Biometrika (2021), **103**, 1, *pp*. 1–15
Printed in Great Britain

Soft calibration for selection bias problems under mixed-effects models

BY CHENYIN GAO, SHU YANG

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
cgao6@ncsu.edu syang24@ncsu.edu

AND JAE KWANG KIM

Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.
jkim@iastate.edu

SUMMARY

Calibration weighting has been widely used to correct selection biases in non-probability sampling, missing data, and causal inference. The main idea is to calibrate the biased sample to the benchmark by adjusting the subject weights. However, hard calibration can produce enormous weights when an exact calibration is enforced on a large set of extraneous covariates. This article proposes a soft calibration scheme, in which the outcome and the selection indicator follow mixed-effects models. The scheme imposes an exact calibration on the fixed effects and an approximate calibration on the random effects. On the one hand, our soft calibration has an intrinsic connection with best linear unbiased prediction, which results in a more efficient estimation compared to hard calibration. On the other hand, soft calibration weighting estimation can be envisioned as penalized propensity score weight estimation, with the penalty term motivated by the mixed-effects structure. The asymptotic distribution and a valid variance estimator are derived for soft calibration. We demonstrate the superiority of the proposed estimator over other competitors in simulation studies and a real-data application.

Some key words: Inverse propensity score weighting; Latent ignorability; Penalized optimization; Restricted maximum likelihood estimation.

1. INTRODUCTION

Calibration weighting, or benchmark weighting, is popular in survey sampling, where probability sampling weights are adjusted to match the known population totals of the auxiliary variables for a possible efficiency gain (Deville & Särndal, 1992). The idea of calibration is related to the generalized regression estimator, a model-assisted estimator in survey sampling (Cassel et al., 1976; Särndal et al., 1992), which has later been extended to the functional model-assisted estimator (Cardot & Josserand, 2011), optimal model calibration (Wu & Sitter, 2001), calibration weighting using instrumental variables (Estevao & Särndal, 2000), empirical likelihood calibration (Wu & Rao, 2006), and multi-source data calibration (Yang & Ding, 2019).

In addition to gaining precision, calibration weighting has been widely used to correct selection bias in various contexts, including finite-population inferences using non-probability samples, missing data, and causal inference. Skinner (1999), Lundström & Särndal (1999), Dev-

ille (2000), Kott (2006) and Lee & Valliant (2009) employed calibration weighting to adjust for selection bias in non-probability samples by enforcing covariate similarity between the non-probability sample and a probability sample; see Yang & Kim (2020) for a comprehensive review. For missing-at-random data, inverse propensity score weighting creates a weighted sample that resembles the complete version of the original sample. Instead of directly inverting the propensity score, calibration weighting imposes conditions to emulate complete data and gains robustness against model misspecification (Han & Wang, 2013; Chen & Haziza, 2017; Lee et al., 2021, 2022). Similarly, for causal inference under the ignorability of treatment assignment, the purpose of calibration weighting is to achieve the covariate balance between treatment groups, thus mitigating confounding biases (Hainmueller, 2012; Anastasiade & Tillé, 2017). For example, the covariate balance propensity score introduced by Imai & Ratkovic (2014) uses a balancing measure as an objective function to estimate the propensity score.

Most existing works aim to calibrate all available auxiliary variables to known finite-population totals, a process known as hard calibration. However, hard calibration may not be necessary when there are many covariates, especially if some covariates are not predictive of the outcome. Over-calibration, or improper application of calibration weighting on too many variables, can lead to variance inflations (Kang & Schafer, 2007). To address this problem, subsequent research has sought to use penalization (Guggemos & Tillé, 2010; Athey et al., 2018; Ning et al., 2020) or regularization (Zubizarreta, 2015; Wong & Chan, 2018; Wang et al., 2022) to ease the calibration constraints on a subset of covariates, which we refer to as regularized calibration. Chattopadhyay et al. (2020) proposed minimal dispersion approximately balancing weights by optimizing some user-specified function. Other attempts have been made to reduce the range of calibration weights directly by trimming, smoothing, or stabilizing (Lazzeroni & Little, 1998; Yang & Ding, 2018). Many of these methods adopt mixed-effects modeling, which is particularly useful in small area estimation (Torabi & Rao, 2008), longitudinal data inference (Verbeke, 2000; Weiss, 2005), handling clustered data with cluster-specific nonignorable missingness (Kim et al., 2016), and causal inference with unmeasured cluster-level confounders (Yang, 2018).

In this article, we focus on the settings with the shared parameter/random-effects models of the outcome and the selection indicator (Follmann & Wu, 1995). The sample inclusion indicator in survey sampling, the response indicator in the missing data context, and the treatment assignment in causal inference are all examples of the selection indicator. As a result, our framework applies to a wide range of problems. The selection indicator in the shared parameter models is latently ignorable in the sense that the selection indicator and outcome are conditionally independent given the observed covariates and the unobserved random effects, entailing nonignorable selection. Under the linear mixed-effects model, we propose a soft calibration algorithm that enforces an exact calibration on fixed effects, see (6a), and an approximate calibration on random effects, see (6b). Our soft calibration exploits the correlation structure of random effects to construct the regularized constraints, which is different from typical regularized calibration methods that leverage sparsity or smoothness conditions (Tan, 2020; Ning et al., 2020). The soft calibration constraints are seemingly intricate but arise naturally from two paths towards constructing the best linear unbiased predictor $\widehat{\theta}_{\text{blup}}$, a minimization problem in (4) and a prediction approach in (5). Thus, the produced estimator has an intrinsic connection to $\widehat{\theta}_{\text{blup}}$ and can be more efficient than the hard-calibration estimator, especially when random effects weakly affect the outcome. Furthermore, the dual problem (7) of soft calibration also establishes a link between soft calibration and penalized propensity score weight estimation, leading to a ridge-type regression (Guggemos & Tillé, 2010).

The calibration weights are well-known to be obtained by optimizing the user-specified loss function, which is related to the modeling of the propensity scores. Because the constrained optimization formulation (6) separates the loss function from the calibration conditions, we can impose relaxed calibration conditions while forcefully bounding the range of weights by changing the loss function. Next, we can show that the soft-calibration estimator is consistent if either the outcome follows a linear mixed-effects model or the propensity score model is correctly specified. The asymptotic distribution and a valid variance estimator for the soft-calibration estimators are then established. Furthermore, augmentations with flexible outcome modeling can be used in conjunction with soft calibration to correct the remaining bias, if any. Finally, a data-adaptive approach aided by cross fitting is proposed to select the optimal tuning parameter that minimizes the finite-sample mean squared error. Proofs of all results are provided in the Supplementary Material.

2. BASIC SETUP

2.1. *Notations, ignorability, and hard calibration*

To fix ideas, we consider estimating the population mean of a study variable based on a non-probability sample and extend it to clustered missing data analysis in §3.3. Suppose that we have a finite population $\mathcal{F}_N = \{(x_i, y_i) : i \in \mathbb{U}\}$ with population size N and index set $\mathbb{U} = \{1, \ldots, N\}$, independently and identically following a super-population model ζ. We assume that x_i is available in the finite population, but the study variable y_i is observed only in the sample. Let $\mathbb{S} \subset \mathbb{U}$ be the index set of the sample of size n. Define the selection indicator δ_i as $\delta_i = 1$ if $i \in \mathbb{S}$ and 0 otherwise. The propensity score for unit i being selected in the sample is $\pi_i = \mathrm{pr}(\delta_i = 1 \mid x_i)$, which is unknown for the non-probability sample. For ease of presentation, we summarize all notations in Table 1 for reference.

