# Propensity Score Weighting for Causal Inference with Clustered Data

Shu Yang*

**Abstract**

Propensity score weighting is a tool for causal inference to adjust for measured confounders in observational studies. In practice, data often present complex structures, such as clustering, which make propensity score modeling and estimation challenging. In addition, for clustered data, there may be unmeasured cluster-level covariates that are related to both the treatment assignment and outcome. When such unmeasured cluster-specific confounders exist and are omitted in the propensity score model, the subsequent propensity score adjustment may be biased. In this article, we propose a calibration technique for propensity score estimation under the latent ignorable treatment assignment mechanism, i. e., the treatment-outcome relationship is unconfounded given the observed covariates and the latent cluster-specific confounders. We impose novel balance constraints which imply exact balance of the observed confounders and the unobserved cluster-level confounders between the treatment groups. We show that the proposed calibrated propensity score weighting estimator is doubly robust in that it is consistent for the average treatment effect if either the propensity score model is correctly specified or the outcome follows a linear mixed effects model. Moreover, the proposed weighting method can be combined with sampling weights for an integrated solution to handle confounding and sampling designs for causal inference with clustered survey data. In simulation studies, we show that the proposed estimator is superior to other competitors. We estimate the effect of School Body Mass Index Screening on prevalence of overweight and obesity for elementary schools in Pennsylvania.

*Keywords*: Calibration; Inverse probability weighting; Survey sampling; Unmeasured confounding.

*Dept. of Statistiscs, North Carolina State University, syang24@ncsu.edu

# 1 Introduction

A main statistical approach to causal inference is built on the potential outcomes framework (Rubin, 1974), in which a causal effect is defined as a comparison of the potential outcomes of the same units under different treatment levels. Observational studies are often used to infer causal effects in medical and social science studies. In observational studies, there often is confounding by indication: some covariates are predictors of both the treatment and outcome. One implication is that the covariate distributions differ between the treatment groups. Under the assumption of ignorable treatment assignment and that all confounders are observed, the causal effect of treatments can be obtained by comparing the outcomes for units from different treatment groups, adjusting for the observed confounders. Rosenbaum and Rubin (1983) further demonstrated the central role of the propensity score, and showed that adjusting for the propensity score removes confounding bias. An extensive literature thereafter proposed a number of estimators based on the propensity score, including *matching* (Rosenbaum and Rubin, 1985, Stuart, 2010, Abadie and Imbens, 2016), *weighting* (Hirano and Imbens, 2001, Bang and Robins, 2005, Cao, Tsiatis and Davidian, 2009), and *stratification* (Rosenbaum and Rubin, 1984, Yang, Imbens, Cui, Faries and Kadziola, 2016). In particular, propensity score weighting can be used to create a weighted population where the covariate distributions are balanced between the treatment groups, on average. Therefore, under some assumptions, the comparison between the weighted outcomes has a causal interpretation; see Imbens and Rubin (2015) for a textbook discussion.

Propensity score weighting has been mainly developed and applied in settings with independently and identically distributed (i.i.d.) data. However, in many research areas, data often present complex structures, such as clustering. Clustering can be formed in diverse ways. First, clusters are created by experimental design. The classical examples are given in educational and health studies, where students are nested in schools (e.g. Hong and Raudenbush, 2006) and patients are grouped in hospitals (e.g. Griswold, Localio and Mulrow, 2010). Second, clusters are induced by high-level features, such as common environmental and contextual factors, shared by individual units (e.g. Li, Zaslavsky and Landrum, 2013). For a motivating example, we examine the 2007–2010 body mass index (BMI) surveillance data from Pennsylvania Department of Health to estimate the effect of School Body Mass Index Screening (SBMIS) on the annual overweight and obesity prevalence in elementary schools in Pennsylvania. The data set includes 493 schools (units) in Pennsylvania, which are clustered by two factors: type of community (rural, suburban, and urban), and population density (low, median, and high). The data structure is schools nested within high-level of environments.

In this article, we address the problem of estimating the average treatment

effect from clustered data, where the treatment is administered at the unit level, the covariates are measured at both unit and cluster levels (where cluster-level covariates are cluster characteristics shared by all units within clusters), and finally the outcome is measured at the unit level. Even if we collect a rich set of unit-level covariates, there may be unobserved cluster-level covariates that are related to both the treatment assignment and outcome. This problem is ubiquitous in clustered data where data are collected sufficiently at the unit level, however insufficient information is available at the cluster level. In our motivating example, we have unit-level school covariates including the baseline prevalence of overweight and obesity and percentage of reduced and free lunch. However, certain key contextual factors, such as accessibility to and quality of care, socioeconomic and environmental variables, which can be very different across clusters, are preceivably important factors for schools implementing prevention screening strategy and children's obesity rate. Unfortunately, these cluster-specific confounders are not available. When such unmeasured confounders exist and are omitted in the propensity score model, the subsequent analysis may be biased.

We make the stable unit and treatment version assumption (SUTVA; Rubin, 1978). Under the SUTVA, potential outcomes for each unit are not affected by the treatments assigned to other units. This assumption is not uncommon. In our application, the treatment was implemented school-wise. The potential outcomes for one school are likely to be unaffected by the treatments implemented at other schools, and therefore the SUTVA is plausible. However, in other settings, this assumption may not hold. A classical example is given in infectious diseases (Ross, 1916, Hudgens and Halloran, 2008), where whether one person becomes infected depends on who else in the population is vaccinated. In this article, we will not discuss the case when the SUTVA is violated.

The literature has proposed different methods for clustered data. Oakes (2004) and VanderWeele (2008) used multi-level models for the potential outcomes to draw causal conclusions in neighborhood effect studies. A series of papers has proposed various propensity score matching algorithms with multi-level models for the propensity score; see, e.g., Hong and Raudenbush (2006), Hong and Yu (2007, 2008), Kim and Seltzer (2007), Kelcey (2009), Griswold et al. (2010), Arpino and Mealli (2011), Thoemmes and West (2011), and Kim and Steiner (2015). Recently, Li et al. (2013), Leite, Jimenez, Kaya, Stapleton, MacInnes and Sandbach (2015) and Schuler, Chu and Coffman (2016) examined propensity score weighting methods to reduce selection bias in multi-level observational studies. Xiang and Tarasawa (2015) employed propensity score stratification and multi-level models to balance key covariates between treatment groups of a cross-state sample of students. For comparison of the effectiveness of various propensity score strategies; see, e.g., Su and Cortina (2009), Griswold et al. (2010) and Eckardt (2012). Among these

works, researchers considered different modeling choices for the propensity score and outcome, such as generalized linear fixed/mixed effects models. The fixed effects models create dummy variables for each cluster, regarding the cluster variables as fixed; while the random effects models use random intercepts for each cluster, treating the cluster variables as random. Nonetheless, all existing methods require correct specification of the propensity score and outcome models.

The goal of this article is to develop a novel propensity score weighting method for causal inference with clustered data in the presence of unmeasured cluster-level confounders. An important contribution is to provide a robust construction of inverse propensity score weights under the latent ignorable treatment assignment mechanism; i.e., the treatment-outcome relationship is unconfounded given the observed confounders and the latent cluster-level confounders. The key insight is based on the central role of the propensity score in balancing the covariate distributions between the treatment groups. For propensity score estimation, we then adopt the calibration technique and impose balance constraints for moments of the observed and latent cluster-level confounders between the treatment groups. Because the latent cluster-level confounders are not observed, we impose stronger balance constraints enforcing the sum of weighted treatments equal the cluster size for all clusters, which imply the exact balance of the cluster-level confounders. The proposed propensity score weighting estimator is *doubly robust* (e.g., Robins, Rotnitzky and Zhao, 1994, Lunceford and Davidian, 2004, Bang and Robins, 2005, Kang and Schafer, 2007) in the sense that it is consistent for the average treatment effect if either the propensity score model is correctly specified or the outcome follows a linear mixed effect model. In general cases, if the conditional mean of the outcome given the observed confounders and the latent cluster-level confounders can be well approximated by the power series of confounders, imposing constraints on these power series can also eliminate confounding biases, and the propensity score weighting estimator is consistent. The simulation results demonstrate that the proposed estimator has improved robustness to model misspecification compared to existing methods.

Importantly, our results are in agreement with some existing findings that misspecification of the propensity score model may have minor impact on the bias of the estimator for the average treatment effect (Rubin, 2004). This is especially true in the matching and stratification algorithm, because the estimated propensity score is only used to balance the covariate distributions instead of directly in estimation. Our results suggest that if both individual and cluster-level confounders achieve a good balance between the treatment groups, the proposed weighting estimator for the average treatment effect is robust.

Clustered data often arise in survey sampling. In complex surveys, the challenge is to take design information or design weights into account when developing

propensity score methods for causal inference. The proposed weighting method can be combined with sampling weights for an integrated solution to handle confounding and sampling designs for causal inference with clustered survey data.

This article is organized as follows. Section 2 introduces the data structure and assumptions, defines the estimands, and presents existing inverse probability of treatment weighting estimators for clustered data. Section 3 proposes our estimators. Section 4 presents the main theoretical results. Section 5 extends the proposed calibration estimator to clustered survey data. Section 6 reports a simulation study to evaluate finite sample properties of our estimator. Section 7 applies our methods to investigate the effect of SBMIS on the annual overweight and obesity prevalence in elementary schools in Pennsylvania. A concluding remark is given in Section 8. Finally, proofs of the main theoretical results and additional simulation results are provided in the Appendix.

# 2 Basic setup

## 2.1 Observed data structure

To fix the ideas, we first focus on two-level clustered data. The extension to clustered survey data will be addressed in Section 5.

Suppose we have a two-level data structure where at the first level we have $m$ clusters, and at the second level each cluster $i$ includes $n_i$ units. Denote the sample size by $n = \sum_{i=1}^{m} n_i$. For unit $j$ in cluster $i$, we observe a $p$-dimensional vector of pre-treatment covariates $X_{ij}$, which may include the observed individual and cluster characteristics, a binary treatment variable $A_{ij} \in \{0,1\}$, with 0 and 1 being the labels of control and active treatments, respectively, and lastly an outcome variable $Y_{ij}$.

## 2.2 Potential outcomes and assumptions

We use the potential outcomes framework (Rubin, 1974). Assume that each unit has two potential outcomes: $Y_{ij}(0)$, the outcome that would be realized, possibly contrary to the fact, had the unit received the control treatment, and $Y_{ij}(1)$, the outcome that would be realized, possibly contrary to the fact, had the unit received the active treatment. The observed outcome is $Y_{ij} = Y_{ij}(A_{ij})$. This notation implicitly makes the SUTVA (Rubin, 1978) that there is no interference between units and no versions of each treatment.

Suppose that clusters are random draws from a super-population of clusters, and that for observations within cluster $i$, $\{A_{ij}, X_{ij}, Y_{ij}(0), Y_{ij}(1) : j = 1, \ldots, n_i\}$ i.i.d. follow a cluster-specific super-population model. Our goal is to estimate the average

treatment effect, $\tau = E[n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{Y_{ij}(1) - Y_{ij}(0)\}]$, where the expectation is taken with respect to the super-population model $\xi$ of all random variables, which will be specified later. In the binary case, $\tau$ is called the causal risk difference.

The fundamental problem is that not all potential outcomes can be observed for each unit in the sample; only one potential outcome, the outcome corresponding to the treatment the unit actually followed, can be observed (Holland, 1986). Throughout, we use $Z_1 \perp Z_2 \mid Z_3$ to denote the conditional independence of $Z_1$ and $Z_2$ given $Z_3$ (Dawid, 1979). With unstructured i.i.d. data, Rubin (1974) described the following assumption for identifying the average treatment effect.

