MULTIVARIATE BEHAVIORAL RESEARCH
2022, VOL. 57, NO. 6, 916–939



# Causal Inference with Multilevel Data: A Comparison of Different Propensity Score Weighting Approaches

Alvaro Fuentes, Oliver Lüdtke, and Alexander Robitzsch

Centre for International Student Assessment, Leibniz Institute for Science and Mathematics Education, Kiel, Germany

**ABSTRACT**

Propensity score methods are a widely recommended approach to adjust for confounding and to recover treatment effects with non-experimental, single-level data. This article reviews propensity score weighting estimators for multilevel data in which individuals (level 1) are nested in clusters (level 2) and nonrandomly assigned to either a treatment or control condition at level 1. We address the choice of a weighting strategy (inverse probability weights, trimming, overlap weights, calibration weights) and discuss key issues related to the specification of the propensity score model (fixed-effects model, multilevel random-effects model) in the context of multilevel data. In three simulation studies, we show that estimates based on calibration weights, which prioritize balancing the sample distribution of level-1 and (unmeasured) level-2 covariates, should be preferred under many scenarios (i.e., treatment effect heterogeneity, presence of strong level-2 confounding) and can accommodate covariate-by-cluster interactions. However, when level-1 covariate effects vary strongly across clusters (i.e., under random slopes), and this variation is present in both the treatment and outcome data-generating mechanisms, large cluster sizes are needed to obtain accurate estimates of the treatment effect. We also discuss the implementation of survey weights and present a real-data example that illustrates the different methods.

**KEYWORDS**

Causal inference; propensity scores; multilevel data; weighting; calibration weights

In the last decades, propensity score methods have received significant attention for estimating treatment effects with non-experimental, observational data in psychology and educational research (e.g., Morgan & Winship, 2014; Schafer & Kang, 2008). Propensity score methods aim to balance the distribution of observed covariates between the treatment and control group, in order to ensure that an estimated treatment effect is not due to differences in observed characteristics between the groups (Austin, 2011; Rosenbaum & Rubin, 1983). In practical applications, the balance of the covariate distributions is achieved by matching observations on the propensity score (Stuart, 2010), stratifying them according to quantiles of the propensity score (Lunceford & Davidian, 2004), or reweighting the sample using functions of the propensity score (Hirano et al., 2003). A relatively small propensity score literature focuses on designs where lower-level units (e.g., students, employees; level 1) are nested within higher-level units (e.g., classrooms, firms; level 2), and recommendations for the use of propensity score methods with multilevel data are still scarce (see Hong, 2015; Hong & Raudenbush, 2006; Kim & Seltzer, 2007; Leite et al., 2015, 2019; Thoemmes & West, 2011), particularly for data structures that are typical in psychological research.

The purpose of this article is to evaluate different propensity score weighting methods for estimating treatment effects in data that have a multilevel structure. We study multilevel scenarios in which individuals are nested in clusters and nonrandomly assigned to either a treatment or control condition (i.e., binary treatment variable) at the individual level (level 1). With treatment assignment at level 1, it is crucial to determine which level-1 and level-2 covariates are potential confounders. Thus, it has been shown in previous research that the propensity score model should take the multilevel structure into account (e.g., Arpino & Mealli, 2011; Li et al.,

2013). The present study focuses on propensity weighting methods, which can be easily combined with the sampling weights that are often included in the analysis of large-scale survey data (e.g., school achievement studies) to obtain a representative sample of the population (Dong et al., 2020; Stapleton, 2013). In three simulation studies, we compare traditional inverse probability weighting (IPW; i.e., weights determined by the inverse probability of receiving the treatment that was actually received) with two alternative methods that have been proposed to stabilize IPW estimators, particularly in scenarios with extreme weights: trimming IPW weights (Lee et al., 2011), and overlap weights (Li et al., 2018). Also, we evaluate two recently introduced versions of calibration weights (Kim et al., 2017; Yang, 2018), and a clustered estimator that estimates the treatment effect separately within each cluster (Li et al., 2013). Calibration weights have the attractive feature that they directly balance the distribution of level-1 and (unmeasured) level-2 covariates when determining the weights (see Hainmueller, 2012; Imai & Ratkovic, 2014).

In our review of these propensity weighting methods, we put particular emphasis on three issues. First, we discuss the ability of these methods to control for unmeasured level-2 confounders, the so-called "unmeasured context" problem (Arpino & Mealli, 2011). More specifically, we investigate how the estimation of the propensity scores (i.e., fixed-effects models or multilevel random-effects models) that are used to compute the different weights affect the performance of the different weighting methods. However, we assume that all relevant level-1 covariates are observed. Second, we evaluate the performance of the different methods under heterogeneous treatment effects, which may arise from interactions of the treatment with level-1 and/or level-2 covariates. Third, we clarify the role of level-1 covariate effects that vary across clusters (i.e., random slopes), and show that random slopes need to be present in the treatment as well as the outcome model in order to deteriorate estimates of the treatment effects.

It should be emphasized that in our discussion of causal inference with two-level data, we focus on the case that nonrandom treatment assignment occurs at level 1. In many research designs, clusters of individuals are selected to participate in different treatments (e.g., schools are assigned to different programs), and nonrandom treatment assignment occurs at the group level (level 2). With treatment assignment at level 2, potential confounder variables are located at the group level (e.g., school resources, student characteristics aggregated at the school level), and covariates at level 1 are not relevant. Methods for estimating treatment effects with cluster-level assignment are discussed in Hansen et al. (2014), Keele et al. (2020), and Page et al. (2020).

The article is organized as follows. We begin with a brief description of the potential outcomes framework. We then describe the role of the propensity score and its estimation in the context of multilevel data. Next, we compare different propensity score weighting methods, discuss how each achieves covariate balance, and review previous findings from the literature concerning their performance. We then present the results of three simulation studies. In Study 1, the treatment assignment mechanism is a multilevel random-intercept model, and the treatment effect is homogeneous in the population. In Study 2, the treatment assignment mechanism is again a multilevel random-intercept model, but we introduce a heterogeneous treatment effect and allow for endogeneity at the cluster level. A brief section then describes the role of covariate-by-cluster interactions, before Study 3 investigates the role of random slopes. Finally, we discuss the implementation of survey weights and present a real-data example that illustrates the different methods.