Table 1. *Summary of the notations*

Notation	Definition
y_i, x_i, x_{1i}, x_{2i}	Individuals of study variable and covariate for unit i, $x_i = (x_{1i}^{\mathrm{T}}, x_{2i}^{\mathrm{T}})^{\mathrm{T}}$
$Y_{\mathbb{U}}, Y_{\mathbb{S}}$	Vectors of study variable, $Y_{\mathbb{U}} = (y_1, \cdots, y_N)^{\mathrm{T}}$, $Y_{\mathbb{S}} = \{y_i : i \in \mathbb{S}\}$
$X_{\mathbb{U}}, X_{1,\mathbb{U}}, X_{2,\mathbb{U}}$	Matrices of covariate for finite population $\mathbb{U}$, $X_{\mathbb{U}} = (X_{1,\mathbb{U}}, X_{2,\mathbb{U}}) \in \mathbb{R}^{N\times(p+q)}$
$X_{\mathbb{S}}, X_{1,\mathbb{S}}, X_{2,\mathbb{S}}$	Matrices of covariate for selected sample $\mathbb{S}$, $X_{\mathbb{S}} = (X_{1,\mathbb{S}}, X_{2,\mathbb{S}}) \in \mathbb{R}^{n\times(p+q)}$
$E_\delta(\cdot), E_\zeta(\cdot), E(\cdot)$	Expectations with respect to the selection δ, the model ζ, and both
$\mathrm{var}_\delta(\cdot), \mathrm{var}_\zeta(\cdot), \mathrm{var}(\cdot)$	Variances with respect to the selection δ, the model ζ, and both
$o(\cdot)$	$a_n = o(b_n)$ implies $a_n/b_n \to 0$ when $n \to \infty$
$O(\cdot)$	$a_n = O(b_n)$ implies $a_n/b_n \to C_0$ when $n \to \infty$ for some constant C_0
$o_{\mathbb{P}}(\cdot), O_{\mathbb{P}}(\cdot)$	Small and big order terms with respect to both the selection δ and model ζ

The goal is to estimate $\theta_N = N^{-1}\sum_{i\in\mathbb{U}} y_i$, and we consider a weighted estimator given by

$$\widehat{\theta}_w = \frac{1}{N}\sum_{i\in\mathbb{S}} w_i y_i. \tag{1}$$

If y_i follows the linear regression model $y_i = x_i^{\mathrm{T}}\beta + e_i$ with $E_\zeta(e_i \mid x_i) = 0$ and $\mathrm{var}_\zeta(e_i \mid x_i) = \sigma_e^2$, we may impose the following condition on the weights:

$$\sum_{i\in\mathbb{S}} w_i x_i = \sum_{i\in\mathbb{U}} x_i, \tag{2}$$

which is a sufficient condition for model calibration (Wu & Sitter, 2001) in the sense that $\sum_{i\in\mathbb{S}} w_i\widehat{y}_i = \sum_{i\in\mathbb{U}} \widehat{y}_i$, where $\widehat{y}_i$ is a prediction based on the linear model. If the sampling mech-

anism is ignorable with $\delta_i \perp\!\!\!\perp y_i \mid x_i$, condition (2) is sufficient for the unbiasedness of $\widehat{\theta}_w$. To find the optimal calibration estimator that minimizes the mean squared error of $\widehat{\theta}_w$ while satisfying (2) under the linear regression model, it suffices to minimize

$$E_\zeta\{(\widehat{\theta}_w - \theta_N)^2 \mid X_\mathbb{U}, \mathbb{S}\} = \frac{1}{N^2}\mathrm{var}_\zeta\left\{\sum_{i\in\mathbb{U}}(\delta_i w_i - 1)e_i \mid X_\mathbb{U}, \mathbb{S}\right\} = \frac{\sigma_e^2}{N^2}\sum_{i\in\mathbb{S}}(w_i-1)^2 + \text{const.},$$

where const. represents a constant that does not depend on $w = \{w_i : i \in \mathbb{S}\}$. Thus, we can formulate the hard calibration weighting problem as finding the minimizer of the square loss function $\sum_{i\in\mathbb{S}}(w_i-1)^2$ subject to condition (2).

2.2. *Mixed-effects models and latent ignorability*

We now partition x_i into two vectors x_{1i} (including an intercept) and x_{2i} with $\dim(x_{1i}) = p$ and $\dim(x_{2i}) = q$, related to fixed effects and random effects, respectively. This setup is particularly relevant in small area estimation, where x_{1i} is a low-dimensional vector of feature variables and x_{2i} is a possibly high-dimensional vector of small area indicators.

In these settings, selection ignorability can be restrictive because it excludes area-specific effects that affect both y_i and δ_i. To overcome this issue, we consider a linear mixed-effects super-population model:

$$y_i = x_{1i}^\mathrm{T}\beta + x_{2i}^\mathrm{T}u + e_i, \quad u \sim N\left(0, D_q\sigma_u^2\right), \quad e_i \sim N(0, q_i^{-1}\sigma_e^2), \quad u \perp\!\!\!\perp e_i \mid x_i, \tag{3}$$

where u is a q-dimensional vector of random effects with a positive-definite covariance matrix D_q, e_i is the heteroscedastic random error with known q_i^{-1}, and σ_e^2 and σ_u^2 characterize the variances of individual errors and random effects, respectively. Typically, we consider $q_i = 1$ for $i \in \mathbb{S}$ but unequal q_i's are also desired in some situations; see Remark 5 in Devaud & Tillé (2019). Next, we make the following assumptions for the sampling mechanism.

Assumption 1 (*Latent ignorability*). The sampling mechanism is ignorable given (x_i, u): $\delta_i \perp\!\!\!\perp y_i \mid (x_i, u)$ for all $i \in \mathbb{U}$.

Assumption 2 (*Positivity*). $0 < \underline{d} < Nn^{-1}\mathrm{pr}(\delta_i = 1 \mid x_i, u) < \overline{d} < 1$ for all x_i and u.

Assumption 1 leads to shared parameter/random-effects models of δ_i and y_i. In the missing data context with clustered data, it is called cluster-specific nonignorable missingness (Yuan & Little, 2007). In the context of causal inference, it is called cluster-specific nonignorable treatment assignment (Yang, 2018). Assumption 1 relaxes the ignorability assumption by allowing unobserved random effects to affect both y_i and δ_i. Assumption 2 implies that the sample support $\{x_i : i \in \mathbb{S}\}$ coincides with the support of x_i in the population.

2.3. *Soft calibration for the best linear unbiased predictor*

Under model (3) and Assumptions 1-2, we wish to develop the optimal calibration estimator $\widehat{\theta}_w$ by minimizing the mean squared error. Following Hirshberg et al. (2019)'s minimax imbalance strategy, we minimize

$$\sup_{\beta\in\mathcal{M}} E_\zeta\{(\widehat{\theta}_w - \theta_N)^2 \mid X_\mathbb{U}, \mathbb{S}\} = \sup_{\beta\in\mathcal{M}} \frac{1}{N^2}(w^\mathrm{T}X_{1,\mathbb{S}} - 1_N^\mathrm{T}X_{1,\mathbb{U}})\beta\beta^\mathrm{T}(w^\mathrm{T}X_{1,\mathbb{S}} - 1_N^\mathrm{T}X_{1,\mathbb{U}})^\mathrm{T}$$
$$+ \frac{\sigma_e^2}{N^2}\left\{\sum_{i\in\mathbb{S}} q_i^{-1}(w_i-1)^2 + \gamma^{-1}(w^\mathrm{T}X_{2,\mathbb{S}} - 1_N^\mathrm{T}X_{2,\mathbb{U}})D_q(w^\mathrm{T}X_{2,\mathbb{S}} - 1_N X_{2,\mathbb{U}})^\mathrm{T}\right\} \tag{4}$$

with respect to w, where $\mathcal{M}$ is a convex subset of $\mathbb{R}^p$ that contains the true β. Since $\mathcal{M}$ may be unbounded without prior knowledge, the minimax problem results in an exact calibration condition $w^{\mathrm{T}}X_{1,\mathbb{S}} = 1_N^{\mathrm{T}}X_{1,\mathbb{U}}$ to diminish the first term of the above equation. The remaining objective function (4) leads to a generalized ridge regression problem (Bardsley & Chambers, 1984) augmented with a data-dependent penalty, where $\gamma^{-1} = \sigma_u^2/\sigma_e^2$ determines the level of calibration for $X_{2,\mathbb{S}}$: if γ is close to zero, the calibration for $X_{2,\mathbb{S}}$ is nearly exact; and if γ is large, the calibration for $X_{2,\mathbb{S}}$ is greatly relaxed.