**Assumption 1 (Ignorability)** $\{Y_{ij}(0), Y_{ij}(1)\} \perp A_{ij} \mid X_{ij}$.

Assumption 1 indicates that all confounders are included in $X_{ij}$.

For clustered data, confounding may vary across clusters and are related to some cluster-level covariates that are not always observable. In these cases, Assumption 1 does not hold. Instead, we assume a cluster-specific latent variable $U_i$ that summarizes the effect of unobserved cluster-level confounders, and consider the following modified ignorability assumption.

**Assumption 2 (Latent ignorability)** $\{Y_{ij}(0), Y_{ij}(1)\} \perp A_{ij} \mid X_{ij}, U_i$.

Under Assumption 2, the propensity score becomes $\text{pr}\{A_{ij} = 1 \mid X_{ij}, U_i, Y_{ij}(0), Y_{ij}(1)\} = \text{pr}(A_{ij} = 1 \mid X_{ij}, U_i) \equiv e(X_{ij}, U_i)$. Moreover, we make the standard positivity assumption for the propensity score.

**Assumption 3 (Positivity)** *There exist constants $\underline{e}$ and $\bar{e}$ such that, with probability 1, $0 < \underline{e} < e(X_{ij}, U_i) < \bar{e} < 1$.*

**Remark 1** *In our setting, the treatment is assigned at the unit level. Assumption 3 implies that each unit in each cluster has a positive probability to receive either treatment or control. Therefore, our setting does not apply to the settings with a cluster-level treatment where all units in one cluster receive one treatment level. For these settings, we refer the interested readers to Stuart (2007) and VanderWeele (2008).*

Under Assumption 2, we write the joint distribution of $\{(A_{ij}, X_{ij}, U_i, Y_{ij}) : i = 1, \ldots, m; j = 1, \ldots n_i\}$ as

$$\prod_{i=1}^{m} f(U_i) \prod_{j=1}^{n_i} \Big( f(X_{ij} \mid U_i) \left\{ f_1(Y_{ij} \mid X_{ij}, U_i) e(X_{ij}, U_i) \right\}^{A_{ij}} \times \left[ f_0(Y_{ij} \mid X_{ij}, U_i) \{1 - e(X_{ij}, U_i)\} \right]^{1 - A_{ij}} \Big),$$

where $f_a(\cdot \mid X_{ij}, U_i)$ is a conditional distribution of $Y_{ij}(a)$ given $(X_{ij}, U_i)$, for $a = 0, 1$.

The literature has considered generalized linear mixed effects models for $f_a(Y_{ij} \mid X_{ij}, U_i)$ and $e(X_{ij}, U_i)$; see, e.g., Arpino and Mealli (2011), Thoemmes and West (2011), Li et al. (2013). Following the literature, we assume generalized linear mixed effects models for the outcome and propensity score.

**Assumption 4 (Outcome model)** *The potential outcome $Y_{ij}(a)$ follows a generalized linear mixed effects model with a random intercept $U_i$ as*

$$\mu_{ij}(a) = g_a(X_{ij}^{\mathrm{T}} \beta_a + U_i), \tag{1}$$

*where $\mu_{ij}(a) = E\{Y_{ij}(a) \mid X_{ij}, U_i\}$, $g_a(\cdot)$ is an unspecified inverse link function, and $\beta_a$ is a $p$-dimensional vector.*

**Assumption 5 (Propensity score model)** *The actual treatment $A_{ij}$ given $(X_{ij}, U_i)$ follows a generalized linear mixed effects model with a random intercept $U_i$ as*

$$e(X_{ij}, U_i; \eta) = h(X_{ij}^{\mathrm{T}} \eta + U_i), \tag{2}$$

*where $h(\cdot)$ is an unspecified inverse link function, and $\eta$ is a $q$-dimensional vector of parameters.*

There are two different model specifications regarding the cluster-level confounders. The fixed effects model treats $U_i$ as fixed but unknown parameters across clusters. In this fixed-effects approach, treatment assignment is an ignorable process, which complies with Assumption 1 given that $X_{ij}$ stacks all observed confounders and cluster dummy variables. On the other hand, the random effects model treats $U_i$ as random and i.i.d. drawn from a distribution. The difference between the two modeling strategies has been addressed in both statistics and econometrics literature; see, e.g., Baltagi (1995) and Wooldridge (2002). Briefly, there are both statistical and logical considerations. First, if the number of clusters is relatively large, the parameter estimates in the fixed effects model are inconsistent (Wallace and Hussain, 1969). In this case, the random effects approach is preferred. Second, the fixed effects approach does not make distributional assumptions of the cluster-level confounders; whereas, the random effects approach assumes that $U_i$ is random and i.i.d. drawn from a distribution.

## 2.3 Inverse probability of treatment weighting estimator

To estimate the average treatment effect $\tau$, let $\nu = (U_1, \ldots, U_m)$ denote the vector of cluster-level confounders. Under (2), the inverse propensity score or probability of treatment weighting (IPTW) estimator of $\tau$ can be expressed as

$$\hat{\tau}_{\text{IPTW}}(\nu, \eta) = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ \frac{A_{ij} Y_{ij}}{e(X_{ij}, U_i; \eta)} - \frac{(1 - A_{ij}) Y_{ij}}{1 - e(X_{ij}, U_i; \eta)} \right\}. \tag{3}$$

Under Assumptions 2 and 3, if the propensity score is known, it is straightforward to verify that $\hat{\tau}_{\text{IPTW}}(\nu, \eta)$ is unbiased for $\tau$. Moreover, if it is unknown but depends only on fixed parameters, $\hat{\tau}_{\text{IPTW}}(\nu, \eta)$ with the consistently estimated propensity score is asymptotically unbiased for $\tau$. The challenge with clustered data is that $\hat{\tau}_{\text{IPTW}}(\nu, \eta)$ may depend on a growing number of unobserved cluster-level confounders. To resolve this issue, there are several options:

**(i)** Weight based on predicted random intercepts; i.e., treat the $U_i$'s in model (2) as random intercepts, and predict the propensity score as $e(X_{ij}, \hat{U}_i^{\text{ran}}; \hat{\eta}^{\text{ran}})$, where $\hat{U}_i^{\text{ran}}$ is the mode of a predictive distribution for $U_i$ given the observed $A_{ij}$ and $X_{ij}$, and $\hat{\eta}^{\text{ran}}$ is the maximum likelihood estimator of $\eta$.

**(ii)** Weight based on estimated fixed intercepts; i.e., treat the $U_i$'s in model (2) as fixed intercepts, and estimate the propensity score as $e(X_{ij}, \hat{U}_i^{\text{fix}}; \hat{\eta}^{\text{fix}})$, where $\hat{U}_i^{\text{fix}}$ and $\hat{\eta}^{\text{fix}}$ are maximum likelihood estimators.

Let $\hat{\tau}_{\text{IPTW}}(\nu, \eta)$ in (3) be denoted as $\hat{\tau}_{\text{ran}}$ or $\hat{\tau}_{\text{fix}}$ when the propensity score is predicted under option (i) or estimated under option (ii), respectively. The two approaches suffer several drawbacks. First, to obtain $\hat{\tau}_{\text{ran}}$ often involves numerical integration, which can be computationally heavy. Second, the predicted value of the propensity score does not guarantee the balance of covariate distributions between the treatment groups, due to the shrinkage of random intercepts toward zero. Lastly, it is well-known that under (2), $\hat{\tau}^{\text{fix}}$ is not consistent as $m$ increases, because when treating $U_i$ as fixed, the number of parameters has an order similar to $m$ (Skinner et al., 2011).

# 3 Proposed methodology

To motivate our estimation of the propensity score, we note

$$E\left\{ \frac{A}{e(X,U)} \begin{pmatrix} X \\ U \end{pmatrix} \right\} = E\left[ E\left\{ \frac{A}{e(X,U)} \mid X, U \right\} \begin{pmatrix} X \\ U \end{pmatrix} \right] = E\left\{ \begin{pmatrix} X \\ U \end{pmatrix} \right\}, \tag{4}$$

and

$$E\left\{\frac{1-A}{1-e(X,U)}\begin{pmatrix} X \\ U \end{pmatrix}\right\} = E\left[E\left\{\frac{1-A}{1-e(X,U)} \mid X,U\right\}\begin{pmatrix} X \\ U \end{pmatrix}\right] = E\left\{\begin{pmatrix} X \\ U \end{pmatrix}\right\}. \quad (5)$$

Equations (4) and (5) clarify the central role of the propensity score in balancing the covariate distributions between the treatment groups in the super-population. For simplicity of exposition, let $e_{ij}$ be the propensity score for unit $j$ in cluster $i$, and let $\hat{e}_{ij}$ be the corresponding estimate. We consider the propensity score estimate to satisfy the following constraints:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{A_{ij}}{\hat{e}_{ij}}X_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1-A_{ij}}{1-\hat{e}_{ij}}X_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}X_{ij}, \quad (6)$$

$$\sum_{j=1}^{n_i}\frac{A_{ij}}{\hat{e}_{ij}} = \sum_{j=1}^{n_i}\frac{1-A_{ij}}{1-\hat{e}_{ij}} = \sum_{j=1}^{n_i}1 = n_i, \quad (i=1,\ldots,m). \quad (7)$$

Note that (6) is the empirical version of (4). The empirical version of (5) is

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{A_{ij}}{\hat{e}_{ij}}U_i = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1-A_{ij}}{1-\hat{e}_{ij}}U_i = \sum_{i=1}^{m}\sum_{j=1}^{n_i}U_i, \quad (8)$$

which however is infeasible because the $U_i$'s are unobserved. Instead, we impose (7) which implies (8), without the need to observe the $U_i$'s.

To obtain the propensity score estimate that achieves (6) and (7), we use the calibration technique in the following steps:

**Step 1.** Obtain an initial propensity score estimate $\hat{e}^0_{ij}$ under a working propensity score model, e.g. a logistic linear mixed effects model. This in turn provides an initial set of inverse propensity score weights, $\mathbb{W}^0 = \{d_{ij} : i = 1,\ldots,m; j = 1,\ldots,n_i\}$, where $d_{ij} = 1/e^0_{ij}$ if $A_{ij} = 1$ and $d_{ij} = 1/(1-e^0_{ij})$ if $A_{ij} = 0$.