## Potential outcomes and ignorability assumption

Consider a two-level structure in which a sample of $N$ units is grouped into $J$ clusters (e.g., $N$ students grouped by the $J$ schools they attend), each with $n_j$ units indexed $i = 1, 2, \ldots, n_j$. We assume a binary treatment variable $T_{ij}$, such that $T_{ij} = 1$ if unit $i$ in cluster $j$ is treated and $T_{ij} = 0$ otherwise. In the potential outcomes framework (Imbens & Rubin, 2015), each unit has two potential outcomes: $Y_{ij}(1)$ is the potential outcome under the treatment condition ($T_{ij} = 1$), and $Y_{ij}(0)$ is the potential outcome under the control condition ($T_{ij} = 0$). We further assume that, for each unit, the observed outcome equals the potential outcome under the observed treatment status, i.e., $Y_{ij} = Y_{ij}(T_{ij})$, and thus write the observed outcomes as $Y_{ij} = Y_{ij}(1) \cdot T_{ij} + Y_{ij}(0) \cdot (1 - T_{ij})$.

The average treatment effect (ATE) is then defined as:

$$\tau = \mathrm{E}\big(Y_{ij}(1) - Y_{ij}(0)\big) = \mu_1 - \mu_0 \qquad (1)$$

where $\mu_1$ and $\mu_0$ are the average potential outcomes under the treatment and control status, respectively. Since none of the units are observed under both the treatment and control conditions simultaneously, the ATE is not identified without further assumptions (Holland, 1986). Non-experimental research proceeds by conditioning on a set of observed covariates so

that the two potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$ are independent of the treatment $T_{ij}$. More formally, let $\mathbf{X}_{ij}$ and $\mathbf{V}_j$ denote vectors of level-1 and level-2 covariates. In the context of our study, it is instructive to further decompose the level-2 covariates $\mathbf{V}_j$ into an observed part $\mathbf{Z}_j$ and an unobserved part $\mathbf{W}_j$, i.e., $\mathbf{V}_j = (\mathbf{Z}_j, \mathbf{W}_j)$. If contextual effects of level-1 covariates are present, the observed part $\mathbf{Z}_j$ also includes the corresponding cluster means of the level-1 covariates. It can be shown that the ATE is identified under the ignorability assumption (see Rosenbaum & Rubin, 1983):

$$Y_{ij}(1), Y_{ij}(0) \perp T_{ij} | \mathbf{X}_{ij}, \mathbf{V}_j \tag{2}$$

which states that the potential outcomes are independent of the treatment given the covariates. This assumption is also labeled the unconfoundedness, conditional independence or selection on observables assumption in the literature (Hernán & Robins, 2020; Imbens, 2004; Morgan & Winship, 2014).[1] Because the ignorability assumption in Equation (2) also involves the unobserved cluster-level variables $\mathbf{W}_j$, Yang (2018) used the term *latent ignorability*. Note that the vector of observed cluster-level variables $\mathbf{Z}_j$ can also include information about cluster membership, and that cluster indicator variables can be used to represent the effects of unobserved cluster level confounders, as will be discussed in the next section.

The goal is then to estimate the ATE from the data ($Y_{ij}$, $T_{ij}$, $\mathbf{X}_{ij}$, $\mathbf{V}_j$). If the ignorability assumption in Equation (2) holds, the ATE can be identified as follows:

$$\tau = \mathrm{E}\big[\mathrm{E}\big(Y_{ij}|\mathbf{X}_{ij}, \mathbf{V}_j, T_{ij} = 1\big) - \mathrm{E}\big(Y_{ij}|\mathbf{X}_{ij}, \mathbf{V}_j, T_{ij} = 0\big)\big]. \tag{3}$$

The expected values of the potential outcomes in the treatment and control conditions (i.e., $\mu_1$ and $\mu_0$) can be determined by averaging the conditional expectation of the outcome given the observed covariates and the treatment status across the covariate distribution. Thus, the ignorability assumption ensures that the ATE can be estimated from the observed data. However, it should be emphasized that the ignorability assumption cannot be empirically tested and needs to be justified by substantive knowledge (Aronow & Miller, 2019). In this article, we assume that ignorability holds at level 1 (i.e., all important covariates at level 1 were measured) and focus on the role of level-2 covariates.

[1] A second assumption is needed (positivity assumption; see Rosenbaum & Rubin, 1983) which states that in the population the probability of receiving the treatment given the covariates is between 0 and 1, i. e. $0 < \mathrm{P}(T_{ij} = 1 | \mathbf{X}_{ij}, \mathbf{V}_j) < 1$. This assumption implies that there exists sufficient overlap in the covariate distributions between the treatment and control groups.

It is often reasonable to assume that the treatment effect varies across different subgroups, in which case the conditional ATE (CATE; Imbens, 2004) defines the ATE conditional on covariate values (i.e., $\mathbf{X}_{ij} = \mathbf{x}$ and $\mathbf{V}_j = \mathbf{v}$):

$$\tau_{\mathrm{CATE}}(\mathbf{x}, \mathbf{v}) = \mathrm{E}\big(Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}, \mathbf{V}_j = \mathbf{v}, T_{ij} = 1\big) - \mathrm{E}\big(Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}, \mathbf{V}_j = \mathbf{v}, T_{ij} = 0\big). \tag{4}$$

If $\tau_{\mathrm{CATE}}(\mathbf{x}, \mathbf{v})$ is a constant function of the covariate values, the ATE is said to be *homogeneous*; otherwise, the treatment effect is labeled as *heterogeneous* (e.g., Morgan & Winship, 2014). Note that the ATE in Equation (3) is obtained by averaging the CATE across the covariate distribution, i.e., $\mathrm{E}[\tau_{\mathrm{CATE}}(\mathbf{X}_{ij}, \mathbf{V}_j)] = \tau$.

## Propensity scores

A useful summary measure of the covariates is the propensity score, defined as the conditional probability of treatment given the covariates:

$$\pi_{ij} = \pi_{ij}(\mathbf{X}_{ij}, \mathbf{V}_j) = \mathrm{P}\big(T_{ij} = 1|\mathbf{X}_{ij}, \mathbf{V}_j\big) \tag{5}$$

Rosenbaum and Rubin (1983) showed that it suffices to condition on the propensity score, rather than on the covariates themselves, in order to fulfill the ignorability assumption that the potential outcomes are independent of the treatment:

$$Y_{ij}(1), Y_{ij}(0) \perp T_{ij} | \pi_{ij}(\mathbf{X}_{ij}, \mathbf{V}_j) \tag{6}$$