In addition, the minimum of (4) should coincide with $\widehat{\theta}_{\mathrm{blup}} = N^{-1}\sum_{i\in\mathbb{U}}(x_{1i}^{\mathrm{T}}\widehat{\beta} + x_{2i}^{\mathrm{T}}\widehat{u})$, where $(\widehat{\beta}, \widehat{u})$ is the solution to the following score equations for the linear mixed-effects model:

$$\begin{pmatrix} \sum_{i\in\mathbb{S}} q_i x_{1i}x_{1i}^{\mathrm{T}} & \sum_{i\in\mathbb{S}} q_i x_{1i}x_{2i}^{\mathrm{T}} \\ \sum_{i\in\mathbb{S}} q_i x_{2i}x_{1i}^{\mathrm{T}} & \sum_{i\in\mathbb{S}} q_i x_{2i}x_{2i}^{\mathrm{T}} + \gamma D_q^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} \sum_{i\in\mathbb{S}} q_i x_{1i}y_i \\ \sum_{i\in\mathbb{S}} q_i x_{2i}y_i \end{pmatrix}. \tag{5}$$

By rewriting $\widehat{\theta}_{\mathrm{blup}}$ as a weighted estimator $w^{\mathrm{T}}Y_{\mathbb{S}}$, the weights satisfy

$$\begin{aligned} w^{\mathrm{T}}X_{\mathbb{S}} &= 1_N^{\mathrm{T}}X_{\mathbb{U}} \left\{ \sum_{i\in\mathbb{S}} q_i x_i x_i^{\mathrm{T}} + \gamma \mathrm{diag}(0, D_q^{-1}) \right\}^{-1} \sum_{i\in\mathbb{S}} q_i x_i x_i^{\mathrm{T}} \\ &= 1_N^{\mathrm{T}}X_{\mathbb{U}} \left[I_{p+q} - \gamma \left\{ \sum_{i\in\mathbb{S}} q_i x_i x_i^{\mathrm{T}} + \gamma \mathrm{diag}(0, D_q^{-1}) \right\}^{-1} \mathrm{diag}(0, D_q^{-1}) \right], \end{aligned}$$

where the second equality is derived by repeatedly applying the Woodbury matrix identity. Therefore, minimizing (4) can be reformulated as a constrained optimization with exact calibration on x_{1i} and approximate calibration on x_{2i}:

$$\begin{aligned} \min_{w} \quad & \sum_{i\in\mathbb{U}} \delta_i Q(w_i) = \sum_{i\in\mathbb{S}} q_i^{-1}(w_i - 1)^2, \\ \text{s.t.} \quad & \sum_{i\in\mathbb{S}} w_i x_{1i} = \sum_{i\in\mathbb{U}} x_{1i}, \qquad (6a) \\ & \sum_{i\in\mathbb{S}} w_i x_{2i} = \sum_{i\in\mathbb{U}} x_{2i} + \sum_{i\in\mathbb{U}} M_{\mathbb{S}}^{\mathrm{T}} x_{1i} + \sum_{i\in\mathbb{U}} R_{\mathbb{S}}^{\mathrm{T}} x_{2i}, \qquad (6b) \end{aligned}$$

where $M_{\mathbb{S}} = -\gamma D_{12} D_q^{-1}$, $R_{\mathbb{S}} = -\gamma D_{22} D_q^{-1}$, and $\{\sum_{i\in\mathbb{S}} q_i x_i x_i^{\mathrm{T}} + \gamma \mathrm{diag}(0, D_q^{-1})\}^{-1} = [D_{11}, D_{12} \mid D_{21}, D_{22}]$. The solution is denoted by $\widehat{w}^{(\mathrm{SQ})} = \{\widehat{w}_i^{(\mathrm{SQ})} : i \in \mathbb{S}\}$, giving rise to $\widehat{\theta}_w^{(\mathrm{SQ})} = N^{-1}\sum_{i\in\mathbb{S}} \widehat{w}_i^{(\mathrm{SQ})} y_i$, where the superscript SQ reflects the use of the square loss.

Proposition 1 reveals the intrinsic connection between soft calibration based on square loss and $\widehat{\theta}_{\mathrm{blup}}$ under the mixed-effects model (3).

PROPOSITION 1. *Under Assumptions 1 and 2 and the model (3), we have* $\widehat{\theta}_w^{(\mathrm{SQ})} = \widehat{\theta}_{\mathrm{blup}}$ *for fixed* $\gamma = \sigma_e^2/\sigma_u^2$.

Through the lens of $\widehat{\theta}_{\mathrm{blup}}$ derived from (4) or (5), the soft-calibration estimator is optimal under model (3) and consistent under any sampling design that satisfies the latent ignorability by Proposition 1.

2.4. *Soft calibration for penalized propensity score weight estimation*

In the proof of Proposition 1, we show that the square loss function is equivalent to assuming a linear regression model for the calibration weight. However, it is possible to obtain negative values that may not be acceptable to practitioners. One advantage of casting the soft-calibration

estimator as a solution to the constrained optimization problem (6) is that it directly leads to a mixed-effects model for the calibration weight through the link function $w(\cdot)$, which allows flexible estimation by adopting other loss functions $Q(\cdot)$. In particular, we consider the *dual problem* of (6) for optimization purposes, which is to minimize a penalized convex function:

$$G(c) = -\sum_{i\in\mathbb{U}} \delta_i Q\{w(c^{\mathrm{T}}x_i)\} + \left\{\sum_{i\in\mathbb{S}} w(c^{\mathrm{T}}x_i)x_i^{\mathrm{T}} - (1_N^{\mathrm{T}}X_{1,\mathbb{U}}, 1_N^{\mathrm{T}}X_{2,\mathbb{U}} + NT_r)\right\} c \tag{7a}$$

$$= \sum_{i\in\mathbb{U}} \delta_i g(c^{\mathrm{T}}x_i) - (1_N^{\mathrm{T}}X_{1,\mathbb{U}})c_1 - (1_N^{\mathrm{T}}X_{2,\mathbb{U}} + NT_r)c_2, \tag{7b}$$

where $g(\cdot)$ is the convex conjugate function of $Q(\cdot)$, $T_r = N^{-1}\sum_{i\in\mathbb{U}}(x_{1i}^{\mathrm{T}}M_{\mathbb{S}} + x_{2i}^{\mathrm{T}}R_{\mathbb{S}})$ is the adjustment for soft calibration, and $c = (c_1^{\mathrm{T}}, c_2^{\mathrm{T}})^{\mathrm{T}}$ is a vector of Lagrange multipliers with $c_2 = D_\delta u$ for a suitable invertible matrix D_δ, featuring a shared random-effects model with the outcome (Gao, 2004). Table 2 provides some examples of loss functions $Q(\cdot)$ and their associated $g(\cdot)$ and $w(\cdot)$. These loss functions belong to a general class of empirical minimum discrepancy measures (Read & Cressie, 2012), which can be considered as measuring the aggregate distance between the weights w and a n-vector of uniform weights 1_n.