**Step 2.** Modify the initial set of weights $\mathbb{W}^0$ to a new set of weights $\mathbb{W} = \{\alpha_{ij} : i = 1,\ldots,m; j = 1,\ldots,n_i\}$ by minimizing the Kullback-Leibler distance (Kullback and Leibler, 1951) of $\mathbb{W}^0$ and $\mathbb{W}$:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}G(\alpha_{ij},d_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}\log\frac{\alpha_{ij}}{d_{ij}}, \quad (9)$$

subject to (6) and (7). By the Lagrange Multipliers technique, the minimizer of (9) subject to (6) and (7) is

$$\alpha_{ij}(\lambda_1,\lambda_2) = \frac{n_i A_{ij} d_{ij}\exp(\lambda_1^{\mathrm{T}}X_{ij}A_{ij})}{\sum_{j=1}^{n_i}A_{ij}d_{ij}\exp(\lambda_1^{\mathrm{T}}X_{ij}A_{ij})} + \frac{n_i(1-A_{ij})d_{ij}\exp\{\lambda_2^{\mathrm{T}}X_{ij}(1-A_{ij})\}}{\sum_{j=1}^{n_i}(1-A_{ij})d_{ij}\exp\{\lambda_2^{\mathrm{T}}X_{ij}(1-A_{ij})\}}, \quad (10)$$

where $(\lambda_1, \lambda_2)$ is the solution to the following equation

$$\hat{Q}(\lambda_1, \lambda_2) = \begin{pmatrix} \hat{Q}_1(\lambda_1, \lambda_2) \\ \hat{Q}_2(\lambda_1, \lambda_2) \end{pmatrix} = \begin{pmatrix} n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ A_{ij} \alpha_{ij}(\lambda_1, \lambda_2) - 1 \right\} X_{ij} \\ n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ (1 - A_{ij}) \alpha_{ij}(\lambda_1, \lambda_2) - 1 \right\} X_{ij} \end{pmatrix} = 0. \quad (11)$$

**Step 3.** Obtain the propensity score estimate as

$$\hat{e}_{ij} = \alpha_{ij}(\hat{\lambda}_1, \hat{\lambda}_2)^{-A_{ij}} \{1 - \alpha_{ij}(\hat{\lambda}_1, \hat{\lambda}_2)\}^{-1+A_{ij}}.$$

Finally, our proposed IPTW estimator is

$$\hat{\tau}_{\text{IPTW}} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ \frac{A_{ij} Y_{ij}}{\hat{e}_{ij}} - \frac{(1 - A_{ij}) Y_{ij}}{1 - \hat{e}_{ij}} \right\}. \quad (12)$$

**Remark 2 (Calibration)** *Calibration has been used in many scenarios. In survey sampling, calibration is widely used to integrate auxiliary information in estimation or handle nonresponse; see, e.g., Wu and Sitter (2001), Chen, Sitter and Wu (2002), Särndal and Lundström (2005), Kott (2006), Chang and Kott (2008) and Kim, Kwon and Paik (2016). In causal inference, calibration has been used such as in Constrained Empirical Likelihood (Qin and Zhang, 2007), Entropy Balancing (Hainmueller, 2012), Inverse Probability Tilting (Graham, Pinto and Egel, 2012), and Covariate Balance Propensity Score of Imai and Ratkovic (2014). Chan, Yam and Zhang (2015) showed that estimation of average treatment effects by empirical balance calibration weighting can achieve global efficiency. However, all these works were developed in settings with i.i.d. variables and they assumed that there are no unmeasured confounders. Our article is the first to use calibration for causal inference with unmeasured cluster-level confounders.*

**Remark 3 (Distance function)** *In Step 2 of the calibration algorithm, different distance functions, other than the Kullback-Leibler distance, can be considered. For example, if we choose $G(\alpha_{ij}, d_{ij}) = d_{ij}(\alpha_{ij}/d_{ij} - 1)^2$, then the minimum distance estimation leads to generalized regression estimation (Park and Fuller, 2012) of the $\alpha_{ij}$'s. If we choose $G(\alpha_{ij}, d_{ij}) = -d_{ij} \log(\alpha_{ij}/d_{ij})$, then it leads to empirical likelihood estimation (Newey and Smith, 2004). We use the Kullback–Leibler distance function, which leads to exponential tilting estimation (Kitamura and Stutzer, 1997, Imbens, Johnson and Spady, 1998, Schennach, 2007). The advantage of using the Kullback-Leibler distance is that the resulting weights are always non-negative.*

*Also, with Kullback-Leibler distance, the calibration constraint (7) can be built into a closed form expression for the weights, and thus avoiding solving a large number of equations. This reduces the computation burden greatly, when there is a large number of clusters.*

**Remark 4 (Nonparametric methods)** *It is worth commenting on the existing robust nonparametric methods and the advantages of our estimator. In the i.i.d. data setting, many nonparametric and machine learning methods have been proposed to capture the complex relationship of different variables without parametric assumptions, such as generalized boosted regression, causal trees, random forest, and neural networks. Indeed, many studies have shown the superiority of these methods to the parametric propensity score estimation; see, e.g., McCaffrey, Ridgeway and Morral (2004), Setoguchi, Schneeweiss, Brookhart, Glynn and Cook (2008), Lee, Lessler and Stuart (2010), Pirracchio, Petersen and van der Laan (2014). However, these data-driven methods assume that all confounders are observed, and therefore they can not handle unobserved cluster-level confounders, unlike our proposed method.*

# 4 Main results

To discuss the asymptotic properties of the proposed estimator, we assume that the cluster sample sizes satisfy the condition that $\min_{1\le i\le m} n_i \to \infty$ and $\sup_{1\le i\le m} n_i = O(n^{1/2})$. To show the double robustness of the proposed estimator $\hat{\tau}_{\text{IPTW}}$, we distinguish two cases and indicate different roles of calibration in estimation. We provide a heuristic argument below and relegate the technical details to the Appendix.

First, consider the case when the initial propensity score model is correctly specified. The weighting estimator with the initial propensity score estimates is then consistent for the average treatment effect. In this case, the role of calibration is to improve estimation efficiency by incorporating additional covariate information. This role of calibration has been demonstrated extensively in the survey literature (e.g., Deville and Särndal 1992) to modify the initial design weights to incorporate known auxiliary information.

Second, consider the case when the outcome models are linear mixed effects models:

$$E\left\{Y_{ij}(a) \mid X_{ij}, U_i\right\} = X_{ij}^{\mathrm{T}}\beta_a + U_i + e_{a,ij},$$

where the $e_{a,ij}$'s are independent with $E(e_{a,ij} \mid X_{ij}, U_i) = 0$, for $a = 0, 1$. In this case, the role of calibration is to balance the confounders between the treatment groups for reducing the selection bias. We note that $\hat{e}_{ij}$ does not depend on outcome

variables, and therefore under Assumptions 2, 4 and 5, $\hat{e}_{ij} \perp Y_{ij}(1) \mid X_{ij}, U_i$. Then, we have

$$E\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\frac{A_{ij}}{\hat{e}_{ij}}-1\right)Y_{ij}(1)\right\} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}E\left[\left(\frac{A_{ij}}{\hat{e}_{ij}}-1\right)E\left\{Y_{ij}(1)\mid X_{ij},U_i\right\}\right]$$
$$= E\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\frac{A_{ij}}{\hat{e}_{ij}}-1\right)\left(X_{ij}^{\mathrm{T}}\beta_1+U_i\right)\right\}=0, \quad (13)$$

where the last equality follows by the constraints (6) and (7). Using Assumption 2 and (13), it follows

$$E\left(\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{A_{ij}}{\hat{e}_{ij}}Y_{ij}\right) = E\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{A_{ij}}{\hat{e}_{ij}}Y_{ij}(1)\right\} = E\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}Y_{ij}(1)\right\}. \quad (14)$$

Similarly, we establish

$$E\left(\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1-A_{ij}}{1-\hat{e}_{ij}}Y_{ij}\right) = E\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}Y_{ij}(0)\right\}. \quad (15)$$

Combining (14) and (15), we have $E(\hat{\tau}_{\mathrm{IPTW}}) = \tau$, which yields the unbiasedness of $\hat{\tau}_{\mathrm{IPTW}}$. Under standard regularity conditions specified in the Appendix, we show that $\hat{\tau}_{\mathrm{IPTW}}$ is consistent for $\tau$.

In general cases, if the conditional mean of the outcome given the observed confounders and the latent cluster-level confounders can be well approximated by the power series of confounders, imposing constraints on these power series can also eliminate confounding biases, and the propensity score weighting estimator is consistent.

We derive the asymptotic distribution of $\hat{\tau}_{\mathrm{IPTW}}$ in the following theorem and relegate the proof to the Appendix.

**Theorem 1** *Suppose that Assumptions 2, 3, and the regularity conditions specified in the Appendix hold. Suppose further that the cluster sample sizes $n_i$, for $i = 1,\ldots,m$, satisfy the condition that $\min_{1\le i\le m} n_i \to \infty$ and $\sup_{1\le i\le m} n_i = O(n^{1/2})$. If the outcome model (1) is a linear mixed effects model or the propensity score model (2) is correctly specified, the proposed propensity score weighting estimator in (12), subject to constraints (6) and (7), satisfies*

$$V_1^{-1/2}(\hat{\tau}_{\mathrm{IPTW}} - \tau) \to \mathcal{N}(0,1),$$

*in distribution, as* $n \to \infty$, *where* $V_1 = \mathrm{var}\left(n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tau_{ij}\right)$,

$$\tau_{ij} = \{\alpha_{ij}(\lambda_1^*,\lambda_2^*)A_{ij}(Y_{ij}-B_1^{\mathrm{T}}X_{ij})+B_1^{\mathrm{T}}X_{ij}\} - \{\alpha_{ij}(\lambda_1^*,\lambda_2^*)(1-A_{ij})(Y_{ij}-B_2^{\mathrm{T}}X_{ij})+B_2^{\mathrm{T}}X_{ij}\},$$

$$B_1 = E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}A_{ij}Y_{ij}X_{ij}^{\mathrm{T}}\right] \times E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}A_{ij}X_{ij}X_{ij}^{\mathrm{T}}\right]^{-1},$$

$$B_2 = E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}(1-A_{ij})Y_{ij}X_{ij}^{\mathrm{T}}\right] \times E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}(1-A_{ij})X_{ij}X_{ij}^{\mathrm{T}}\right]^{-1},$$

*and* $(\lambda_1^*,\lambda_2^*)$ *satisfies* $E\{\hat{Q}(\lambda_1^*,\lambda_2^*)\}=0$ *with* $\hat{Q}(\lambda_1,\lambda_2)$ *defined in (11).*

The asymptotic result in Theorem 1 allows for variance estimation of $\hat{\tau}_{\mathrm{IPTW}}$. We now discuss variance estimation. Let $\hat{\tau}_{ij} = \alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\{A_{ij}(Y_{ij}-\hat{B}_1^{\mathrm{T}}X_{ij})-(1-A_{ij})(Y_{ij}-\hat{B}_2^{\mathrm{T}}X_{ij})\}+(\hat{B}_1-\hat{B}_2)^{\mathrm{T}}X_{ij}$, where

$$\hat{B}_1 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\left\{1-\frac{\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)}{n_i}\right\}A_{ij}Y_{ij}X_{ij}^{\mathrm{T}} \times\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\left\{1-\frac{\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)}{n_i}\right\}A_{ij}X_{ij}X_{ij}^{\mathrm{T}}\right]^{-1},$$

$$\hat{B}_2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\left\{1-\frac{\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)}{n_i}\right\}(1-A_{ij})Y_{ij}X_{ij}^{\mathrm{T}} \times\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\left\{1-\frac{\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)}{n_i}\right\}(1-A_{ij})X_{ij}X_{ij}^{\mathrm{T}}\right]^{-1}.$$

Let $\hat{\tau}_i = n_i^{-1}\sum_{j=1}^{n_i}\hat{\tau}_{ij}$ and $\hat{V}_i = (n_i-1)^{-1}\sum_{j=1}^{n_i}(\hat{\tau}_{ij}-\hat{\tau}_i)^2$. The variance estimator can be constructed as

$$\hat{V}(\hat{\tau}_{\mathrm{IPTW}}) = \frac{1}{n}\left\{\frac{1}{m-1}\sum_{i=1}^{m}(\hat{\tau}_i-\hat{\tau}_{\mathrm{IPTW}})^2+\frac{1}{m}\sum_{i=1}^{m}\hat{V}_i\right\}.$$

# 5 Extension to clustered survey data

In this section, we extend the proposed propensity score weighting estimator to clustered survey data. Consider a finite population $\mathscr{F}_N$ with $M$ clusters and $N_i$ units in the $i$th cluster, where $N = \sum_{i=1}^{M} N_i$ denotes the population size. We assume that in the finite population, $\{A_{ij}, X_{ij}, Y_{ij}(0), Y_{ij}(1) : i = 1, \ldots, M; j = 1, \ldots, N_i\}$ follows the super-population model $\xi$ as described in Section 2. We are interested in estimating the population average treatment effect $\tau = E[N^{-1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \{Y_{ij}(1) - Y_{ij}(0)\}]$.