In practice, the propensity scores $\pi_{ij}$ have to be estimated from data, and previous research has consistently emphasized the importance of taking the multilevel structure into account at this estimation stage (e.g., Arpino & Mealli, 2011; Li et al., 2013; Steiner et al. 2013; Thoemmes & West, 2011). To this end, propensity score estimates are typically obtained with either the logistic fixed-effects or logistic random-effects specifications from the multilevel modeling literature. The two approaches mainly differ in how they deal with the effects of unobserved confounders at the cluster level. To further describe these two methods, we introduce the following multilevel logistic random-intercept model as a data-generating mechanism for the propensity scores (Snijders & Bosker, 2012):

$$g(\pi_{ij}) = g\big(\mathrm{P}\big(T_{ij} = 1|\mathbf{X}_{ij}, \mathbf{Z}_j, \mathbf{W}_j\big)\big) = \gamma_0 + \mathbf{X}_{ij}\boldsymbol{\gamma}_{\mathbf{X}} + \mathbf{Z}_j\boldsymbol{\gamma}_{\mathbf{Z}} + \mathbf{W}_j\boldsymbol{\gamma}_{\mathbf{W}} + U_{0j} \tag{7}$$

where $\gamma_0$ is the intercept, $\boldsymbol{\gamma}_{\mathbf{X}}$ are the effects of the covariates at the individual level, $\boldsymbol{\gamma}_{\mathbf{Z}}$ and $\boldsymbol{\gamma}_{\mathbf{W}}$ are the effects of the observed and unobserved covariates at

the cluster level, and $g$ denotes the logit link function. The random effects $U_{0j}$ are assumed to have zero mean and are uncorrelated with the covariates: $\text{Cov}(\mathbf{W}_j, U_{0j}) = 0$, $\text{Cov}(\mathbf{X}_{ij}, U_{0j}) = 0$, and $\text{Cov}(\mathbf{Z}_j, U_{0j}) = 0$. Note that $\mathbf{W}_j$ is not observed (e.g., unmeasured school resources) and has the potential to distort the estimation of treatment effects. Also note that we make the simplifying assumption that the effects of the level-1 covariates are constant across clusters (i.e., no random slopes). In the later section "Extension to Models with Covariate-by-Cluster Interactions," we discuss the more general case of treatment assignment models in which the effects of level-1 covariates vary across clusters.

In the fixed-effects modeling approach, a logistic regression model is specified for estimating the propensity scores:

$$g(\pi_{ij}) = \gamma_{0,\text{FE}} + \mathbf{X}_{ij}\boldsymbol{\gamma}_{\mathbf{X},\text{FE}} + U_{0j,\text{FE}} \tag{8}$$

Here, $U_{0j,\text{FE}}$ are cluster-specific effects, estimated by introducing a set of cluster-specific dummy variables that take values of 1 when a unit belongs to the cluster and 0 otherwise (Allison, 2009). The parameter estimates $\hat{\boldsymbol{\gamma}}_{\mathbf{X},\text{FE}}$ and $\hat{U}_{0j,\text{FE}}$ are then used to compute predicted probabilities of treatment $\hat{\pi}_{ij,\text{FE}} = g^{-1}(\hat{\gamma}_{0,\text{FE}} + \mathbf{X}_{ij}\hat{\boldsymbol{\gamma}}_{\mathbf{X},\text{FE}} + \hat{U}_{0j,\text{FE}})$. Previous simulation studies (e.g., Arpino & Mealli, 2011) have shown that the fixed-effects approach is able to remove confounding at the cluster level. This has the advantage that researchers do not need to measure the relevant level-2 covariates. However, with small cluster sizes (e.g., 10 level-1 units per level-2 unit), the estimated fixed effects can yield extreme predicted probabilities and unstable results (Li et al., 2013).

In the random-effects modeling approach, a multilevel logistic random-intercept model is specified for estimating the propensity scores:

$$g(\pi_{ij}) = \gamma_{0,\text{RE}} + \mathbf{X}_{ij}\boldsymbol{\gamma}_{\mathbf{X},\text{RE}} + \mathbf{Z}_j\boldsymbol{\gamma}_{\mathbf{Z},\text{RE}} + U_{0j,\text{RE}} \tag{9}$$

The random intercepts $U_{0j,\text{RE}}$ are assumed to be normally distributed. Because the unobserved cluster-level variables $\mathbf{W}_j$ are not included in the model, the random intercepts $U_{0j,\text{RE}}$ will, in general, be correlated with the level-1 and level-2 variables, i.e., $\text{Cov}(\mathbf{X}_{ij}, U_{0j,\text{RE}}) \neq 0$, and $\text{Cov}(\mathbf{Z}_j, U_{0j,\text{RE}}) \neq 0$. Thus, the random-effects model will be misspecified in the presence of unknown group-level confounders (Ebbes et al., 2004). However, for larger cluster sizes (e.g., 50 or larger), the estimated slopes of level-1 covariates and the estimated random intercepts of the random-effects model approximate the estimates of the fixed-effects model (e.g., Kreft & de Leeuw, 1998); this yields predicted probabilities $\hat{\pi}_{ij,\text{RE}}$ that converge against $\hat{\pi}_{ij,\text{FE}}$. We now discuss how the estimated probabilities are used to construct weighting estimators of the treatment effect.

## Propensity score weighting estimators

The estimated propensity scores, $\hat{\pi}_{ij}$, are used to compute weights, $\hat{\omega}_{ij}$, which are in turn used to construct estimators of the treatment effect. In general, the propensity score weighting estimators are of the form (Li et al., 2018):

$$\hat{\tau} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij}T_{ij}Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij}T_{ij}} - \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij}(1-T_{ij})Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij}(1-T_{ij})} \tag{10}$$

These are weighted averages of the outcome among the treated and non-treated units, where the first term is an estimate of $\mu_1$, and the second term estimates $\mu_0$ (see Equation (1)). Note that the estimated weights $\hat{\omega}_{ij}$ are functions of the observed covariates. If the propensity score model is correctly specified, the following balancing conditions are asymptotically fulfilled for the covariates:

$$\begin{aligned} \text{E}(\hat{\omega}_{ij}T_{ij}\boldsymbol{X}_{ij}) &= \text{E}(\hat{\omega}_{ij}(1-T_{ij})\boldsymbol{X}_{ij}) = \text{E}(\boldsymbol{X}_{ij}) \\ \text{E}(\hat{\omega}_{ij}T_{ij}\boldsymbol{V}_j) &= \text{E}(\hat{\omega}_{ij}(1-T_{ij})\boldsymbol{V}_j) = \text{E}(\boldsymbol{V}_j) \end{aligned} \tag{11}$$

In words, the weights create a pseudopopulation in which the treatment indicator $T_{ij}$ is independent of the covariates. In empirical applications, these balancing conditions are checked by comparing the weighted means of the observed covariates across the treatment and control conditions (e.g., Imbens & Rubin, 2015). However, it needs to be pointed out that balance does not imply that the ignorability assumption holds because balance on observed variables does not imply balance on unobserved variables (i.e., unmeasured confounder variables at level 1 or level 2).