Table 2. *Correspondence of loss functions $Q(w_i)$, the convex conjugate functions $g(z_i)$ and the weight models $w(z_i)$ when weights are adjusted to satisfy the calibration constraints for the first moments of x_i*

	$Q(w_i)$	$g(z_i)$	$w(z_i)$
Squared loss	$q_i^{-1}(w_i-1)^2/2$	$z_i + q_i z_i^2/2$	$1 + q_i z_i$
Entropy divergence	$q_i^{-1}\{w_i\log(w_i) - w_i + 1\}$	$q_i^{-1}\{\exp(q_i z_i) - 1\}$	$\exp(q_i z_i)$
Empirical Likelihood	$q_i^{-1}\{-\log(w_i) - 1 + w_i\}$	$-q_i^{-1}\log(1 - q_i z_i)$	$(1 - q_i z_i)^{-1}$
Maximum entropy	$q_i^{-1}(w_i - 1)\{\log(w_i - 1) - 1\}$	$z_i + q_i^{-1}\exp(q_i z_i)$	$1 + \exp(q_i z_i)$

PROPOSITION 2. *If $\widehat{c}$ is the minimizer of (7b), the calibration weights $w(\widehat{c}^{\mathrm{T}}x_i)$ attain the soft calibration conditions (6a) and (6b).*

Proposition 2 is justified since (7b) gives a dual optimization for solving the constrained optimization in (6). Furthermore, the penalized estimation in (7a) is closely related to the L_2 penalized propensity score weight estimator, which is, however, not optimal as its penalty term does not account for the correlation structure of the mixed effects; see §B.3 of the Supplementary Material for numerical details. In view of the Lagrangian function (7a), the soft-calibration estimator enforces an exact calibration on x_{1i} while penalizing a large discrepancy of imbalances between $\sum_{i\in\mathbb{S}} w_i x_{2i}$ and $\sum_{i\in\mathbb{U}} x_{2i}$, thus avoiding posing overly stringent constraints.

Remark 1. Let $\mathbb{A} = \{w : w^{\mathrm{T}}X_{\mathbb{S}} = 1_N^{\mathrm{T}}X_{\mathbb{U}} + (0_p^{\mathrm{T}}, NT_r)\}$ be a set of solutions to the soft calibration conditions. Assume that $Q(w)$ is strictly convex and smooth, defined in $\mathbb{W}$ that includes 1. Assume that $\mathbb{W}$ is either a compact set or an open set with $\lim_{w\to\partial\mathbb{W}} |Q(w)| = \infty$, where $\partial\mathbb{W}$ denotes the boundary of the set $\mathbb{W}$, (7) has a unique optimum with probability 1 when $\mathbb{A}\cap\mathbb{W} \neq \emptyset$.

In finite samples, a unique optimum of (7) may not exist due to conflicting conditions imposed for calibration. For example, calibration weights are restricted to an overly bounded support $\mathbb{W}$ to reduce the impact of outliers; see §B.2, which might render $\mathbb{A}\cap\mathbb{W}$ empty. One remedy for this issue is to adopt a Moore-Penrose generalized inverse (Devaud & Tillé, 2019) for the Newton-type method to achieve a solution even when $\mathbb{A}\cap\mathbb{W} = \emptyset$.

3. MAIN THEORY

3.1. *Bias correction and asymptotic properties*

In this section, we establish the asymptotic properties of $\widehat{\theta}_w$ under the general loss function $Q(w)$ and adopt the joint randomization framework for inference, which considers both the super-population mixed-effects model ζ and the sampling mechanism δ (Isaki & Fuller, 1982). Before delving into the technical details, we assume the following regularity conditions.

Assumption 3 (*Regularity conditions*). (a) The matrices $n^{-1}X_{\mathbb{S}}^{\mathrm{T}}X_{\mathbb{S}} = \Sigma_n$ for any sample $\mathbb{S}$, and $N^{-1}X_{\mathbb{U}}^{\mathrm{T}}X_{\mathbb{U}} = \Sigma_N$ are positive-definite; (b) There exists some constant C such that $\|x_i\|^2 < qC$ for all $i \in \mathbb{U}$; (c) The finite population is a random sample of a super-population model (3) satisfying $N^{-1}\sum_{i\in\mathbb{U}} y_i^{2+\alpha} < \infty$ for some $\alpha > 0$ with $N \to \infty$.

Assumptions 3(a) and (b) are standard regularity conditions related to the auxiliary variables (Portnoy, 1984; Dai et al., 2018; Chauvet & Goga, 2022). Assumption 3(c) requires the moment conditions to employ the central limit theorem. In contrast to hard calibration, the inexact calibration scheme for x_{2i} involves a correction term on the right-hand side of (6b), incurring an additional term in $\widehat{\theta}_w - \theta_N$:

$$\widehat{\theta}_w - \theta_N = N^{-1}\gamma_n(1_N^{\mathrm{T}}X_{1,\mathbb{U}}D_{12} + 1_N^{\mathrm{T}}X_{2,\mathbb{U}}D_{22})D_q^{-1}u + N^{-1}\sum_{i\in\mathbb{U}}(\delta_i w_i - 1)e_i, \qquad (8)$$

where γ_n is considered as a finite-sample tuning parameter for γ. In §3.2, we propose a data-adaptive approach to select γ_n that minimizes the estimated mean squared error of the soft-calibration estimator.

The following theorem characterizes the asymptotic properties of $\widehat{\theta}_w$.

THEOREM 1. *Suppose Assumptions 1-3, the conditions for $Q(w)$ in Remark 1 hold and $\gamma_n = o(n^{1/2}q^{-1/2})$, the soft-calibration estimator $\widehat{\theta}_w$ satisfies $\widehat{\theta}_w - \theta_N = N^{-1}\sum_{i\in\mathbb{U}}\psi_i(c^*) - \theta_N + o_{\mathbb{P}}(n^{-1/2})$, where c^* is the solution to $E\{\partial G(c)/\partial c \mid X_{\mathbb{U}}, u\} = 0$,*

$$\psi_i(c^*) = B(c^*)x_{i,\mathrm{SC}} + \delta_i w(c^{*\mathrm{T}}x_i)\eta_i(c^*), \quad \eta_i(c^*) = y_i - B(c^*)x_i, \qquad (9)$$

$B(c^*) = \{\sum_{i\in\mathbb{U}}\delta_i w'(c^{*\mathrm{T}}x_i)x_i y_i\}\{\sum_{i\in\mathbb{U}}\delta_i w'(c^{*\mathrm{T}}x_i)x_i x_i^{\mathrm{T}}\}^{-1}$, and $x_{i,\mathrm{SC}} = \{x_{1i}^{\mathrm{T}}, x_{1i}^{\mathrm{T}}M_{\mathbb{S}} + x_{2i}^{\mathrm{T}}(I_q + R_{\mathbb{S}})\}^{\mathrm{T}}$. As a result, if either the outcome y_i follows a linear mixed-effects model or $Q(w)$ entails a correct propensity score model, we have $n^{1/2}(\widehat{\theta}_w - \theta_N) \to N(0, V_1 + V_2)$ as $n \to \infty$, where

$$V_1 = \lim_{n\to\infty}\frac{n}{N^2}E_\zeta\left[\mathrm{var}_\delta\left\{\sum_{i\in\mathbb{U}}\delta_i w(c^{*\mathrm{T}}x_i)\eta_i(c^*) \mid X_{\mathbb{U}}, u, Y_{\mathbb{S}}\right\} \mid X_{\mathbb{U}}\right],$$

and

$$V_2 = \lim_{n\to\infty}\frac{n}{N^2}\mathrm{var}_\zeta\left[E_\delta\left\{\sum_{i\in\mathbb{U}}\psi_i(c^*) \mid X_{\mathbb{U}}, u, Y_{\mathbb{S}}\right\} \mid X_{\mathbb{U}}\right].$$

Theorem 1 states that $\widehat{\theta}_w$ is doubly robust as its consistency requires the outcome following a linear mixed-effects model or the propensity score being correctly specified. We now estimate V_1 and V_2 by $\widehat{V_1}$ and $\widehat{V_2}$, respectively, in Theorem 2.