We assume that the sample is selected according to a two-stage cluster sampling design. Specifically, at the first stage, cluster $i$ is sampled with the first-order inclusion probability $\pi_i = \text{pr}(i \in S_I)$, where $S_I$ is the index set for the sampled clusters. Let $\pi_{ij} = \text{pr}\{(i, j) \in S_I\}$ be the second-order inclusion probability for clusters $i$ and $j$ being sampled. At the second stage, given that cluster $i$ was selected at the first stage, unit $j$ is sampled with conditional probability $\pi_{j|i} = \text{pr}(j \in S_{II} \mid i \in S_I)$, where $S_{II}$ is the index set for the sampled units. The final sample size is $n = \sum_{i \in S_I} n_i$. The design weight for unit $j$ in cluster $i$ be $\omega_{ij} = (\pi_i \pi_{j|i})^{-1}$, which reflects the number of units for cluster $i$ in the finite population this unit $j$ represents. We assume that the design weights are positive and known throughout the sample. Also, let $\pi_{kl|i} = \text{pr}\{(k, l) \in S_{II} \mid i \in S_I\}$ be the second-order inclusion probability for units $k$ and $l$ being sampled given that cluster $i$ was selected. The second-order inclusion probabilities, $\pi_{ij}$ and $\pi_{kl|i}$, are often used for variance estimation.

For clustered survey data, if the propensity score $e(X_{ij}, U_i)$ is known, we can express the IPTW estimator of $\tau$ as

$$\hat{\tau}_{\text{IPTW}} = \frac{1}{N} \sum_{i \in S_I} \sum_{j=1}^{n_i} \omega_{ij} \left\{ \frac{A_{ij} Y_{ij}}{e(X_{ij}, U_i)} - \frac{(1 - A_{ij}) Y_{ij}}{1 - e(X_{ij}, U_i)} \right\}. \tag{16}$$

Let $E_\xi$ and $E_p$ denote expectation under the super-population model and the sampling design, respectively. It is easy to verify that

$$\begin{aligned} E(\hat{\tau}_{\text{IPTW}}) &= E_\xi \{E_p(\hat{\tau}_{\text{IPTW}})\} \\ &= E_\xi \left[ \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left\{ \frac{A_{ij} Y_{ij}}{e(X_{ij}, U_i)} - \frac{(1 - A_{ij}) Y_{ij}}{1 - e(X_{ij}, U_i)} \right\} \right] = \tau. \end{aligned}$$

In practice, because the propensity score $e(X_{ij}, U_i)$ is often unknown, (16) is not feasible. To estimate the propensity score, we now require the propensity score

estimate $\hat{e}_{ij}$ satisfy the following design-weighted moment constraints

$$\sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}\frac{A_{ij}}{\hat{e}_{ij}}X_{ij} = \sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}\frac{1-A_{ij}}{1-\hat{e}_{ij}}X_{ij} = \sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}X_{ij}, \tag{17}$$

$$\sum_{j=1}^{n_i}\omega_{ij}\frac{A_{ij}}{\hat{e}_{ij}} = \sum_{j=1}^{n_i}\omega_{ij}\frac{1-A_{ij}}{1-\hat{e}_{ij}} = N_i, \quad (i\in S_I). \tag{18}$$

These moment constraints (17) and (18) are the sample version of (4) and (5), respectively.

To obtain the propensity score estimate that achieves (17) and (18), we use the calibration technique in the following steps:

**Step 1.** Obtain an initial propensity score estimate $\hat{e}^0_{ij}$ under some working propensity score model, e.g. a logistic linear mixed effect model, each unit weighted by the design weight $\omega_{ij}$. This in turn provides an initial set of inverse propensity score weights, $\mathbb{W}^0 = \{d_{ij} : i = 1,\ldots,m; j = 1,\ldots,n_i\}$, where $d_{ij} = 1/e^0_{ij}$ if $A_{ij} = 1$ and $d_{ij} = 1/(1-e^0_{ij})$ if $A_{ij} = 0$.

**Step 2.** Modify the initial set of weights $\mathbb{W}^0$ to a new set of weights $\mathbb{W} = \{\alpha_{ij} : i = 1,\ldots,m; j = 1,\ldots,n_i\}$ by minimizing $\sum_{i=1}^m\sum_{j=1}^{n_i}\omega_{ij}\alpha_{ij}\log(\alpha_{ij}/d_{ij})$, subject to (17) and (18). By Lagrange Multiplier, $\alpha_{ij}$ can be obtained as

$$\alpha_{ij}(\lambda_1,\lambda_2) = \frac{N_iA_{ij}d_{ij}\exp(\lambda_1X_{ij}A_{ij})}{\sum_{j=1}^{n_i}\omega_{ij}A_{ij}d_{ij}\exp(\lambda_1X_{ij}A_{ij})} + \frac{N_i(1-A_{ij})d_{ij}\exp\{\lambda_2X_{ij}(1-A_{ij})\}}{\sum_{j=1}^{n_i}\omega_{ij}(1-A_{ij})d_{ij}\exp\{\lambda_2X_{ij}(1-A_{ij})\}}, \tag{19}$$

where $(\lambda_1,\lambda_2)$ is the solution to the following equation

$$\hat{Q}(\lambda_1,\lambda_2) = \begin{pmatrix}\hat{Q}_1(\lambda_1,\lambda_2)\\ \hat{Q}_2(\lambda_1,\lambda_2)\end{pmatrix} = \begin{pmatrix} N^{-1}\sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}\left\{A_{ij}\alpha_{ij}(\lambda_1,\lambda_2)-1\right\}X_{ij}\\ N^{-1}\sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}\left\{(1-A_{ij})\alpha_{ij}(\lambda_1,\lambda_2)-1\right\}X_{ij}\end{pmatrix} = 0. \tag{20}$$

**Step 3.** Obtain the propensity score estimate as

$$\hat{e}_{ij} = \alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)^{-A_{ij}}\{1-\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\}^{-1+A_{ij}}.$$

Finally, our proposed IPTW estimator is

$$\hat{\tau}_{\text{IPTW}} = \frac{1}{N} \sum_{i \in S_I} \sum_{j=1}^{n_i} \omega_{ij} \left\{ \frac{A_{ij} Y_{ij}}{\hat{e}_{ij}} - \frac{(1 - A_{ij}) Y_{ij}}{1 - \hat{e}_{ij}} \right\}. \tag{21}$$

In the above procedure, the design wights are used in both the propensity score estimates and the weighting estimator.

We now consider the asymptotic property of $\hat{\tau}_{\text{IPTW}}$ in (21). We use an asymptotic framework, where the sample size $n$ indexes a sequence of finite populations and samples (Fuller, 2009; Section 1.3), such that the population size $N$ increases with $n$. In addition, we have the following regularity conditions for the sampling mechanism.

**Assumption 6** *(i) The first-order inclusion probability $\pi_i \pi_{j|i}$ is positive and uniformly bounded in the sense that there exist positive constants $C_1$ and $C_2$ that do not depend on N, such that $C_1 \leq \pi_i \pi_{j|i} N n^{-1} \leq C_2$, for any i and j; (ii) the sequence of Horvitz-Thompson estimators $\hat{Y}_{\text{HT}} = N^{-1} \sum_{i \in S_I} \sum_{j=1}^{n_i} \omega_{ij} y_i$ satisfies $\text{var}_p(\hat{Y}_{\text{HT}}) = O\left(n^{-1}\right)$ and $\left\{\text{var}_p(\hat{Y}_{\text{HT}})\right\}^{-1/2} (\hat{Y}_{\text{HT}} - \bar{Y}) \mid \mathscr{F}_N \to \mathscr{N}(0,1)$, in distribution, as $n \to \infty$, where $\bar{Y} = N^{-1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_i$ is the population mean of Y, and the reference distribution is the randomization distribution generated by the sampling mechanism.*

Sufficient conditions for the asymptotic normality of the Horvitz-Thompson estimators are discussed in Chapter 1 of Fuller (2009).

**Theorem 2** *Suppose that Assumptions 2–6, and the regularity conditions specified in the Appendix hold. Suppose further that the cluster sample sizes $N_i$, for $i = 1, \ldots, M$, satisfy the condition that $\min_{1 \leq i \leq M} N_i \to \infty$ and $\sup_{1 \leq i \leq M} N_i = O(N^{1/2})$. If the outcome model (1) is a linear mixed effects model or the propensity score model (2) is correctly specified, the proposed propensity score weighting estimator in (21), subject to constraints (17) and (18), satisfies*

$$V_2^{-1} (\hat{\tau}_{\text{IPTW}} - \tau) \to \mathscr{N}(0,1),$$

*in distribution, as $n \to \infty$, where $V_2 = \text{var}\left( N^{-1} \sum_{i \in S_I} \sum_{j=1}^{n_i} \omega_{ij} \tau_{ij} \right)$,*

$$\begin{aligned} \tau_{ij} = & \{ \alpha_{ij}(\lambda_1^*, \lambda_2^*) A_{ij} (Y_{ij} - B_1^{\mathrm{T}} X_{ij}) + B_1^{\mathrm{T}} X_{ij} \} \\ & - \{ \alpha_{ij}(\lambda_1^*, \lambda_2^*) (1 - A_{ij}) (Y_{ij} - B_2^{\mathrm{T}} X_{ij}) + B_2^{\mathrm{T}} X_{ij} \}, \end{aligned}$$

$$B_1 = E\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}A_{ij}Y_{ij}X_{ij}^{\mathrm{T}}\right]$$
$$\times E\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}A_{ij}X_{ij}X_{ij}^{\mathrm{T}}\right]^{-1},$$

$$B_2 = E\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}(1-A_{ij})Y_{ij}X_{ij}^{\mathrm{T}}\right]$$
$$\times E\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}(1-A_{ij})X_{ij}X_{ij}^{\mathrm{T}}\right]^{-1},$$

*and* $(\lambda_1^*,\lambda_2^*)$ *satisfies* $E\{\hat{Q}(\lambda_1^*,\lambda_2^*)\}=0$ *with* $\hat{Q}(\lambda_1,\lambda_2)$ *defined in (20).*

For variance estimation of $\hat{\tau}_{\mathrm{IPTW}}$, let $\hat{\tau}_{ij}=\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\{A_{ij}(Y_{ij}-\hat{B}_1^{\mathrm{T}}X_{ij})-(1-A_{ij})(Y_{ij}-\hat{B}_2^{\mathrm{T}}X_{ij})\}+(\hat{B}_1-\hat{B}_2)^{\mathrm{T}}X_{ij}$, where

$$\hat{B}_1 = \sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\left\{1-\frac{\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)}{n_i}\right\}A_{ij}Y_{ij}X_{ij}^{\mathrm{T}}$$
$$\times\left[\sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\left\{1-\frac{\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)}{n_i}\right\}A_{ij}X_{ij}X_{ij}^{\mathrm{T}}\right]^{-1},$$

$$\hat{B}_2 = \sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\left\{1-\frac{\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)}{n_i}\right\}(1-A_{ij})Y_{ij}X_{ij}^{\mathrm{T}}$$
$$\times\left[\sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)\left\{1-\frac{\alpha_{ij}(\hat{\lambda}_1,\hat{\lambda}_2)}{n_i}\right\}(1-A_{ij})X_{ij}X_{ij}^{\mathrm{T}}\right]^{-1}.$$

Let $\hat{\tau}_i=\sum_{j=1}^{n_i}\pi_{j|i}^{-1}\hat{\tau}_{ij}$ and

$$\hat{V}_i=\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\frac{\pi_{kl|i}-\pi_{k|i}\pi_{l|i}}{\pi_{kl|i}}\frac{\hat{\tau}_{ik}}{\pi_{k|i}}\frac{\hat{\tau}_{il}}{\pi_{l|i}}.$$

The variance estimator can be constructed as

$$\hat{V}(\hat{\tau}_{\mathrm{IPTW}})=\frac{1}{N^2}\left(\sum_{i\in S_I}\sum_{j\in S_I}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{\hat{\tau}_i}{\pi_i}\frac{\hat{\tau}_j}{\pi_j}+\sum_{i\in S_I}\frac{\hat{V}_i}{\pi_i}\right).$$

# 6 Simulation studies

We conduct two simulation studies to evaluate the finite-sample performance of the proposed estimator, assessing its robustness against model misspecification in Section 6.1 and the robustness against omitting a unit-level confounder in Section 6.2.