For the inverse probability weighting (IPW) estimator, $\hat{\tau}_{\text{IPW}}$, the weights are defined as:

$$\hat{\omega}_{ij,\text{IPW}} = \begin{cases} 1/\hat{\pi}_{ij} & \text{for } T_{ij} = 1 \\ 1/(1-\hat{\pi}_{ij}) & \text{for } T_{ij} = 0 \end{cases}. \tag{12}$$

The weight of each unit is the inverse of the probability of assignment to the condition it was assigned to. As a result, individuals who are very unlikely to be assigned to treatment are upweighted in the treatment condition and downweighted in the control condition, and vice versa.

Though the weights in Equation (12) are widely used, they are known to exhibit large variability, especially in the case of small to moderate samples and

when the distribution of the covariates strongly differs between the treatment and control groups (Cole & Hernán, 2008; Harder et al., 2010). Indeed, while treated units with small propensities and control units with large propensities make for ideal counterfactual-type comparisons, their estimated probability values can be too extreme, in the sense that some of the weights implied are unreasonably large. Trimmed weights have been proposed to stabilize the IPW estimator (Crump et al., 2009; Lee et al., 2011)[2] by choosing a cutoff value $c$ and setting all weights larger than the cutoff value to zero:

$$\hat{\omega}_{ij,\,\mathrm{IPW-T}} = \hat{\omega}_{ij,\,\mathrm{IPW}} \mathbf{1}_{\{\hat{\omega}_{ij,\,\mathrm{IPW}} < c\}}, \tag{13}$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. The trimmed weight $\hat{\omega}_{ij,\,\mathrm{IPW-T}}$ equals $\hat{\omega}_{ij,\,\mathrm{IPW}}$ if $\hat{\omega}_{ij,\,\mathrm{IPW}}$ is smaller than $c$, and it is 0 otherwise. Trimmed weights are then used to obtain an IPW trimming estimator $\hat{\tau}_{\mathrm{IPW-T}}$ for the ATE as in Equation (10). Importantly, trimming implies a redefinition of the causal estimand (i.e., ATE in Equation (1)), since $\hat{\tau}_{\mathrm{IPW-T}}$ is aimed at a different target population: the population of units with mild probabilities of treatment. For example, with $c = 10$, only units with propensity scores that lie between .10 and .90 are considered. Thus, different trimming parameters lead to different target distributions, leaving it up to the analyst to choose an appropriate cutoff. Still, this is not a problem under homogenous treatment effects, that is, treatment effects that are constant across the distribution of the covariates.

A more principled approach uses overlap weights proposed by Li et al. (2018):

$$\hat{\omega}_{ij,\,\mathrm{OW}} = \begin{cases} 1 - \hat{\pi}_{ij} & \text{for } T_{ij} = 1 \\ \hat{\pi}_{ij} & \text{for } T_{ij} = 0 \end{cases} \tag{14}$$

for the treated and control units, respectively. The overlap weights upweight units with propensity scores close to .5, and downweight units with extreme propensity scores instead of completely discarding them. The estimator obtained from the overlap weights, $\hat{\tau}_{\mathrm{OW}}$, therefore focuses on the overlapping area of the propensity distributions of the treated and control samples. While this also redefines the target population, the overlap is often a meaningful area on the support of the propensity score, as it represents a subpopulation that had nontrivial probabilities for both being among the treated and the controls (Mao et al., 2019). Li et al. (2018; see also Li, Thomas, & Li, 2019) show that $\hat{\tau}_{\mathrm{OW}}$ has two desirable features: weighting by overlap weights achieves an exact balance of the covariates between treatment and control groups, and $\hat{\tau}_{\mathrm{OW}}$ achieves minimum asymptotic variance under certain conditions. However, to the best of our knowledge, the performance of overlap weights has not been investigated in the context of multilevel data.

[2]Alternatively, truncation or winsorization of weights has been proposed (Leite, 2016). Instead of discarding units with extreme weights, truncation assigns the cut-off value to units with weights above the cut-off (e.g., all units with weights larger than $c = 10$ obtain a weight of 10). In Simulation Study 1, we also applied truncation and found that it did not substantially improve the performance of the IPW estimator.

### Calibration estimator

Importantly, if the ignorability assumption holds (see Equation (2)), any set of weights that yields a pseudopopulation where the balance conditions of Equation (11) are fulfilled will result in an unbiased estimator of the average treatment effect, regardless of how the weights are obtained (i.e., whether they are constructed from estimates of the propensity score or not). The basic idea of calibration weights is to directly incorporate these balancing conditions in the construction of the weights. Hainmueller (2012) and Imai and Ratkovic (2014) were among the first to exploit calibration conditions to construct weights in the single-level literature and, more recently, Kim et al. (2017) and Yang (2018) extended these ideas to settings with clustered data. Specifically, calibration weights $\hat{\omega}_{ij,\,\mathrm{CAL}}$ must fulfill sample analogs of the balancing conditions for the level-1 covariates

$$\begin{aligned} \sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\,\mathrm{CAL}} T_{ij} \mathbf{X}_{ij} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\,\mathrm{CAL}} (1 - T_{ij}) \mathbf{X}_{ij} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathbf{X}_{ij} \end{aligned} \tag{15}$$

and the level-2 covariates

$$\begin{aligned} \sum_{j=1}^{J} \mathbf{V}_j \sum_{i=1}^{n_j} \hat{\omega}_{ij,\,\mathrm{CAL}} T_{ij} &= \sum_{j=1}^{J} \mathbf{V}_j \sum_{i=1}^{n_j} \hat{\omega}_{ij,\,\mathrm{CAL}} (1 - T_{ij}) \\ &= \sum_{j=1}^{J} n_j \mathbf{V}_j \end{aligned} \tag{16}$$

Since not all variables in $\mathbf{V}_j$ are observed, the empirical balancing condition cannot be directly evaluated. Instead, a sufficient condition for Equation (16) is:

$$\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} T_{ij} = \sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} (1 - T_{ij}) = \sum_{i=1}^{n_j} 1 = n_j \quad (j = 1, ..., J) \tag{17}$$

Under this condition, the within-cluster sum of weights for treated units equals the within-cluster sum of weights for the controls, which equals the cluster size. The weighted sample is, therefore, a pseudopopulation in which the proportion of treated is constant across clusters, which implies that a cluster's treatment prevalence is uncorrelated with any level-2 confounders. The calibration estimator $\hat{\tau}_{\mathrm{CAL}}$ is obtained by inserting the weights $\hat{\omega}_{ij,\mathrm{CAL}}$ into Equation (10). We now show that the calibration estimator provides unbiased estimates of the ATE if the ignorability assumption is fulfilled.

