampling: Design and analysis*. 2nd ed. Boca Raton: Chapman and Hall/CRC.

Mercer, A., A. Lau, and C. Kennedy. January 2018. *For weighting online opt-in samples, what matters most?* Pew Research Center. http://www.pewresearch.org/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Robbins, M. W., B. Ghosh-Dastidar, and R. Ramchand. 2021. Blending probability and nonprobability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology* 9 (5):1114–45. doi:10.1093/jssam/smaa037.

Schonlau, M., A. Van Soest, and A. Kapteyn. 2007. Are 'Webographic' or attitudinal questions useful for adjusting estimates from Web surveys using propensity scoring?

Schonlau, M., K. Zapert, L. P. Simon, K. H. Sanstad, S. M. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner, and S. H. Berry. 2004. A comparison between responses from a propensity-weighted web survey and an identical RDD survey. *Social Science Computer Review* 22 (1):128–38. doi:10.1177/0894439303256551.

Valliant, R. 2020. Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology* 8 (2):231–63. doi:10.1093/jssam/smz003.

Valliant, R, and J. A. Dever. 2011. Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research* 40 (1):105–37. doi:10.1177/0049124110392533.

Vehovar, V., V. Toepoel, and S. Steinmetz. 2016. Non-probability sampling. In *The Sage Handbook of Survey Methods*, ed. C. Wolf, D. Joye, T.W. Smith, and Y. Fu, 329–45. Los Angeles, LA: SAGE.