## 6.1 Robustness against model misspecification

We first generate finite populations and then select a sample from each finite population using a two-stage cluster sampling design. In the first setting, we specify the number of clusters in the population to be $M = 10,000$, and the size of the $i$th cluster size $N_i$ to be the integer part of $500 \times \exp(2+U_i)/\{1+\exp(2+U_i)\}$, where $U_i \sim \mathcal{N}(0,1)$. The cluster sizes range from 100 to 500. The potential outcomes are generated according to linear mixed effects models, $Y_{ij}(0) = X_{ij} + U_i + e_{ij}$ and $Y_{ij}(1) = X_{ij} + \tau + \tau U_i + e_{ij}$, where $\tau = 2$, $X_{ij} \sim \mathcal{N}(0,1)$, $e_{ij} \sim \mathcal{N}(0,1)$, $U_i$, $X_{ij}$, and $e_{ij}$ are independent, for $i = 1,\ldots,M$, $j = 1,\ldots,N_i$. The parameter of interest is $\tau$. We consider three propensity score models, $\text{pr}(A_{ij} = 1 \mid X_{ij}; U_i) = h(\gamma_0 + \gamma_1 U_i + X_{ij})$, with $h(\cdot)$ being the inverse logit, probit and complementary log-log link function, for generating $A_{ij}$. We set $(\gamma_0, \gamma_1)$ to be $(-0.5, 1)$, $(-0.25, 0.5)$ and $(-0.5, 0.1)$ for the above three propensity score models, respectively. The observed outcome is $Y_{ij} = A_{ij}Y_{ij}(1) + (1-A_{ij})Y_{ij}(0)$. From each realized population, $m$ clusters are sampled by Probability-Proportional-to-Size (PPS) sampling with the measure of size $N_i$. So the first-order inclusion probability of selecting cluster $i$ is equal to $\pi_i = mN_i/\sum_{i=1}^{m} N_i$, which implicitly depends on the unobserved random effect. Once the clusters are sampled, the $n_i$ units in the $i$th selected cluster are sampled by Poisson sampling with the corresponding first-order inclusion probability $\pi_{j|i} = n_e z_{ij}/(\sum_{j=1}^{M_i} z_{ij})$, where $z_{ij} = 0.5$ if $e_{ij} < 0$ and 1 if $e_{ij} > 0$. With this sampling design, the units with $e_{ij} > 0$ are sampled with a chance twice as big as the units with $e_{ij} < 0$. We consider four combinations of $m$ and $n_e$: (i) $(m, n_e) = (50, 50)$; (ii) $(m, n_e) = (100, 30)$, representing a large number of small clusters; (iii) $(m, n_e) = (30, 100)$; and (iv) $(m, n_e) = (5, 100)$, representing a small number of large clusters.

In the second setting, all data-generating mechanisms are the same as in the first setting, except that the potential outcomes are generated according to logistic linear mixed effects models, $Y_{ij}(0) \sim \text{Bernoulli}(p_{ij}^0)$ with $\text{logit}(p_{ij}^0) = X_{ij} + U_i$ and $Y_{ij}(1) \sim \text{Bernoulli}(p_{ij}^1)$ with $\text{logit}(p_{ij}^1) = X_{ij} + \tau + \tau u_i$. Moreover, in the 2-stage cluster sampling, $\pi_{j|i} = n_e z_{ij}/(\sum_{j=1}^{M_i} z_{ij})$, where $z_{ij} = 0.5$ if $Y_{ij} = 0$ and 1 if $Y_{ij} = 1$.

With this sampling design, the units with $Y_{ij} = 1$ are sampled with a chance twice as big as the units with $Y_{ij} = 0$.

We compare four estimators for $\tau$: (i) $\hat{\tau}_{\text{simp}}$, the simple design-weighted estimator without propensity score adjustment; (ii) $\hat{\tau}_{\text{fix}}$, the weighting estimator in (3) with the propensity score estimated by a logistic linear fixed effects model with fixed cluster intercepts; (iii) $\hat{\tau}_{\text{ran}}$, the weighting estimator in (3) with the propensity score estimated by a logistic linear mixed effects model with random cluster intercepts; (iv) $\hat{\tau}_{\text{IPTW}}$, the proposed estimator with calibrations (17) and (18).

Table 1 shows biases, variances and coverages for 95% confidence intervals from $1,000$ simulated data sets. The simple estimator shows large biases across difference scenarios, even adjusting for sampling design. This suggests that the covariate distributions are different between the treatment groups in the finite population, contributing to the bias. $\hat{\tau}_{\text{fix}}$ works well under Scenario 1 with the linear mixed effects model for the outcome and the logistic linear mixed effects model for the propensity score; however, its performance is not satisfactory in other scenarios. Moreover, $\hat{\tau}_{\text{fix}}$ shows the largest variance among the four estimators in most of scenarios. This is because for a moderate or large number of clusters, there are too many free parameters, and hence the propensity score estimates may not be stable. For $\hat{\tau}_{\text{ran}}$, we assume that the cluster effect is random, which reduces the number of free parameters. As a result, $\hat{\tau}_{\text{ran}}$ shows less variability than $\hat{\tau}_{\text{fix}}$. Nonetheless, both $\hat{\tau}_{\text{fix}}$ and $\hat{\tau}_{\text{ran}}$ cannot control the bias well. The proposed estimator shows small bias and good empirical coverage across all scenarios. Notably, to compute $\hat{\tau}_{\text{IPTW}}$, we used a working model, a logistic linear model, to provide an initial set of weights. When the true propensity score is probit or complementary log-log model, $\hat{\tau}_{\text{IPTW}}$ still has small biases. This suggests that our proposed estimator achieves improved robustness compared to the existing weighting estimators.

## 6.2 Robustness against omitting a unit-level (higher moment of) confounder

The data generating mechanisms are the same as in Section 6.1, except that in the potential outcomes and the treatment assignment models, we use a squared covariate instead of the original covariate. Now, for the potential outcomes models, we have in the first setting, $Y_{ij}(0) = X_{ij}^2 + U_i + e_{ij}$ and $Y_{ij}(1) = X_{ij}^2 + \tau + \tau U_i + e_{ij}$, where $\tau = 2$, $X_{ij} \sim \mathcal{N}(0,1)$, $e_{ij} \sim \mathcal{N}(0,1)$, $U_i$, $X_{ij}$, and $e_{ij}$ are independent, for $i = 1,\ldots,M$, $j = 1,\ldots,N_i$; while in the second setting, $Y_{ij}(0) \sim \text{Bernoulli}(p_{ij}^0)$ with $\text{logit}(p_{ij}^0) = X_{ij}^2 + U_i$ and $Y_{ij}(1) \sim \text{Bernoulli}(p_{ij}^1)$ with $\text{logit}(p_{ij}^1) = X_{ij}^2 + \tau + \tau u_i$. For three propensity score models, we now have $\text{pr}(A_{ij} = 1 \mid X_{ij}; U_i) = h(\gamma_0 + \gamma_1 U_i +$

$X_{ij}^2$), with $h(\cdot)$ being the inverse logit, probit and complementary log-log link function, for generating $A_{ij}$. We set $(m, n_e) = (50, 50)$. In the proposed method, the calibration condition (17) is imposed only for the first moment of $X_{ij}$. This represents the case of omitting a unit-level confounder.

Table 2 shows biases, variances and coverages for 95% confidence intervals from $1,000$ simulated data sets. The proposed estimator $\hat{\tau}_{\text{IPTW}}$ does not control bias well in some scenarios, similar to all other estimators. This is in line with the consensus in the causal inference literature that all unit-level confounders, including higher moments if present, must be properly controlled for in order to obtain a consistent causal effect estimator.

# 7 An application

Ethical approval: The conducted research uses an existing de-identified dataset and is not considered as human subject research by the authors' institutional review board. We examine the 2007–2010 BMI surveillance data from Pennsylvania Department of Health to investigate the effect of School Body Mass Index Screening (SBMIS) on the annual overweight and obesity prevalence in elementary schools in Pennsylvania. Early studies have shown that SBMIS has been associated with increased parental awareness of child weight (Harris, Kuramoto, Schulzer and Retallack, 2009, Ebbeling, Feldman, Chomitz, Antonelli, Gortmaker, Osganian and Ludwig, 2012). However, there have been mixed findings about the effect of screening on reducing prevalence of overweight and obesity (Harris et al., 2009, Thompson and Card-Higginson, 2009). The data set includes 493 schools in Pennsylvania. The baseline is the school year 2007. Previous studies (e.g. Peyer, Welk, Bailey-Davis, Yang and Kim, 2015) have shown that two high-level contextual factors are strongly associated with school policies for SBMIS: type of community (rural, suburban, and urban), and population density (low, median, and high). Therefore, we cluster schools according to these two factors. This results in five clusters: rural-low, rural-median, rural-high, suburban-high, and urban-high, with cluster sample sizes $n_1 = 33$, $n_2 = 118$, $n_3 = 116$, $n_4 = 84$, and $n_5 = 142$, respectively.

Let $A = 1$ if the school implemented SBMIS, and $A = 0$ if the school did not. In this data set, 63% of schools implemented SBMIS, and the percentages of schools implemented SBMIS across the clusters range from 45% to 70%, indicating cluster-level heterogeneity of treatment. The outcome variable $Y$ is the annual overweight and obesity prevalence for each school in the school year 2010. The prevalence is calculated by dividing the number of students with BMI$>$ 85th by the total number of students screened for each school. Therefore, the outcome was

measured for each school. For each school, we obtain school characteristics including the baseline prevalence of overweight and obesity ($X_1$), and percentage of reduced and free lunch ($X_2$).

For a direct comparison, the average difference of the prevalence of overweight and obesity for schools that implemented SBMIS and those that did not is 8.78%. This unadjusted difference in the prevalence of overweight and obesity ignores differences in schools and clusters. To take the cluster-level heterogeneity of treatment into account, we consider three propensity score models: (i) a logistic linear fixed effects model with linear predictors including $X_1$, $X_2$, and a fixed intercept for each cluster; (ii) a logistic linear mixed effects model with linear predictors including $X_1$, $X_2$, and a random intercept for each cluster; (iii) the proposed calibrated propensity score. Using the estimated propensity score, we estimate the average treatment effect $\tau$ by the weighting method.