THEOREM 2. *Under the assumptions in Theorem 1, we have* $\widehat{V}_1 = nN^{-2}\sum_{i\in\mathbb{S}} w(\widehat{c}^{\mathrm{T}}x_i)^2\eta_i(\widehat{c})^2 \to V_1$ *and* $\widehat{V}_2 = nN^{-2}\sum_{i\in\mathbb{S}} w(\widehat{c}^{\mathrm{T}}x_i)(y_i - x_{1i}^{\mathrm{T}}\widehat{\beta})^2 \to V_2$ *in probability, where* $\widehat{\beta} = D_{11}\sum_{i\in\mathbb{S}} q_i x_{1i} y_i + D_{12}\sum_{i\in\mathbb{S}} q_i x_{2i} y_i$

Theorem 2 estimates V_1 and V_2 by applying the standard variance estimator formula with c^* replaced by $\widehat{c}$. As Shao & Steel (1999) show that the order of V_2/V_1 is $O(n/N)$; thus if the sampling fraction n/N is negligible, we only need to estimate V_1.

Remark 2. In Theorem 1, we need $\gamma_n = o(n^{1/2}q^{-1/2})$ to make the bias term (8) negligible. If the bias term does not dwindle away, one can use a bias-corrected estimator $\widehat{\theta}_{\mathrm{bc}}$ to correct the remaining bias after soft calibration weighting. Denote $\widehat{\theta}_{\mathrm{bc}} = \widehat{\theta}_w - N^{-1}\sum_{i\in\mathbb{U}}\{\delta_i w(\widehat{c}^{\mathrm{T}}x_i) - 1\}\widehat{\mu}_i$, which combines soft calibration with the fitted outcomes $\widehat{\mu}_i$ by flexible modeling, similar to Ben-Michael et al. (2021) and Avagyan & Vansteelandt (2021).

As an example, if we combine the soft-calibration estimator with best linear unbiased prediction $\widehat{\mu}_i = x_{1i}^{\mathrm{T}}\widehat{\beta} + x_{2i}^{\mathrm{T}}\widehat{u}$, γ_n is allowed to grow faster with n than requested in Theorem 1 under the linear mixed-effects model, implying that $\widehat{\theta}_{\mathrm{bc}}$ is more robust than $\widehat{\theta}_w$ against the rate requirement for γ_n. Other choices for outcome models can also effectively reduce the left-over bias as long as they can approximate the true outcome $E_\zeta(y_i \mid x_i)$ well enough. A detailed discussion of its asymptotic properties is deferred to §A.8 of the Supplementary Material.

3.2. *Data-adaptive tuning parameter selection*

To properly choose the tuning parameter γ_n, we propose a data-adaptive cross-fitting strategy that targets minimizing the mean squared error of the soft-calibration estimator $\widehat{\theta}_w$. Specifically, we divide the data into $\mathcal{B}$ disjoint groups $\mathcal{I}_b, b = 1, \cdots, \mathcal{B}$. Let $\widehat{c}_{-k}$ and $\widehat{\beta}_{-k}$ denote the estimator of c^* and β computed using the observations from all the folds except the k-th fold based on the soft conditions with the tuning parameter γ_n. The estimated mean squared error will be

$$\mathrm{MSE}(\widehat{\theta}_w;\gamma_n) = \frac{1}{\mathcal{B}}\sum_{k=1}^{\mathcal{B}}\left[\left\{\frac{\mathcal{B}}{N}\sum_{i\in\mathcal{I}_k}\delta_i w(\widehat{c}_{-k}^{\mathrm{T}}x_i)y_i\right\} - \theta_N\right]^2$$
$$+ \frac{1}{\mathcal{B}}\sum_{k=1}^{\mathcal{B}}\frac{\mathcal{B}^2}{N^2}\left[\sum_{i\in\mathcal{I}_k}\delta_i w(\widehat{c}_{-k}^{\mathrm{T}}x_i)^2\{y_i - B(\widehat{c}_{-k})x_i\}^2 + \sum_{i\in\mathcal{I}_k}\delta_i w(\widehat{c}_{-k}^{\mathrm{T}}x_i)(y_i - x_{1i}^{\mathrm{T}}\widehat{\beta}_{-k})^2\right],$$

where the unknown parameter θ_N is approximated by the hard-calibration estimator $\widehat{\theta}_{\mathrm{hc}}$ as a proxy. Given this cross-fitting scheme, $\mathrm{MSE}(\widehat{\theta}_w;\gamma_n)$ is able to approximate the true mean squared error with negligible bias. A similar strategy has been used by Xiao et al. (2013) for tuning parameter selection in other contexts. We select γ_n by minimizing the estimated mean squared error of $\widehat{\theta}_w$ over a discrete grid $\{\gamma_n^* \times 10^j : j = -5, \ldots, 5\}$, where γ_n^* is a user-provided value. Our tuning strategy involves specifying γ_n^* and one candidate can be $\widehat{\sigma}_e^2/\widehat{\sigma}_u^2$, where $\widehat{\sigma}_e^2$ and $\widehat{\sigma}_u^2$ are the restricted maximum likelihood estimators of σ_e^2 and σ_u^2, respectively (Golub et al., 1979).

3.3. *Cluster-specific Nonignorable Missingness*

We now consider one important extension of latent ignorability to cluster-specific nonignorable missingness, and another extension to causal inference in the presence of unmeasured cluster-level confounders is presented in §B.5. Following the conventional notations for clustered data, consider the finite population $\mathcal{F}_N = \{(x_{ij}, y_{ij}, \delta_{ij}) : i = 1, \ldots, K, j = 1, \ldots, N_i\}$, where

i indexes the cluster and j indexes the unit within each cluster, y_{ij} is the outcome of interest for the j-th unit in cluster i, which is subject to missingness, $x_{ij} \in \mathbb{R}^p$ is the vector of observed covariates, δ_{ij} is the response indicator with value one if y_{ij} is observed and zero otherwise, and $N = \sum_{i=1}^{K} N_i$ is the population size. The parameter of interest is $\theta_N = N^{-1} \sum_{i=1}^{K} \sum_{j=1}^{N_i} y_{ij}$. We consider the two-stage cluster sampling: in the first stage, k clusters are selected from K clusters with cluster sampling weights d_i, and in the second stage, a random sample of n_i units is selected from each sampled cluster i with unit sampling weights N_i/n_i. The sample size is $n = \sum_{i=1}^{k} n_i$. Assume the outcome follows the linear mixed-effects model

$$y_{ij} = x_{ij}^{\mathrm{T}}\beta + a_i + e_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}a + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i,$$

where $a = (a_1, \ldots, a_k)^{\mathrm{T}}$ are the latent cluster-specific random effects, and $z_{ij} = s_i$ with s_i being the canonical coordinate basis for $\mathbb{R}^k$ as the cluster indicator. Here, x_{ij}, z_{ij} and a are the counterparts of x_{1i}, x_{2i} and u in §2.