### Unbiasedness of calibration estimators

Let us assume that a multilevel random-intercept model holds for the continuous outcome (see Yang, 2018):

$$Y_{ij}(t) = \beta_{0,t} + \mathbf{X}_{ij}\boldsymbol{\beta}_{\mathbf{X},t} + \mathbf{V}_j\boldsymbol{\beta}_{\mathbf{V},t} + U_{j,t} + e_{ij,t} (t = 0, 1) \tag{18}$$

where $\mathrm{E}(U_{j,t}) = \mathrm{E}(e_{ij,t}) = 0$. The random intercept is allowed to be correlated with $\mathbf{X}$ and $\mathbf{V}$, while residuals $e_{ij,t}$ are uncorrelated with these covariates. If the data-generating model in Equation (18) holds, it can be shown that the calibration estimator $\hat{\tau}_{\mathrm{CAL}}$ provides an unbiased estimate of the ATE. Let us denote by $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\mu}_{\mathbf{V}}$, the expected values of $\mathbf{X}$ and $\mathbf{V}$, respectively. The ATE is then obtained by inserting Equation (18) into Equation (1) and taking expectations

$$\begin{aligned}\tau &= \mathrm{E}\left[Y_{ij}(1) - Y_{ij}(0)\right] \\ &= \beta_{0,1} - \beta_{0,0} + \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\beta}_{\mathbf{X},1} - \boldsymbol{\beta}_{\mathbf{X},0}) + \boldsymbol{\mu}_{\mathbf{V}}(\boldsymbol{\beta}_{\mathbf{V},1} - \boldsymbol{\beta}_{\mathbf{V},0})\end{aligned} \tag{19}$$

When the ignorability condition of Equation (2) holds, and the balancing conditions of Equations (15) and (17) are met (also note that $\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} T_{ij} = N$ directly follows), we obtain for the first term in $\hat{\tau}_{\mathrm{CAL}}$ :

$$\begin{aligned}&\mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} T_{ij} Y_{ij}\right] \\ &= \mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} T_{ij} Y_{ij}(1)\right] \\ &= \mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,CAL} T_{ij} (\beta_{0,1} + \mathbf{X}_{ij}\boldsymbol{\beta}_{\mathbf{X},1} + \mathbf{V}_j\boldsymbol{\beta}_{\mathbf{V},1} + U_{j,1} + e_{ij,1})\right] \\ &= \mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j} (\beta_{0,1} + \mathbf{X}_{ij}\boldsymbol{\beta}_{\mathbf{X},1} + \mathbf{V}_j\boldsymbol{\beta}_{\mathbf{V},1})\right] \\ &= \mathrm{N}\left[\beta_{0,1} + \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\beta}_{\mathbf{X},1} + \boldsymbol{\mu}_{\mathbf{V}}\boldsymbol{\beta}_{\mathbf{V},1}\right].\end{aligned} \tag{20}$$

Hence, we arrive at

$$\mathrm{E}\left[\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} T_{ij} Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} T_{ij}}\right] = \beta_{0,1} + \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\beta}_{\mathbf{X},1} + \boldsymbol{\mu}_{\mathbf{V}}\boldsymbol{\beta}_{\mathbf{V},1} \tag{21}$$

Similarly, we *obtain* for the second term in the calibration estimator of the treatment effect:

$$\begin{aligned}&\mathrm{E}\left[\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} (1 - T_{ij}) Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} (1 - T_{ij})}\right] \\ &= \beta_{0,0} + \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\beta}_{\mathbf{X},0} + \boldsymbol{\mu}_{\mathbf{V}}\boldsymbol{\beta}_{\mathbf{V},0}\end{aligned} \tag{22}$$

Now, by subtracting Equation (22) from Equation (21), the ATE in Equation (19) is obtained, and hence, the treatment effect estimator based on calibration weights is unbiased, if the balancing *conditions* are correctly specified (i.e., all relevant level-1 covariates are included in Equation (15))

### Calibration estimators for multilevel data

Two approaches for computing calibration weights in multilevel-data settings are available in the literature. The two approaches differ in the number of parameters used to obtain weights that fulfill the balancing conditions. First, Kim et al. (2017) introduced the following calibration weights:

$$\hat{\omega}_{ij,\mathrm{CAL1}} = \begin{cases} 1 + n_{0j}\dfrac{\exp\left\{\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}\right\}}{\sum_{h=1}^{n_j} T_{hj}\exp\left\{\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}\right\}} & \text{for } T_{ij} = 1 \\ 1 + n_{1j}\dfrac{\exp\left\{-\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}\right\}}{\sum_{h=1}^{n_j}(1 - T_{hj})\exp\left\{-\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}\right\}} & \text{for } T_{ij} = 0 \end{cases} \tag{23}$$

where $n_{1j}$ and $n_{0j}$ are the number of treated and control units in the $j$th cluster, respectively, and $\hat{\boldsymbol{\lambda}}$ is a vector of coefficients for the level-1 covariates. In Appendix A it is shown how the estimation equations for $\hat{\omega}_{ij,\mathrm{CAL1}}$ are obtained from a multilevel random-intercept model that includes the balancing conditions as additional estimation constraints.

Yang (2018) followed a different approach, which uses an initial vector of weights $\omega^*_{ij}$ (e.g., weights constructed from the propensities of an initial working model) to arrive at the calibration weights:

$$\hat{\omega}_{ij,\mathrm{CAL2}} = \begin{cases} n_j\dfrac{\omega^*_{ij}\exp\left\{\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}_1\right\}}{\sum_{h=1}^{n_j}\omega^*_{hj}T_{hj}\exp\left\{\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}_1\right\}} & \text{for } T_{ij} = 1 \\ n_j\dfrac{\omega^*_{ij}\exp\left\{\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}_0\right\}}{\sum_{h=1}^{n_j}\omega^*_{hj}(1 - T_{hj})\exp\left\{\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}_0\right\}} & \text{for } T_{ij} = 0 \end{cases}, \tag{24}$$

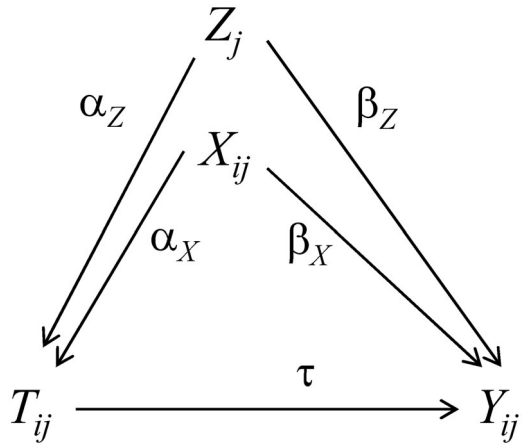


**Figure 1.** Schematic description of the data-generating model of simulation study 1.

where $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_0$ are vectors of coefficients for the level-1 covariates, obtained by minimizing a loss function (the Kullback-Leibler distance) subject to the balancing conditions in Equations (15) and (17) (see Appendix B).