Table 3 displays the standardized differences of means for $X_1$ and $X_2$ between the treated and control groups for each cluster and the whole population, standardized by the standard errors in the whole population. Without any adjustment, there are large differences in means for $X_1$ and $X_2$. For this specific data set, the three methods for modeling and estimating the propensity score are similar in balancing the covariate distributions between the treated and control groups. All three propensity score weighting methods improve the balance for $X_1$ and $X_2$. Table 4 displays point estimates and variance estimates based on 500 bootstrap replicates. The simple estimator shows that the screening has a significant effect in reducing the prevalence of overweight and obesity. However, this may be due to confounders. After adjusting for the confounders, the screening does not have a significant effect. Given the different sets of assumptions for the different methods, this conclusion is reassuring.

# 8 Discussion

We provide a doubly robust construction of inverse propensity score weights by imposing the exact balance of unit- and (observed and unobserved) cluster-level covariate distributions between the treatment groups. When either the treatment assignment is correctly specified or the outcome follows a linear mixed effects model, we show that consistent estimation of the average treatment effect is possible. Our simulation examines the robustness property of the proposed estimator under various data generating mechanisms. The results suggest that if all confounders in the linear predictors of the treatment and outcome models (including all higher moments) achieve a good balance between two treatment groups, the proposed estimator is robust. The balance conditions help to satisfy the underlying latent ignorable

treatment assignment assumption, and may be particularly useful in the case where not sufficient cluster-level confounders are available. In this case, misspecification of the propensity score model has little impact on the bias of the causal effect estimator.

Moreover, our simulation results also indicate that robustness may not hold in the case where higher moments of unit-level confounders are present however are omitted in the balance constraints. This is similar to the case when there are unmeasured unit-level confounders. We therefore emphasize that unbiased estimation of the average treatment effect requires that all unit-level confounders be sufficiently controlled for.

It is well known that the IPTW estimator is sensitive to near-zero values of the estimated propensity score (e.g. Robins, Sued, Lei-Gomez and Rotnitzky, 2007, Kang and Schafer, 2007, Cao et al., 2009). Our proposed estimator prevents the occurrence of extreme values of weights through the calibration constraints where the weights within each cluster are positive and sum to the cluster sample sizes. Therefore, it is unlikely that some units receive extremely large weights that dominate.

We have focused on two-level data with a binary treatment and the average treatment effect over the full population in this article. Our proposed method can be easily generalized to many other scenarios not discussed here, such as multilevel data, multiple treatments (Yang et al., 2016) or other causal estimands, e.g., the average causal effects over a subset of population (Crump, Hotz, Imbens and Mitnik, 2006, Li, Morgan and Zaslavsky, 2017, Yang and Ding, 2018), including the average causal effect on the treated.

The IPTW estimator is not efficient in general. Semiparametric efficiency bounds for estimating the average treatment effects in the setting with i.i.d. random variables were derived by Hahn (1998). He showed that the efficient influence function for the average treatment effect depends on both the propensity score and the outcome model. An important implication is that combining the propensity score model and the outcome regression model can improve efficiency of the IPTW estimator. For clustered data, because the data are correlated through the random cluster variables, the efficiency theory established for the i.i.d. data is not applicable. It remains an interesting avenue for future research to develop the semiparametric efficiency theory for clustered data.

# Acknowledgments

The author acknowledges the support in part by Ralph E. Powe Junior Faculty Enhancement Award from Oak Ridge Associated Universities and NSF grant DMS 1811245. The author would like to thank Peng Ding and Fan Li for insightful and

fruitful discussions. The author is grateful to the Associated Editor and two referees for helpful comments and suggestions that have greatly improved the original submission. Conflict of interest: Authors state no conflict of interest.

# Appendix.

# Regularity conditions and proof of Theorem 1

We formulate the proposed estimator as a Z-estimator (e.g., van der Vaart, 2000), which invokes the standard Z-estimation theory to show the asymptotic properties. Write $\hat{\tau}_{\mathrm{IPTW}}(\lambda_1,\lambda_2) = n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\lambda_1,\lambda_2)Y_{ij}$, where $\alpha_{ij}(\lambda_1,\lambda_2)$ is defined in (10). The proposed estimator is $\hat{\tau}_{\mathrm{IPTW}}(\hat{\lambda}_1,\hat{\lambda}_2)$, where $(\hat{\lambda}_1,\hat{\lambda}_2)$ satisfies $\hat{Q}(\hat{\lambda}_1,\hat{\lambda}_2) = 0$, where $\hat{Q}(\lambda_1,\lambda_2)$ is defined in (11). Define $Q(\lambda_1,\lambda_2) = \lim_{n\to\infty} E\{\hat{Q}(\lambda_1,\lambda_2)\}$, and define $(\lambda_1^*,\lambda_2^*)$ that satisfy $Q(\lambda_1^*,\lambda_2^*) = 0$. Denote $A \cong B$ as $A = B + o_P(n^{-1/2})$, where the reference distribution is the super-population model, as $n \to \infty$.

We impose the following regularity conditions.

**Condition A1** *$\hat{Q}(\lambda_1,\lambda_2) \to Q(\lambda_1,\lambda_2)$ in probability uniformly for $(\lambda_1,\lambda_2) \in \mathscr{B}$ as $n \to \infty$, and there exists a unique $(\lambda_1^*,\lambda_2^*) \in \mathscr{B}$ such that $Q(\lambda_1^*,\lambda_2^*) = 0$;*

**Condition A2** *$\partial\hat{\tau}_{\mathrm{IPTW}}(\lambda_1,\lambda_2)/\partial(\lambda_1^{\mathrm{T}},\lambda_2^{\mathrm{T}})$ and $\partial\hat{Q}(\lambda_1,\lambda_2)/\partial(\lambda_1^{\mathrm{T}},\lambda_2^{\mathrm{T}})$ are continuous at $(\lambda_1,\lambda_2) \in \mathscr{B}$ almost surely;*

**Condition A3** *The matrix*

$$E\left\{\frac{\partial\hat{Q}(\lambda_1^*,\lambda_2^*)}{\partial(\lambda_1^{\mathrm{T}},\lambda_2^{\mathrm{T}})^{\mathrm{T}}}\right\} = E\left[\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(\lambda_1^*,\lambda_2^*)\left\{1-\frac{\alpha_{ij}(\lambda_1^*,\lambda_2^*)}{n_i}\right\}A_{ij}X_{ij}X_{ij}^{\mathrm{T}}\right]$$

*is invertible;*

**Condition A4** *$E||X_{ij}||^3 < \infty$, $E|Y_{ij}(0)|^3 < \infty$, and $E|Y_{ij}(1)|^3 < \infty$.*

The convergence in Condition A1 is uniform convergence. That is, given $\varepsilon > 0$, there exists $n_0 = n_0(\varepsilon)$ such that $\mathrm{pr}\left\{|\hat{Q}(\lambda_1,\lambda_2) - Q(\lambda_1,\lambda_2)| > \varepsilon\right\} \le \varepsilon$ holds for all $n \ge n_0$ and $(\lambda_1,\lambda_2) \in \mathscr{B}$. A sufficient condition for the uniform convergence is that $|\alpha_{ij}(\lambda_1,\lambda_2)| < M$ for all $(i,j)$ and $(\lambda_1,\lambda_2) \in \mathscr{B}$, where $M$ is a constant. Condition A4 specifies moment conditions for the central limit theorem. Conditions A1–A4 are standard regularity conditions on Z-estimation; see, e.g., van der Vaart

(2000). However, the regularity conditions may be difficult to check in practice. We give an extreme example where a certain condition is violated. For example, if two covariates in $X$ are perfectly correlated, then Condition A3 fails to hold. Aside from extreme cases, the regularity conditions are often satisfied for the models we are considering and reasonable choices of covariates in practice.

Under Conditions A1–A4, using the standard linearization technique, we obtain

$$\begin{aligned}\hat{\tau}_{\mathrm{IPTW}}(\hat{\lambda}_1,\hat{\lambda}_2) &\cong \hat{\tau}_{\mathrm{IPTW}}(\lambda_1^*,\lambda_2^*) \\ &\quad -E\left\{\frac{\partial\hat{\tau}_{\mathrm{IPTW}}(\lambda_1^*,\lambda_2^*)}{\partial(\lambda_1^{\mathrm{T}},\lambda_2^{\mathrm{T}})}\right\}E\left\{\frac{\partial\hat{Q}(\lambda_1^*,\lambda_2^*)}{\partial(\lambda_1^{\mathrm{T}},\lambda_2^{\mathrm{T}})}\right\}^{-1}\hat{Q}(\lambda_1^*,\lambda_2^*) \\ &\equiv \hat{\tau}_{\mathrm{IPTW}}(\lambda_1^*,\lambda_2^*)-B_1^{\mathrm{T}}\hat{Q}_1(\lambda_1^*,\lambda_2^*)-B_2^{\mathrm{T}}\hat{Q}_2(\lambda_1^*,\lambda_2^*). \qquad \text{(A1)}\end{aligned}$$

First, consider the case when the initial propensity score model is correctly specified and consistently estimated. We have $e_{ij}^0 \cong e_{ij}$ and $\lambda_1^* = \lambda_2^* = 0$. This is because with $\lambda_1^* = \lambda_2^* = 0$, $\lim_{n\to\infty} E\{\hat{Q}(\lambda_1^*,\lambda_2^*)\} = 0$. We now evaluate the terms in (A1) further. We express $\hat{\tau}_{\mathrm{IPTW}}(0,0)$ as

$$\begin{aligned}& n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\alpha_{ij}(0,0)Y_{ij} \\ =\ & n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{d_{ij}A_{ij}Y_{ij}}{n_i^{-1}\sum_{j=1}^{n_i}d_{ij}A_{ij}} - n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{d_{ij}(1-A_{ij})Y_{ij}}{n_i^{-1}\sum_{j=1}^{n_i}d_{ij}(1-A_{ij})} \\ =\ & n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{d_{ij}A_{ij}Y_{ij}(1)}{n_i^{-1}\sum_{j=1}^{n_i}d_{ij}A_{ij}} - n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{d_{ij}(1-A_{ij})Y_{ij}(0)}{n_i^{-1}\sum_{j=1}^{n_i}d_{ij}(1-A_{ij})} \\ \cong\ & n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{e_{ij}^{-1}A_{ij}Y_{ij}(1)}{n_i^{-1}\sum_{j=1}^{n_i}e_{ij}^{-1}A_{ij}} - n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{(1-e_{ij})^{-1}(1-A_{ij})Y_{ij}(0)}{n_i^{-1}\sum_{j=1}^{n_i}(1-e_{ij})^{-1}(1-A_{ij})} \\ \cong\ & \tau, \qquad \text{(A2)}\end{aligned}$$

where the third line follows by the consistency assumption, the forth line follows by the assumption that the initial propensity score model is correctly specified, and the last line follows by the strong law of large numbers and the condition of $\min_{1\le i\le m} n_i \to \infty$. Also, following the similar argument, we obtain

$$\hat{Q}_1(0,0) = n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{A_{ij}\alpha_{ij}(0,0)-1\right\}X_{ij} \cong 0, \qquad \text{(A3)}$$

$$\hat{Q}_2(0,0) = n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{(1-A_{ij})\alpha_{ij}(0,0)-1\right\}X_{ij} \cong 0. \qquad \text{(A4)}$$

Combining (A1)–(A4), we obtain $\hat{\tau}_{\mathrm{IPTW}}(\hat{\lambda}_1,\hat{\lambda}_2) \cong \tau$.