In the presence of missing data, the sample average of the observed y_{ij} even adjusted for sampling design weights may be biased for θ_N due to the selection bias associated with the respondents. To correct such selection bias, the calibrated propensity score method proposed by Kim et al. (2016) imposes the following hard calibration constraints for both fixed effects and cluster effects:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}\delta_{ij}w_{ij}x_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}x_{ij}, \tag{10}$$

and $\sum_{j=1}^{n_i} d_{ij}\delta_{ij}w_{ij} = \sum_{j=1}^{n_i} d_{ij}$ for $i = 1, \ldots, k$ with $d_{ij} = d_i N_i n_i^{-1}$. The calibration constraints for the cluster effects may be stringent when the clusters weakly affect the outcome and may be relaxed to the following under soft calibration

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}\delta_{ij}w_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}, \tag{11}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}\delta_{ij}w_{ij}z_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}z_{ij} + \sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}M_{\mathbb{S}}^{\mathrm{T}}x_{ij} + \sum_{j=1}^{n_i} d_{ij}R_{\mathbb{S}}^{\mathrm{T}}z_{ij}, \tag{12}$$

where (11) is still an exact constraint forcing the weighted estimator of the population size to be the same as the design-weighted estimator, and (12) is an approximate calibration for cluster effects. The adjustment in (12) relaxes the requirement of an exact calibration of cluster effects, which can be beneficial when the outcome has relatively homogeneous cluster-specific effects, that is, the ratio σ_e^2/σ_u^2 is large. Thus, our soft-calibration estimator of θ_N is $\widehat{\theta}_w = N^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}\delta_{ij}w(\widehat{c}^{\mathrm{T}}x_{ij})y_{ij}$, where $w(\widehat{c}^{\mathrm{T}}x_{ij})$ is obtained by minimizing a given loss function subject to the soft calibration constraints (10), (11) and (12).

COROLLARY 1. *Under Assumptions 1(a), 3, other regularity conditions in Assumption S3 of the Supplementary Material, and $\gamma_n = o(n^{1/2}q^{-1/2})$, if either the outcome y_{ij} follows a linear mixed-effects model or $Q(w)$ entails a correct propensity score model, we have $n^{1/2}(\widehat{\theta}_w - \theta_N) \to N(0, V_1)$ as $n \to \infty$ and $n/N \to f \in [0, 1)$, where $V_1 = \lim_{n\to\infty} nN^{-2}\mathrm{var}_p\left\{\sum_{i=1}^{k} d_i\psi_i(c^*) \mid \mathcal{F}_N\right\}$,*

$$\psi_i(c^*) = \frac{N_i}{n_i}\sum_{j=1}^{n_i}\left\{B(c^*)x_{ij,\mathrm{SC}} + \delta_{ij}w(c_0^{*\mathrm{T}}x_{ij} + c_1^{*\mathrm{T}}z_{ij})\eta_{ij}(c^*)\right\}, \quad c^* = (c_0^{*\mathrm{T}}, c_1^{*\mathrm{T}})^{\mathrm{T}},$$

and $\eta_{ij}(c^*) = y_{ij} - B(c^*)(x_{ij}^{\mathrm{T}}, z_{ij}^{\mathrm{T}})^{\mathrm{T}}$ *with* $\mathrm{var}_p(\cdot)$ *being the variance under the clustered sampling design and* $\{B(c^*), x_{ij,\mathrm{SC}}\}$ *defined in §A.5 of the Supplementary Material.*

The results in Corollary 1 are similar to that of Theorem 1 except that V_2 under two-stage cluster sampling is negligible compared to V_1 even though n/N or some cluster sampling fractions n_i/N_i are not negligible (Shao & Steel, 1999) and thus is omitted. For variance estimation, the variance of $\widehat{\theta}_w$ can be consistently estimated as $\widehat{V}_1 = nN^{-2}\sum_{i=1}^{k}\sum_{j=1}^{k}\Omega_{i,j}\psi_i(\widehat{c})\psi_j(\widehat{c})$, where $\Omega_{i,j}$ depends on the cluster sampling scheme at the first stage, $\psi_i(\widehat{c})$ is referred as the pseudo-values with c^* replaced by $\widehat{c}$, and the consistency of $\widehat{V}_1$ can be verified by standard arguments in Kim & Rao (2009).

4. SIMULATION STUDY

In this section, we conduct a simulation study to evaluate the finite-sample performance of our proposed soft-calibration estimator and assess the robustness of its bias-corrected version in the case of cluster-specific nonignorable missingness. First, we generate samples from finite populations using the two-stage cluster sampling mechanism, in which $k = 30$ clusters with cluster sizes $n_i = 200$ are selected from $K = 2000$ clusters.

We consider two generating models for y_{ij}. One is the linear mixed-effects model: $y_{ij} = x_{ij}^{\mathrm{T}}\beta + \lambda_1 a_i + e_{ij}$ with $x_{ij} = (1, x_{1ij}, x_{2ij})^{\mathrm{T}}$ where $\beta = (0, 1, 1)^{\mathrm{T}}$, $x_{1ij} \sim U[-0.75, 0.75]$, $x_{2ij} \sim N(0, 1)$, $a_i \sim N(0, 1)$ and $e_{ij} \sim N(0, 1)$. The other one is a non-linear mixed-effects model $y_{ij} = x_{ij}^{\mathrm{T}}\beta + x_{1ij}^2 + x_{2ij}^2 + 0.1x_{3ij}^{\dagger} + 0.1x_{4ij}^{\dagger} + \lambda_1 a_i + e_{ij}$, where $x_{3ij}^{\dagger}$ and $x_{4ij}^{\dagger}$ are the standardized versions of $x_{3ij} = \exp(x_{1ij})$ and $x_{4ij} = \exp(x_{2ij})$. We consider a logistic propensity score to generate δ_{ij}: $\delta_{ij} \sim \text{Bernoulli}(p_{ij})$, where $\mathrm{logit}(p_{ij}) = x_{ij}^{\mathrm{T}}\alpha + \lambda_2 z_i$ and $\alpha = (-0.25, 1, 1)^{\mathrm{T}}$ with $\mathrm{logit}(\cdot)$ being the logit-link. For illustration, we present a set of (λ_1, λ_2) in Table 3 gauging the between-cluster variation of y_{ij} and δ_{ij}, and additional simulation studies are deferred to §B of the Supplementary Material.

From §2.4, the loss function dictates the propensity score model. For assessing the double robustness of the soft-calibration estimator, we consider two loss functions: the maximum entropy balancing function, i.e., a logistic mixed-effects model for the propensity score, and the square loss function, i.e., a linear mixed-effects model for the inverse of the propensity score. Next, we compute nine estimators for θ_N: (i) $\widehat{\theta}_{\mathrm{sim}}$ the simple average of the observed y_{ij}; (ii, iii) $\widehat{\theta}_{\mathrm{fix}}$ and $\widehat{\theta}_{\mathrm{rand}}$, where p_{ij} is estimated with fixed or random effects for clusters; (iv-vi) $\widehat{\theta}_{\mathrm{hc}}$, $\widehat{\theta}_w^{(\mathrm{SQ})}$ and $\widehat{\theta}_w^{(\mathrm{ME})}$, where w_{ij} achieves the hard calibration conditions under the maximum entropy loss function, the soft calibration conditions under the square loss function or under the maximum entropy loss function; (vii) $\widehat{\theta}_{\mathrm{bc}}$, bias-corrected $\widehat{\theta}_w^{(\mathrm{ME})}$ with $\widehat{\mu}_{ij} = x_{1ij}^{\mathrm{T}}\widehat{\beta} + x_{2ij}^{\mathrm{T}}\widehat{u}$; (viii) $\widehat{\theta}_{\mathrm{cbps}}$, the high-dimensional covariate propensity score balancing method of Ning et al. (2020); and (ix) $\widehat{\theta}_{\mathrm{rcal}}$, the high-dimensional regularized calibration method of Tan (2020).