In simulations, Kim et al. (2017) found their estimator $\hat{\tau}_{\text{CAL1}}$ to be superior to the $\hat{\tau}_{\text{IPW}}$ based on random-effects propensities in terms of both bias and variance in all conditions studied. Yang (2018) pit her estimator $\hat{\tau}_{\text{CAL2}}$ against $\hat{\tau}_{\text{IPW}}$ both with fixed and random-effects propensities and found that, in scenarios with a continuous outcome, it dominated both of the IPW estimators, though with a binary outcome the variance of the calibration estimator was often higher than that of $\hat{\tau}_{\text{IPW}}$ based on a fixed-effects propensity score.

### *Clustered estimator*

Another strategy to control for the confounding effect of level-2 covariates is to compute an estimate of the treatment effect within each cluster and then average those within-cluster estimates (Li et al., 2013). Such an estimator is equivalent to applying the following cluster-normalized weights:

$$\hat{\omega}_{ij,\text{CL}} = \begin{cases} n_j \dfrac{\omega_{ij,\text{IPW}}}{\sum_{h=1}^{n_j} T_{hj}\omega_{hj,\text{IPW}}} & \text{for } T_{ij} = 1 \\ n_j \dfrac{\omega_{ij,\text{IPW}}}{\sum_{h=1}^{n_j} (1 - T_{hj})\omega_{hj,\text{IPW}}} & \text{for } T_{ij} = 0 \end{cases}, \tag{25}$$

One major limitation of the clustered estimator $\hat{\tau}_{\text{CL}}$ is that only the level-2 balance condition (see Equation (17)) is fulfilled exactly, while the level-1 balance condition is only guaranteed asymptotically. Li et al. (2013) showed that, as the cluster size approaches infinity, the bias of $\hat{\tau}_{\text{CL}}$ vanishes. However, with small to moderate cluster sizes (15 to 50 level-1 units), biased estimates of the treatment effects can be obtained (Lee et al., 2019; see also Thoemmes & West, 2011).

We now turn to the results of three simulation studies, which provide a comprehensive evaluation of the different propensity score weighting estimators under various data-generating mechanisms of interest.

## Simulation study 1: homogeneous treatment effect and random intercepts

We begin with a simulation study in which both the treatment and the outcome data-generating mechanisms are random-intercept models, with a treatment effect that is constant across the support of the covariates (i.e., is homogeneous). In this scenario, it is straightforward to observe how the different propensity score weighting estimators deal with confounding information at both levels of analysis. Indeed, random-intercept simulation studies in the literature have shown that propensity scores obtained from a fixed-effects approach can be used to adjust for confounding at level 2, even if the confounding information is unobserved (Arpino & Mealli, 2011). In contrast, random-effects propensities are known to capture level-2 information less accurately, due to the shrinkage of posterior modes in multilevel models; as a consequence, random-effects propensities are deemed reliable only when clusters are large or all level-2 confounders are available (Leite et al., 2015). Conditioning within clusters, like with the cluster-normalized weights in Equation (25), automatically deals with level-2 confounding, but is known to require large clusters to account for confounding at level 1 (Li et al., 2013). To the best of our knowledge, trimmed and overlap weights have not been evaluated in the context of multilevel data, though one would expect their stabilization property to carry over to this setting. Finally, simulations by Kim et al. (2017) and Yang (2018) showed their calibration estimators to perform favorably in various conditions, and we expect them to also outperform the traditional IPW estimator with fixed-effects and random-effects propensities in this setup.

### *Method*

For the data-generating mechanisms, we specified a standardized and normally distributed covariate at level 1 ($X_{ij}$) and another at level 2 ($Z_j$), assumed to be

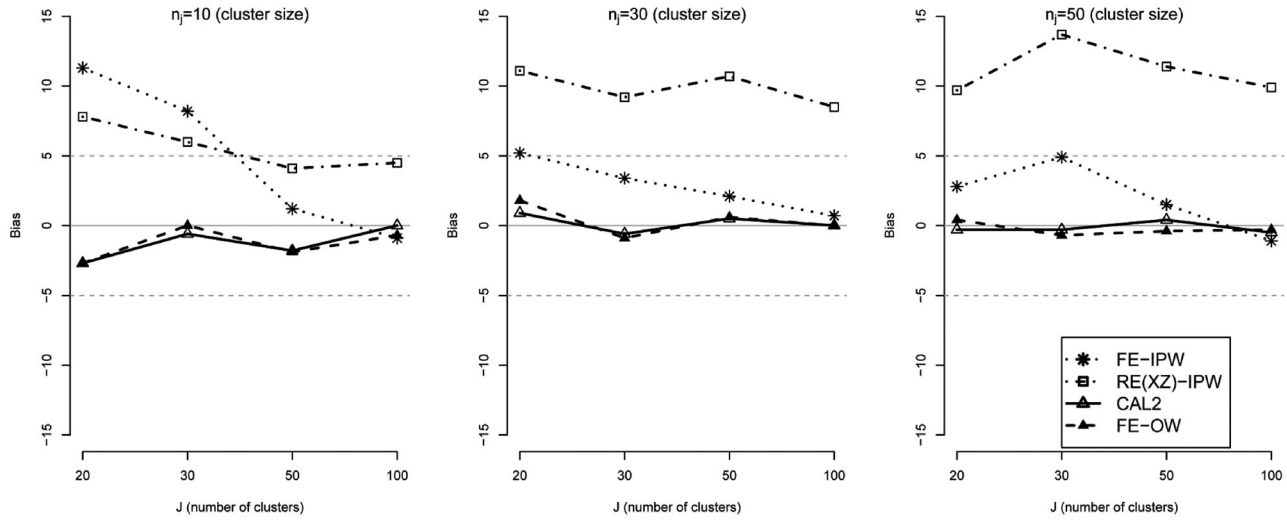


**Figure 2.** Relative Bias of different estimators of the treatment effect as a function of the number of clusters, and cluster sizes $n_j$ = 10 (left panel), $n_j$ = 30 (middle panel) and large cluster sizes $n_j$ = 50 (right panel). FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates $X$ and $Z$; IPW = inverse probability weighting; OW = overlap weights; CAL2 = calibration weights of Yang (2018).