Second, consider the case when the outcome follows a linear mixed effects model. We do not assume that the initial propensity score model is correctly specified, and therefore $\lambda_1^*$ and $\lambda_2^*$ in (A1) are not necessarily zero. Conditions A1–A4 ensure that (A1) is consistent for some parameter. We have shown in Section 4 that the proposed estimator is asymptotically unbiased for $\tau$. It follows that $\hat{\tau}_{\text{IPTW}}(\hat{\lambda}_1, \hat{\lambda}_2) \cong \tau$.

To derive asymptotic variance formula, continuing (A1), we obtain

$$\begin{aligned}
\hat{\tau}_{\text{IPTW}}(\hat{\lambda}_1, \hat{\lambda}_2) &= \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\alpha_{ij}(\lambda_1^*, \lambda_2^*)A_{ij}(Y_{ij} - B_1^{\mathrm{T}}X_{ij}) + B_1^{\mathrm{T}}X_{ij}\right\} \\
&\quad -\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\alpha_{ij}(\lambda_1^*, \lambda_2^*)(1-A_{ij})(Y_{ij} - B_2^{\mathrm{T}}X_{ij}) + B_2^{\mathrm{T}}X_{ij}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tau_{ij},
\end{aligned}$$

where

$$\begin{aligned}
\tau_{ij} = &\left\{\alpha_{ij}(\lambda_1^*, \lambda_2^*)A_{ij}(Y_{ij} - B_1^{\mathrm{T}}X_{ij}) + B_1^{\mathrm{T}}X_{ij}\right\} \\
&- \left\{\alpha_{ij}(\lambda_1^*, \lambda_2^*)(1-A_{ij})(Y_{ij} - B_2^{\mathrm{T}}X_{ij}) + B_2^{\mathrm{T}}X_{ij}\right\}.
\end{aligned}$$

Therefore, $\text{var}(\hat{\tau}_{\text{IPTW}}) = \text{var}(n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tau_{ij})$, denoted as $V_1$.

To establish the asymptotic normality of $\hat{\tau}_{\text{IPTW}}$, we use the central limit theory for dependent variables (Hoeffding, Robbins et al., 1948, Serfling, 1968). Let $\text{var}(\tau_{ij}) = \sigma_\tau^2$ and $\text{cov}(\tau_{ij}, \tau_{ik}) = \nu_\tau$ for $j \neq k$. Arrange the $\tau_{ij}$'s in a $n$-length sequence $\{\tau_{11}, \ldots, \tau_{1n_1}, \tau_{21}, \ldots, \tau_{mn_m}\}$. To simplify the notation, let the $k$th random variable in this sequence be denoted by $\tau_k$, for $k = 1, \ldots, n$. We now consider such sequences $\{\tau_k : k = 1, \ldots, n\}$ are indexed by $n$. By Condition A4, the absolute central moments $E|\tau_k - E(\tau_k)|^3$ are bounded uniformly in $k$. Moreover, by the assumption of $\sup_{1\leq i\leq m} n_i = O(n^{1/2})$, we then have $\text{var}(\sum_{k=a+1}^{a+n}\tau_k) \sim nA^2$, uniformly in $a$, as $n \to \infty$, where $A^2$ is a positive constant. Following Serfling (1968), these are typical criterion for verifying the Lindeberg condition (Loève, 1960), and therefore $V_1^{-1/2}(\hat{\tau}_{\text{IPTW}} - \tau) \to \mathcal{N}(0,1)$, in distribution, as $n \to \infty$.

# Regularity conditions for Theorem 2

For the clustered survey data, we now write $\hat{\tau}_{\text{IPTW}}(\lambda_1, \lambda_2) = N^{-1}\sum_{i\in S_I}\sum_{j=1}^{n_i}\omega_{ij}$ $\times\alpha_{ij}(\lambda_1, \lambda_2)Y_{ij}$, where $\alpha_{ij}(\lambda_1, \lambda_2)$ is defined in (19). The proposed estimator is $\hat{\tau}_{\text{IPTW}}(\hat{\lambda}_1, \hat{\lambda}_2)$, where $(\hat{\lambda}_1, \hat{\lambda}_2)$ satisfies $\hat{Q}(\hat{\lambda}_1, \hat{\lambda}_2) = 0$, where $\hat{Q}(\lambda_1, \lambda_2)$ is defined in

(20). We assume Conditions A1– A4 holds with the new definitions of $\hat{\tau}_{\text{IPTW}}(\lambda_1, \lambda_2)$, $\alpha_{ij}(\lambda_1, \lambda_2)$, and $\hat{Q}(\lambda_1, \lambda_2)$.

# References


Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score, *Econometrica* **84**: 781–807.

Arpino, B. and Mealli, F. (2011). The specification of the propensity score in multilevel observational studies, *Computational Statistics & Data Analysis* **55**: 1770–1780.

Baltagi, B. (1995). *Econometric Analysis of Panel Data*, John Wiley & Sons, Wiley: New York.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models, *Biometrics* **61**: 962–973.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data, *Biometrika* **96**: 723–734.

Chan, K. C. G., Yam, S. C. P. and Zhang, Z. (2015). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting, *J. R. Statist. Soc.* B **78**: 673–700.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika* **95**: 555–571.

Chen, J., Sitter, R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika* **89**: 230–237.

Crump, R., Hotz, V. J., Imbens, G. and Mitnik, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, *Technical report*, 330, National Bureau of Economic Research, Cambridge, MA.

Dawid, A. P. (1979). Conditional independence in statistical theory, *J. R. Statist. Soc. B* **41**: 1–31.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling, *J. Am. Stat. Assoc.* **87**: 376–382.

Ebbeling, C. B., Feldman, H. A., Chomitz, V. R., Antonelli, T. A., Gortmaker, S. L., Osganian, S. K. and Ludwig, D. S. (2012). A randomized trial of sugar-sweetened beverages and adolescent body weight, *N. Engl. J. Med.* **367**: 1407–1416.

Eckardt, P. (2012). Propensity score estimates in multilevel models for causal inference, *Nurs. Res.* **61**: 213–223.

Fuller, W. A. (2009). *Sampling Statistics*, Wiley, Hoboken, NJ.

Graham, B. S., Pinto, C. C. D. X. and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data, *Rev. Econ. Stud.* **79**: 1053–1079.

Griswold, M. E., Localio, A. R. and Mulrow, C. (2010). Propensity score adjustment with multilevel data: setting your sites on decreasing selection bias, *Ann. Intern. Med.* **152**: 393–395.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica* **66**: 315–331.

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

Harris, K. C., Kuramoto, L. K., Schulzer, M. and Retallack, J. E. (2009). Effect of school-based physical activity interventions on body mass index in children: a meta-analysis, *Can. Med. Assoc. J.* **180**: 719–726.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization, *Health Services and Outcomes Research Methodology* **2**: 259–278.

Hoeffding, W., Robbins, H. et al. (1948). The central limit theorem for dependent random variables, *Duke Math. J* **15**: 773–780.

Holland, P. W. (1986). Statistics and causal inference, *J. Am. Stat. Assoc.* **81**: 945–960.

Hong, G. and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data, *J. Am. Stat. Assoc.* **101**: 901–910.

Hong, G. and Yu, B. (2007). Early-grade retention and children's reading and math learning in elementary years, *Educational Evaluation and Policy Analysis* **29**: 239–261.

Hong, G. and Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data, *Dev. Psychol.* **44**: 407–421.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference, *J. Am. Stat. Assoc.* **103**: 832–842.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score, *J. R. Statist. Soc.* B **76**: 243–263.

Imbens, G., Johnson, P. and Spady, R. H. (1998). Information theoretic approaches to inference in moment condition models, *Econometrica* **66**: 333–357.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, Cambridge UK.

Kang, J. D. and 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.

Kelcey, B. M. (2009). *Improving and assessing propensity score based causal inferences in multilevel and nonlinear settings*, PhD thesis, University of Michigan.

Kim, J. K., Kwon, Y. and Paik, M. C. (2016). Calibrated propensity score method for survey nonresponse in cluster sampling, *Biometrika* **103**: 461–473.

Kim, J.-S. and Steiner, P. M. (2015). Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy, *Quantitative Psychology Research*, Springer, pp. 293–306.

Kim, J. and Seltzer, M. (2007). Causal inference in multilevel settings in which selection processes vary across schools, *Technical Report Working Paper 708*, University of California, Los Angeles, Center for the Study of Evaluation.

Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation, *Econometrica* **65**: 861–874.

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

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *The Annals of Mathematical Statistics* **22**: 79–86.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010). Improving propensity score weighting using machine learning, *Stat. Med.* **29**: 337–346.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W. and Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies, *Multivariate Behavioral Research* **50**: 265–284.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2017). Balancing covariates via propensity score weighting, *J. Am. Stat. Assoc.* p. doi:10.1080/01621459.2016.1260466.

Li, F., Zaslavsky, A. M. and Landrum, M. B. (2013). Propensity score weighting with multilevel data, *Stat. Med.* **32**: 3373–3387.

Loève, M. (1960). *Probability Theory*, 2 edn, Van Nostrand: Princeton.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study, *Stat. Med.* **23**: 2937–2960.

McCaffrey, D. F., Ridgeway, G. and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies, *Psychol. Methods* pp. 403–425.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica* **72**: 219–255.

Oakes, J. M. (2004). The (mis) estimation of neighborhood effects: causal inference for a practicable social epidemiology, *Social Science & Medicine* **58**: 1929–1952.

Park, M. and Fuller, W. A. (2012). Generalized regression estimators, *Encyclopedia of Environmetrics* **2**: 1162–1166.

Peyer, K. L., Welk, G., Bailey-Davis, L., Yang, S. and Kim, J.-K. (2015). Factors associated with parent concern for child weight and parenting behaviors, *Childhood Obesity* **11**: 269–274.

Pirracchio, R., Petersen, M. L. and van der Laan, M. (2014). Improving propensity score estimators' robustness to model misspecification using super learner, *Am. J. Epidemiol.* **181**: 108–119.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies, *J. R. Statist. Soc.* B **69**: 101–122.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *J. Am. Stat. Assoc.* **89**: 846–866.

Robins, J., Sued, M., Lei-Gomez, Q. and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable, *Statistical Science* **22**: 544–559.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**: 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score, *J. Am. Stat. Assoc.* **79**: 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, *The American Statistician* **39**: 33–38.

Ross, R. (1916). An application of the theory of probabilities to the study of a priori pathometry. part i, *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* **92**: 204–230.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *J. Educ. Psychol.* **66**: 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization, *Ann. Statist.* **6**: 34–58.

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research, *Pharmacoepidemiology and Drug Safety* **13**: 855–857.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*, John Wiley & Sons: New York.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood, *Ann. Statist.* **35**: 634–672.

Schuler, M. S., Chu, W. and Coffman, D. (2016). Propensity score weighting for a continuous exposure with multilevel data, *Health Services and Outcomes Research Methodology* **16**: 271–292.

Serfling, R. J. (1968). Contributions to central limit theory for dependent variables, *The Annals of Mathematical Statistics* **39**: 1158–1175.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J. and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study, *Pharmacoepidemiology and Drug Safety* **17**: 546–555.