Table 3 reports the simulation results based on 500 Monte Carlo samples. The performance of estimators is evaluated on the basis of biases, variances, mean squared errors, and coverage probabilities. Among all estimators, the simple average estimator $\widehat{\theta}_{\mathrm{sim}}$ shows large biases across all different scenarios. When the cluster factor is included as fixed or random effects, the biases of $\widehat{\theta}_{\mathrm{fix}}$ and $\widehat{\theta}_{\mathrm{rand}}$ are substantially reduced, while their variances remain large. The large variances could be attributed to their overly abundant parameters associated with the cluster indicators. When the random effects weakly affect outcomes (i.e., $\lambda_1 = 0.01$), all soft-calibration estimators outperform $\widehat{\theta}_{\mathrm{hc}}$, indicating their ability to address the issue of over-calibration. In particular, $\widehat{\theta}_w^{(\mathrm{SQ})}$ performs better than $\widehat{\theta}_w^{(\mathrm{ME})}$ under the linear mixed-effects model, which agrees with

Table 3. *Bias* $(\times 10^{-2})$*, variance* $(\times 10^{-3})$*, mean squared error* $(\times 10^{-3})$ *and coverage probability (%) of the estimators under cluster-specific nonignorable missingness based on* 500 *simulated datasets*

	$\widehat{\theta}_{\text{sim}}$	$\widehat{\theta}_{\text{fix}}$	$\widehat{\theta}_{\text{rand}}$	$\widehat{\theta}_{\text{hc}}$	$\widehat{\theta}_{w}^{(\text{SQ})}$	$\widehat{\theta}_{w}^{(\text{ME})}$	$\widehat{\theta}_{\text{bc}}$	$\widehat{\theta}_{\text{cbps}}$	$\widehat{\theta}_{\text{rcal}}$
Linear mixed-effects model with $(\lambda_1, \lambda_2) = (0.01, 1)$									
Bias	21.2	0.02	0.29	0.10	0.16	0.13	0.09	0.17	0.35
VAR	0.23	1.53	1.40	0.78	0.61	0.73	0.74	0.78	0.75
MSE	45.1	1.53	1.41	0.78	0.61	0.73	0.74	0.78	0.76
CP	0.0	94.6	94.2	92.6	93.8	93.0	93.2	–	–
Linear mixed-effects model with $(\lambda_1, \lambda_2) = (0.01, 10)$									
Bias	5.02	0.28	0.01	0.73	0.29	0.27	0.18	0.43	7.44
VAR	0.35	26.4	22.3	4.57	1.49	1.69	2.16	5.88	0.69
MSE	2.88	26.4	22.3	4.62	1.49	1.70	2.16	5.89	6.23
CP	23.8	88.6	87.8	94.2	94.4	92.4	92.2	–	–
Linear mixed-effects model with $(\lambda_1, \lambda_2) = (0.5, 1)$									
Bias	30.3	0.49	1.61	0.64	1.26	1.28	0.63	0.82	2.03
VAR	2.74	10.7	10.2	9.23	9.64	9.84	9.21	10.2	9.79
MSE	94.4	10.7	10.4	9.27	9.80	10.0	9.25	10.3	10.2
CP	0.0	95.0	93.4	94.2	94.0	93.6	94.0	–	–
Non-linear mixed-effects model with $(\lambda_1, \lambda_2) = (0.01, 1)$									
Bias	31.6	0.10	0.38	0.92	8.75	0.03	0.11	0.09	0.59
VAR	1.50	2.42	2.24	1.96	1.72	1.69	1.71	1.71	1.86
MSE	102	2.42	2.25	2.05	9.37	1.69	1.71	1.71	1.89
CP	0.0	94.0	94.0	92.6	0.0	94.4	96.6	–	–

VAR, variance; MSE, mean squared error; CP, coverage probability; We omit calculating the variance estimators for $\widehat{\theta}_{\text{cbps}}$ and $\widehat{\theta}_{\text{rcal}}$ because they are unavailable for the clustered data in their R packages.

the connection between $\widehat{\theta}_{w}^{(\text{SQ})}$ and $\widehat{\theta}_{\text{blup}}$ established in Proposition 1. However, $\widehat{\theta}_{w}^{(\text{SQ})}$ is subject to significant bias when the outcome model is misspecified, leading to an unsatisfactory coverage probability, while $\widehat{\theta}_{w}^{(\text{ME})}$ still exhibits a desirable finite-sample coverage probability, which aligns with our claim of double robustness in Theorem 1 when the propensity score is correctly specified. Although the bias-corrected estimator $\widehat{\theta}_{\text{bc}}$ has a slightly larger mean squared error than $\widehat{\theta}_{w}^{(\text{ME})}$ when $\lambda_1 = 0.01$, it performs better when the data present a larger between-cluster variation of y_{ij} (i.e., $\lambda_1 = 0.5$), which provides empirical support for the robustness of $\widehat{\theta}_{\text{bc}}$ with respect to the rate requirement for γ_n. As expected, both regularized calibration estimators $\widehat{\theta}_{\text{cbps}}$ and $\widehat{\theta}_{\text{rcal}}$ have larger mean squared errors under the linear mixed-effects model since our soft calibration conditions are motivated by linear mixed effects.

Overall, our proposed estimators tend to produce smaller mean squared errors while dealing with cluster-specific missingness, irrespective of possible model misspecification of either outcome or propensity score.

5. AN APPLICATION: EFFECT OF SCHOOL-BASED BMI SCREENING ON CHILDHOOD OBESITY

The epidemic of childhood obesity has been widely publicized (Peyer et al., 2015). Many school districts have implemented coordinated school-based body mass index screening programs to help increase parental awareness of children's body status and promote preventive strategies to reduce the risk of obesity. We use a data set collected by the Pennsylvania Department of Health to evaluate the effect of the program on the annual prevalence of overweight and obesity in elementary schools across Pennsylvania in 2007. The primary goal is to investigate the causal effect of implementing the program on reducing childhood obesity and overweight. Because the implementation of the policy was not randomized, it is essential to control pre-treatment covariates for causal analysis of the effect of the policy. Furthermore, school districts are clustered by geographic and demographic factors. Thus, soft calibration can be used to estimate the causal effect by correcting for cluster-specific confounding bias.

The dataset contains information on 493 elementary schools, which are clustered according to the type of community (rural, suburban, and urban) and the population density (low, moderate, and high). There are six clusters with sample size $n_1 = 65, n_2 = 96, n_3 = 89, n_4 = 29$, $n_5 = 104$, and $n_6 = 4$. For each school, the data consist of the treatment status A_{ij} where $A_{ij} = 1$ if the school has implemented the policy and 0 otherwise, the outcome variable y_{ij}, indicating the annual prevalence of overweight and obesity in each school, and two covariates x_{1ij} and x_{2ij}, the baseline prevalence of overweight children and the percentage of reduced and free lunch, respectively. For estimation, we consider the linear mixed-effects model and the maximum entropy loss function, including covariates x_{1ij}, x_{2ij} and the cluster intercept to model the outcome and weights for $A_{ij} = 0$ and $A_{ij} = 1$, respectively.

Table 4 reports the estimated average treatment effects on the annual prevalence of overweight and obesity along with the estimated variances and 95% confidence intervals. Without any adjustment, $\widehat{\theta}_{\text{sim}}$ shows that the policy has a significant effect in reducing the prevalence of overweight and obesity in schools, which may be subject to confounder bias. After adjusting for confounders through propensity weighting or calibration, all other estimators show that the policy may mildly reduce the prevalence of overweight. Also, $\widehat{\theta}_{\text{hc}}$, $\widehat{\theta}_w^{(\text{ME})}$ and $\widehat{\theta}_{\text{bc}}$ provide similar estimates, but the soft-calibration estimators yield a slightly smaller variance, which can be attributed to the approximate calibration condition on the cluster indicator. As discussed in §C of the Supplementary Material, the cross-fitting strategy selects two small tuning parameters as $\gamma_{n,A=0} = 0.052$ and $\gamma_{n,A=1} = 0.068$. It implies that the correction term on the right-hand side of (6b) is fairly small and a nearly exact calibration should be adopted, as demonstrated by the similarities in the calibration weights in Figure S5. Estimators $\widehat{\theta}_{\text{cbps}}$ and $\widehat{\theta}_{\text{rcal}}$ might not be credible when the sparsity condition is not met, as we have shown in the simulation studies. Based on our analysis, the policy can reduce the average prevalence of overweight and obesity in elementary schools in Pennsylvania, although the statistical evidence is not significant.