uncorrelated (see Figure 1). Treatment assignment follows the multilevel logistic random-intercept model:

$$T_{ij}^{*} = \alpha_0 + \alpha_X X_{ij} + \alpha_Z Z_j + u_{0j} + \varepsilon_{ij} \qquad (26)$$

where an individual was assigned to treatment ($T_{ij}$ = 1) if $T_{ij}^{*} > 0$; $\alpha_X$, and $\alpha_Z$ are the regression coefficients; $u_{0j} \sim \mathrm{N}(0, \sigma_u^2)$ is the residual at level 2, and $\varepsilon_{ij} \sim$ Logistic(0, 1) is the residual at level 1. The intraclass correlation of $X$ ($\mathrm{ICC}_X$) was set to .2. The residual ICC of the treatment indicator was fixed to .2, that is, $\sigma_u^2/(\sigma_u^2 + \pi^2/3) = .2$. We fixed the intercept to zero ($\alpha_0 = 0$), which implies a treated-to-control ratio of 1:1. The explained variation in the treatment assignment model at level 1 is given by $R_{L1}^2 = \left[\alpha_X^2(1 - ICC_X)\right]/\mathrm{Var}_{\mathrm{total}}$, and at level 2 by $R_{L2}^2 = \left[\alpha_X^2 ICC_X + \alpha_Z^2\right]/\mathrm{Var}_{\mathrm{total}}$, where $\mathrm{Var}_{\mathrm{total}} = \alpha_X^2 + \alpha_Z^2 + \sigma_u^2 + \pi^2/3$ is the total variation of the treatment indicator (Snijders & Bosker, 2012; see Rights & Sterba, 2019).

The outcome follows a multilevel random-intercept model:

$$Y_{ij} = \beta_0 + \tau T_{ij} + \beta_X X_{ij} + \beta_Z Z_j + \nu_{0j} + e_{ij}, \qquad (27)$$

where $\beta_X$ and $\beta_Z$ are regression coefficients, and $\nu_{0j}$ and $e_{ij}$ are normally distributed residuals at level 2 and level 1, respectively. The treatment effect $\tau$ (ATE) was set to .30. The residual ICC of the outcome was set to .2, and the intercept was fixed to zero ($\beta_0 = 0$).

**Simulated conditions.** We specified four different conditions for the effect of the covariates on the treatment indicator and the outcome. In each condition, we assumed that the effects of the level-1 covariate and the level-2 covariate were equal for both the treatment and outcome equations, but manipulated the strength of confounding at level 1 and level 2: only confounding at level 2 ($\alpha_X = \beta_X = 0$ and $\alpha_Z = \beta_Z = 1$, which implies $R_{L1}^2 = 0$ and $R_{L2}^2 = .20$ for the treatment equation); only confounding at level 1 ($\alpha_X = \beta_X = .5$ and $\alpha_Z = \beta_Z = 0$, implying $R_{L1}^2 = .05$ and $R_{L2}^2 = .01$); confounding at both levels ($\alpha_X = \beta_X = .5$ and $\alpha_Z = \beta_Z = .5$, implying $R_{L1}^2 = .04$ and $R_{L2}^2 = .07$); and confounding at both levels with a stronger effect of the confounder at level 2 ($\alpha_X = \beta_X = .5$ and $\alpha_Z = \beta_Z = 1$, implying $R_{L1}^2 = .04$ and $R_{L2}^2 = .20$).

The number of clusters was set to $J$ = 50 and 100. Studies with about 50 groups are commonly found in educational and organizational psychology (e.g., Maas & Hox, 2005; Mathieu et al., 2012). The number of units per cluster was set to $n_j$ = 10, 20, 30, and 50. Group sizes of 10 are common in small-group research, whereas group sizes of 30 and 50 are typical of educational psychology research on class or school characteristics.

**Analysis model.** For each of the 4 (different effects of covariates) × 2 (number of clusters) × 4 (number of observations per cluster) = 32 conditions, 1,000 simulated data sets were generated. For each simulated data set, propensity scores were estimated by using three different models. First, to implement the fixed-effects (FE) approach, we specified a logistic regression model, including the level-1 covariate $X$ and a set of $J$–1 cluster-indicator variables (see Equation (8)). In addition, we implemented two variants of the random-effects (RE) approach by specifying two different multilevel logistic regression models:

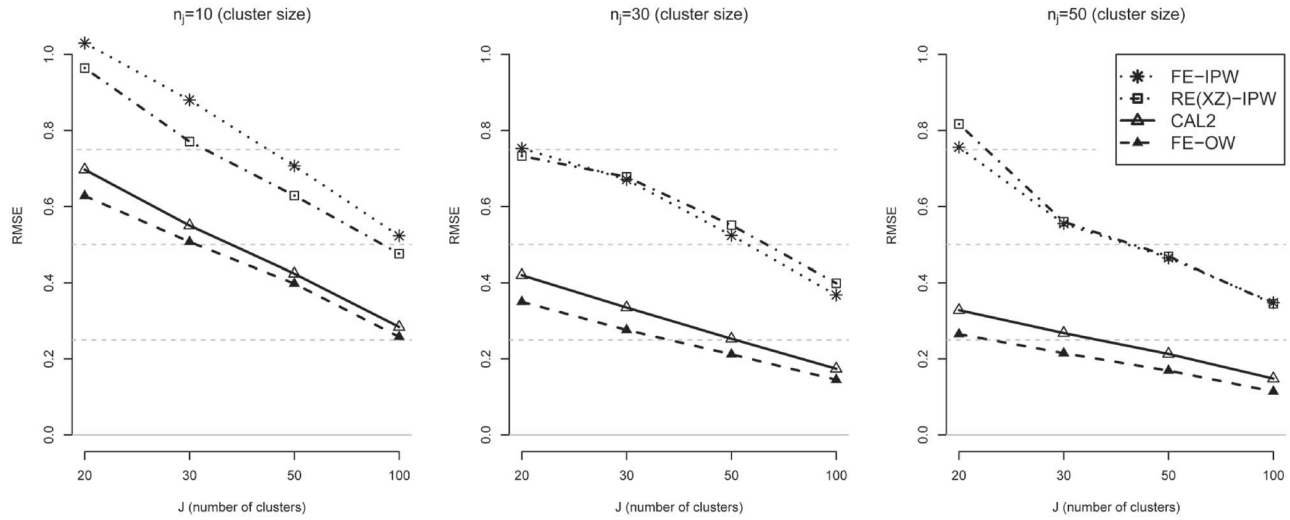


**Figure 3.** Relative RMSE of different estimators of the treatment effect as a function of the number of clusters, and cluster sizes $n_j = 10$ (left panel), $n_j = 30$ (middle panel) and large cluster sizes $n_j = 50$ (right panel). FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates $X$ and $Z$; IPW = inverse probability weighting; OW = overlap weights; CAL2 = calibration weights of Yang (2018).