Skinner, C. J. et al. (2011). Inverse probability weighting for clustered nonresponse, *Biometrika* **98**: 953–966.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets, *Educational Researcher* **36**: 187–198.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward, *Statist. Sci.* **25**: 1–21.

Su, Y.-S. and Cortina, J. (2009). What do we gain? combining propensity score methods and multilevel modeling, *Annual Meeting of the American Political Science Association, Toronto, Canada (2009)*.

Thoemmes, F. J. and West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data, *Multivariate Behavioral Research* **46**: 514–543.

Thompson, J. W. and Card-Higginson, P. (2009). Arkansas' experience: statewide surveillance and parental information on the child obesity epidemic, *Pediatrics* **124**: 73–82.

van der Vaart (2000). *Asymptotic Statistics*, Vol. 3, Cambridge university press, Cambridge: Cambridge University Press.

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research, *Stat. Med.* **27**: 1934–1943.

Wallace, T. D. and Hussain, A. (1969). The use of error components models in combining cross section with time series data, *Econometrica* **37**: 55–72.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT press, Cambridge, MA: MIT Press.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data, *J. Am. Stat. Assoc.* **96**: 185–193.

Xiang, Y. and Tarasawa, B. (2015). Propensity score stratification using multilevel models to examine charter school achievement effects, *Journal of School Choice* **9**: 179–196.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores, *Biometrika* **105**: 487–493.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E. and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments, *Biometrics* **72**: 1055–1065.

Table 1: Simulation results: bias, variance ($\times 10^{-3}$) and coverage (%) of 95% confidence intervals based on $1,000$ Monte Carlo samples; the outcome is linear and logistic linear mixed effects model and the propensity score is logistic, probit or complementary log-log (C-loglog).

| $(m, n_e)$ | $(50,50)$ | | | $(100,30)$ | | | $(30,100)$ | | | $(5,100)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias | var | cvg | bias | var | cvg | bias | var | cvg | bias | var | cvg |
| Scenario 1: Linear outcome & Logistic propensity score | | | | | | | | | | | | |
| $\hat{\tau}_{\text{simp}}$ | -0.37 | 22 | 27.4 | -0.38 | 12 | 8.7 | -0.38 | 35 | 42.3 | -0.38 | 228 | 55.1 |
| $\hat{\tau}_{\text{fix}}$ | -0.01 | 36 | 95.6 | 0.00 | 21 | 95.6 | -0.01 | 42 | 95.2 | -0.01 | 298 | 82.0 |
| $\hat{\tau}_{\text{ran}}$ | 0.14 | 26 | 90.2 | 0.21 | 14 | 64.6 | 0.07 | 37 | 94.7 | 0.07 | 263 | 87.1 |
| $\hat{\tau}_{\text{cal}}$ | 0.01 | 26 | 94.5 | 0.02 | 11 | 95.1 | 0.00 | 33 | 95.6 | 0.00 | 245 | 93.3 |
| Scenario 2: Linear outcome & Probit propensity score | | | | | | | | | | | | |
| $\hat{\tau}_{\text{simp}}$ | -0.29 | 16 | 34.4 | -0.08 | 9 | 2.3 | -0.22 | 30 | 65.6 | -0.30 | 162 | 58.0 |
| $\hat{\tau}_{\text{fix}}$ | 0.08 | 35 | 90.3 | -0.10 | 19 | 4.5 | 0.12 | 69 | 90.4 | 0.07 | 341 | 82.5 |
| $\hat{\tau}_{\text{ran}}$ | 0.24 | 28 | 73.9 | -0.07 | 16 | 29.9 | 0.21 | 60 | 85.5 | 0.15 | 303 | 86.7 |
| $\hat{\tau}_{\text{cal}}$ | 0.01 | 22 | 94.9 | 0.01 | 11 | 95.4 | 0.00 | 33 | 94.6 | 0.00 | 252 | 95.1 |
| Scenario 3: Linear outcome & C-loglog propensity score | | | | | | | | | | | | |
| $\hat{\tau}_{\text{simp}}$ | -0.21 | 20 | 62.0 | -0.21 | 10 | 41.2 | -0.22 | 30 | 65.6 | -0.21 | 240 | 65.0 |
| $\hat{\tau}_{\text{fix}}$ | 0.12 | 48 | 88.8 | 0.12 | 36 | 82.7 | 0.12 | 69 | 90.4 | 0.13 | 445 | 80.6 |
| $\hat{\tau}_{\text{ran}}$ | 0.29 | 38 | 69.1 | 0.36 | 22 | 32.5 | 0.21 | 60 | 85.5 | 0.22 | 441 | 84.6 |
| $\hat{\tau}_{\text{cal}}$ | 0.00 | 21 | 95.3 | 0.00 | 10 | 95.1 | 0.00 | 33 | 94.6 | -0.01 | 248 | 94.1 |
| Scenario 4: Logistic outcome & Logistic propensity score | | | | | | | | | | | | |
| $\hat{\tau}_{\text{simp}}$ | -0.11 | 100 | 9.1 | -0.11 | 540 | 0.5 | -0.11 | 160 | 20.5 | -0.11 | 9 | 62.9 |
| $\hat{\tau}_{\text{fix}}$ | -0.11 | 44 | 0.3 | -0.11 | 38 | 0.1 | -0.11 | 39 | 0.1 | -0.11 | 3 | 30.6 |
| $\hat{\tau}_{\text{ran}}$ | -0.09 | 33 | 1.3 | -0.08 | 21 | 0.5 | -0.10 | 34 | 0.3 | -0.10 | 2 | 45.8 |
| $\hat{\tau}_{\text{cal}}$ | 0.01 | 74 | 96.3 | 0.01 | 55 | 95.2 | 0.01 | 74 | 95.9 | 0.01 | 5 | 94.4 |
| Scenario 5: Logistic outcome & Probit propensity score | | | | | | | | | | | | |
| $\hat{\tau}_{\text{simp}}$ | -0.08 | 58 | 13.1 | -0.08 | 34 | 2.3 | -0.08 | 81 | 25.3 | -0.07 | 5 | 65.9 |
| $\hat{\tau}_{\text{fix}}$ | -0.10 | 93 | 6.9 | -0.10 | 85 | 4.5 | -0.10 | 73 | 3.8 | -0.10 | 5 | 50.5 |
| $\hat{\tau}_{\text{ran}}$ | -0.08 | 67 | 23.0 | -0.07 | 48 | 29.9 | -0.09 | 61 | 8.3 | -0.09 | 4 | 67.0 |
| $\hat{\tau}_{\text{cal}}$ | 0.01 | 89 | 94.7 | 0.01 | 65 | 95.4 | 0.01 | 84 | 95.0 | 0.01 | 6 | 95.4 |
| Scenario 6: Logistic outcome & C-loglog propensity score | | | | | | | | | | | | |
| $\hat{\tau}_{\text{simp}}$ | -0.06 | 0.3 | 3.2 | -0.06 | 0.2 | 1.0 | -0.06 | 0.2 | 3.7 | -0.06 | 2 | 62.0 |
| $\hat{\tau}_{\text{fix}}$ | -0.05 | 0.5 | 44.6 | -0.05 | 0.5 | 43.6 | -0.05 | 0.5 | 43.0 | -0.05 | 3 | 76.8 |
| $\hat{\tau}_{\text{ran}}$ | -0.03 | 0.5 | 95.4 | -0.03 | 0.4 | 97.3 | -0.03 | 0.4 | 92.8 | -0.03 | 3 | 93.4 |
| $\hat{\tau}_{\text{cal}}$ | -0.01 | 0.7 | 95.5 | 0.00 | 0.6 | 95.8 | -0.01 | 0.7 | 95.2 | 0.00 | 5 | 95.9 |

Table 2: Simulation results: bias, variance ($\times 10^{-3}$) and coverage (%) of 95% confidence intervals based on $1,000$ Monte Carlo samples; the outcome is linear and logistic linear mixed effects model and the propensity score is logistic, probit or complementary log-log (C-loglog).

| Scenario | 1 Linear outcome Logistic PS | | | 2 Linear outcome Probit PS | | | 3 Linear outcome C-loglog PS | | |
|---|---|---|---|---|---|---|---|---|---|
| | bias | var | cvg | bias | var | cvg | bias | var | cvg |
| $\hat{\tau}_{\text{simp}}$ | -0.84 | 34 | 2.5 | -0.88 | 22 | 0.6 | -0.66 | 28 | 4.6 |
| $\hat{\tau}_{\text{fix}}$ | 0.00 | 52 | 95.2 | -0.06 | 87 | 92.7 | 0.59 | 121 | 47.8 |
| $\hat{\tau}_{\text{ran}}$ | 0.11 | 41 | 95.7 | 0.08 | 72 | 96.7 | 0.67 | 91 | 33.8 |
| $\hat{\tau}_{\text{cal}}$ | -0.02 | 31 | 94.7 | -0.08 | 42 | 93.9 | 0.33 | 39 | 84.0 |
| Scenario | 4 Logistic outcome Logistic PS | | | 5 Logistic outcome Probit PS | | | 6 Logistic outcome C-loglog PS | | |
| | bias | var | cvg | bias | var | cvg | bias | var | cvg |
| $\hat{\tau}_{\text{simp}}$ | -0.14 | 0.84 | 1.1 | -0.13 | 0.52 | 0.2 | -0.12 | 0.22 | 0.0 |
| $\hat{\tau}_{\text{fix}}$ | -0.10 | 0.80 | 8.9 | -0.11 | 1.79 | 21.4 | -0.01 | 0.86 | 93.9 |
| $\hat{\tau}_{\text{ran}}$ | -0.08 | 0.60 | 15.0 | -0.09 | 1.28 | 32.4 | 0.01 | 0.71 | 99.3 |
| $\hat{\tau}_{\text{cal}}$ | 0.00 | 0.81 | 95.5 | -0.01 | 1.08 | 95.7 | 0.06 | 0.78 | 86.1 |

Table 3: Balance Check

| | | simple | fixed | random | calibration |
|---|---|---|---|---|---|
| $X_1$ | Cluster 1 | 1.68 | -0.22 | 0.68 | 0.20 |
| | Cluster 2 | 1.21 | 0.10 | -0.41 | 0.10 |
| | Cluster 3 | 1.75 | -0.02 | 0.99 | 0.02 |
| | Cluster 4 | 0.86 | -0.04 | -1.05 | 0.02 |
| | Cluster 5 | -0.36 | 0.37 | -1.39 | 0.33 |
| | Whole Pop | 1.28 | -0.02 | -0.02 | 0 |
| $X_2$ | Cluster 1 | 0.48 | 0.02 | 0.30 | 0.03 |
| | Cluster 2 | 0.43 | 0.13 | -0.01 | 0.14 |
| | Cluster 3 | 0.73 | 0.01 | 0.46 | 0.02 |
| | Cluster 4 | 0.18 | -0.08 | -0.34 | -0.07 |
| | Cluster 5 | -0.57 | -0.39 | -1.53 | -0.44 |
| | Whole Pop | 0.39 | -0.003 | -0.001 | 0 |

Table 4: Results: estimate, variance estimate (ve) based on 500 bootstrap replicates, and 95% confidence interval (c.i.)

| | estimate | ve | 95% c.i. |
|---|---|---|---|
| simple | 8.78 | 2.11 | (5.94, 11.63) |
| fixed | 0.47 | 0.44 | (-0.83, 1.77) |
| random | 0.52 | 0.44 | (-0.77, 1.82) |
| calibration | 0.53 | 0.39 | (-0.71, 1.76) |