Table 4. *The estimated average treatment effects of SBMIS on the annual overweight and obesity prevalence in elementary schools across Pennsylvania*

	$\widehat{\theta}_{\text{sim}}$	$\widehat{\theta}_{\text{fix}}$	$\widehat{\theta}_{\text{rand}}$	$\widehat{\theta}_{\text{hc}}$	$\widehat{\theta}_w^{(\text{ME})}$	$\widehat{\theta}_{\text{bc}}$	$\widehat{\theta}_{\text{cbps}}$	$\widehat{\theta}_{\text{rcal}}$
ATE	8.71	0.41	0.43	0.55	0.53	0.54	0.28	0.51
VE	2258.8	467.8	474.5	448.5	445.7	446.0		
CIs	(5.77,11.66)	(-0.93,1.75)	(-0.92,1.78)	(-0.77,1.86)	(-0.78,1.84)	(-0.77,1.85)	–	–

SBMIS, school-based body mass index screening; ATE, average treatment effects; VE, variance estimation ($\times 10^{-3}$); CIs, confidence intervals

ACKNOWLEDGEMENT

This research is supported by the U.S. National Science Foundation and the U.S. National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes all technical proofs, additional simulation results, and other implementation details.

REFERENCES

ANASTASIADE, M.-C. & TILLÉ, Y. (2017). Decomposition of gender wage inequalities through calibration: Application to the swiss structure of earnings survey. *Survey Methodol.* **43**, 211–235.

ATHEY, S., IMBENS, G. W. & WAGER, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *J. R. Statist. Soc. B* **80**, 597–623.

AVAGYAN, V. & VANSTEELANDT, S. (2021). High-dimensional inference for the average treatment effect under model misspecification using penalized bias-reduced double-robust estimation. *Biostat. Epidemiol.* , 1–18.

BARDSLEY, P. & CHAMBERS, R. (1984). Multipurpose estimation from unbalanced samples. *J. R. Statist. Soc. C* **33**, 290–299.

BEN-MICHAEL, E., FELLER, A. & HARTMAN, E. (2021). Multilevel calibration weighting for survey data. *arXiv preprint arXiv:2102.09052* .

CARDOT, H. & JOSSERAND, E. (2011). Horvitz–Thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika* **98**, 107–118.

CASSEL, C. M., SÄRNDAL, C. E. & WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–620.

CHATTOPADHYAY, A., HASE, C. H. & ZUBIZARRETA, J. R. (2020). Balancing vs modeling approaches to weighting in practice. *Statist. Med.* **39**, 3227–3254.

CHAUVET, G. & GOGA, C. (2022). Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. *J. Stat. Plan. Inference* **217**, 177–187.

CHEN, S. & HAZIZA, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104**, 439–453.

DAI, L., CHEN, K., SUN, Z., LIU, Z. & LI, G. (2018). Broken adaptive ridge regression and its asymptotic properties. *J. Mult. Anal.* **168**, 334–351.

DEVAUD, D. & TILLÉ, Y. (2019). Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem. *Test* **28**, 1033–1065.

DEVILLE, J.-C. (2000). Generalized calibration and application to weighting for non-response. In *COMPSTAT*. Springer.

DEVILLE, J.-C. & SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **87**, 376–382.

ESTEVAO, V. M. & SÄRNDAL, C.-E. (2000). A functional form approach to calibration. *J. Off. Statist.* **16**, 379–399.

FOLLMANN, D. A. & WU, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151–168.

GAO, S. (2004). A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. *Statist. Med.* **23**, 211–219.

GOLUB, G. H., HEATH, M. & WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

GUGGEMOS, F. & TILLÉ, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *J. Stat. Plan. Inference* **140**, 3199–3212.

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20**, 25–46.

HAN, P. & WANG, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* **100**, 417–430.

HIRSHBERG, D. A., MALEKI, A. & ZUBIZARRETA, J. R. (2019). Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296* .

IMAI, K. & RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Statist. Soc. B* **76**, 243–263.

ISAKI, C. T. & FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Assoc.* **77**, 89–96.

KANG, J. D. & SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22**, 523–539.

KIM, J. K., KWON, Y. & PAIK, M. C. (2016). Calibrated propensity score method for survey nonresponse in cluster sampling. *Biometrika* **103**, 461–473.

KIM, J. K. & RAO, J. N. K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika* **96**, 917–932.

KOTT, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodol.* **32**, 133–142.

LAZZERONI, L. C. & LITTLE, R. J. (1998). Random-effects models for smoothing poststratification weights. *J. Off. Statist.* **14**, 61–78.

LEE, D., YANG, S. DONG, L., WANG, X., ZENG, D. & CAI, J. (2021). Improving trial generalizability using observational studies. *Biometrics* Available online (https://doi.org/10.1111/biom.13609).

LEE, D., YANG, S. & WANG, X. (2022). Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of Causal Inference* **10**, 415–440.

LEE, S. & VALLIANT, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol. Methods Res.* **37**, 319–343.

LUNDSTRÖM, S. & SÄRNDAL, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *J. Off. Statist.* **15**, 305.

NING, Y., SIDA, P. & IMAI, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika* **107**, 533–554.

PEYER, K. L., WELK, G., BAILEY-DAVIS, L., YANG, S. & KIM, J.-K. (2015). Factors associated with parent concern for child weight and parenting behaviors. *Child. Obes.* **11**, 269–274.

PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. consistency. *Ann. Statist.* **12**, 1298–1309.

READ, T. R. & CRESSIE, N. A. (2012). *Goodness-of-fit statistics for discrete multivariate data*. Springer Science & Business Media.

SÄRNDAL, C. E., SWENSSON, B. & WRETMAN, J. H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SHAO, J. & STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fraction. *J. Am. Statist. Assoc.* **94**, 254–265.

SKINNER, C. J. (1999). Calibration weighting and non-sampling errors. *Research in Off. Statist.* **2**, 33–43.

TAN, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Ann. Statist.* **48**, 811–837.

TORABI, M. & RAO, J. (2008). Small area estimation under a two-level model. *Survey Methodol.* **34**, 11–17.

VERBEKE, G. (2000). Linear mixed models for longitudinal data. In *Linear Mixed Models in Practice*. Springer, pp. 63–153.

WANG, J., WONG, R. K. W., YANG, S. & CHAN, K. C. G. (2022). Estimation of partially conditional average treatment effect by double kernel-covariate balancing. *Elect. J. Statist.* **16**, 4332 – 4378.

WEISS, R. E. (2005). *Modeling Longitudinal Data*. Springer Science & Business Media.

WONG, R. K. W. & CHAN, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika* **105**, 199–213.

WU, C. & RAO, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Can. J. Statist.* **34**, 359–375.

WU, C. & SITTER, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Statist. Assoc.* **96**, 185–193.

XIAO, Y., MOODIE, E. E. & ABRAHAMOWICZ, M. (2013). Comparison of approaches to weight truncation for marginal structural cox models. *Epidemiol. Methods* **2**, 1–20.

YANG, S. (2018). Propensity score weighting for causal inference with clustered data. *J. Causal Inference* **6**.

YANG, S. & DING, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika* **105**, 487–493.

YANG, S. & DING, P. (2019). Combining multiple observational data sources to estimate causal effects. *J. Am. Statist. Assoc.* **115**, 1540–1554.

YANG, S. & KIM, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science* **3**, 625–650.

YUAN, Y. & LITTLE, R. J. (2007). Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *J. R. Statist. Soc. C* **56**, 79–97.

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Statist. Assoc.* **110**, 910–922.