a model that includes both covariates, labeled RE(XZ), and a model that only includes the level-1 covariate, labeled RE(X). Note that in the presence of a level-2 confounder (i.e., conditions in which $Z$ has an effect), the RE(X) model is misspecified. Logistic regression models were estimated with the glm function, and the multilevel logistic regression models were specified in the lme4 package (using the glmer function).

The propensity score predictions from these models were used to construct IPW weights (see Equation (12)), trimmed IPW weights (IPW-T; see Equation (13)), overlap weights (OW; see Equation (14)), and cluster-normalized weights (CL; see Equation (25)). For the trimmed weights, we applied a cutoff value of $c = 20$, including only cases with propensity scores that lie between .05 and .95 (Crump et al., 2009).[3] Thus, 3 (propensity score model) × 4 (type of weights) = 12 different estimators of the ATE were computed by substituting the various weights into Equation (10). Additionally, we implemented the two calibration estimators that were proposed by Kim et al. (2017; see Equation (23)) and Yang (2018; see Equation (24)). In total, 14 estimators of the ATE were compared. The R code for the data-generating model and the different analysis models is provided in Supplements S1, S2, S3, and S4 at https://doi.org/10.17605/OSF.IO/3FERB.

[3]We also computed trimmed weights with cut-off values of $c = 100$, and 10. As expected, higher cut-off values resulted in less biased but more variable estimates. We only report the results for $c = 20$ because they provided a reasonable trade-off between bias and variance.

Note that clusters in which all units had the same treatment status were discarded prior to estimation. The probability of simulating clusters in which all units had the same treatment status is higher for conditions with small cluster sizes, and conditions with strong confounding at level 2.

**Evaluation criteria**. We used two criteria to evaluate the different weighting approaches: relative bias and root mean square error (RMSE). Relative bias was calculated by dividing the empirical raw bias (the difference between the mean parameter estimate and the true population parameter value from each design cell) by the true parameter value. Relative bias of less than .05 in magnitude was considered acceptable and is referred to as approximately unbiased. We assessed the overall accuracy with the (empirical) RMSE, which combines the squared empirical relative bias and variance of the parameter estimates into a measure of overall accuracy.

## Results

Table 1 presents the relative bias and relative RMSE of the 14 different estimators of the ATE for a large number of clusters ($J = 100$; see Supplement S5 for full results). The weighting estimators that rely on FE propensity scores yielded approximately unbiased estimates, even under conditions with a strong confounding influence of the covariate $Z$ at level 2 (i.e., $\alpha_Z = \beta_Z = 1$). In contrast, the estimators based on the RE propensity scores produced biased estimates, particularly in the condition with a strong level-2 confounder and when the misspecified multilevel logistic model is

**Table 1.** Simulation study 1: relative bias and relative RMSE as a function of strength of level-1 and level-2 confounder effects and cluster size for a large number of groups (J = 100).

| | | Bias | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Weight | (0,1) | (.5,0) | (.5,.5) | (.5,1) | (0,1) | (.5,0) | (.5,.5) | (.5,1) |
| $J = 100$, $n_j = 10$ | | | | | | | | | |
| FE | IPW | 0 | −2 | 0 | −2 | .29 | .33 | .41 | .51 |
| FE | IPW-T | 0 | 1 | 3 | −1 | .29 | .27 | .31 | .34 |
| FE | OW | 0 | 1 | 2 | −1 | .27 | .23 | .25 | .25 |
| FE | CL | 0 | **14** | **19** | **19** | .28 | .31 | .35 | .37 |
| | CAL1 | 0 | 3 | **9** | **15** | .28 | .26 | .32 | .36 |
| | CAL2 | 1 | 1 | 3 | −1 | .28 | .24 | .27 | .28 |
| RE(XZ) | IPW | −5 | **10** | **13** | 4 | .31 | .26 | .33 | .49 |
| RE(XZ) | IPW-T | −2 | **8** | **9** | 5 | .29 | .25 | .30 | .32 |
| RE(XZ) | OW | −5 | −4 | −4 | **−8** | .28 | .23 | .25 | .27 |
| RE(XZ) | CL | 0 | **31** | **36** | **37** | .28 | .39 | .45 | .47 |
| RE(X) | IPW | **80** | **10** | **57** | **102** | .85 | .26 | .63 | 1.07 |
| RE(X) | IPW-T | **80** | **8** | **53** | **94** | .85 | .25 | .59 | .99 |
| RE(X) | OW | **68** | −4 | **36** | **67** | .73 | .23 | .43 | .72 |
| RE(X) | CL | 0 | **31** | **37** | **39** | .28 | .39 | .45 | .48 |
| $J = 100$, $n_j = 30$ | | | | | | | | | |
| FE | IPW | 1 | 0 | 0 | 2 | .18 | .18 | .24 | .36 |
| FE | IPW-T | 1 | 0 | 0 | 1 | .17 | .15 | .17 | .21 |
| FE | OW | 1 | 0 | 0 | 0 | .15 | .13 | .14 | .15 |
| FE | CL | 1 | **6** | **8** | **16** | .18 | .17 | .19 | .27 |
| | CAL1 | 1 | 1 | −1 | 0 | .18 | .14 | .16 | .20 |
| | CAL2 | 1 | 0 | −1 | 0 | .18 | .13 | .15 | .18 |
| RE(XZ) | IPW | 2 | **6** | **11** | **8** | .23 | .16 | .22 | .41 |
| RE(XZ) | IPW-T | **10** | 3 | 3 | 3 | .19 | .14 | .17 | .20 |
| RE(XZ) | OW | −2 | −3 | −4 | −5 | .15 | .13 | .15 | .16 |
| RE(XZ) | CL | 1 | **11** | **14** | **22** | .18 | .18 | .21 | .30 |
| RE(X) | IPW | **53** | **6** | **30** | **62** | .55 | .16 | .35 | .68 |
| RE(X) | IPW-T | **53** | 3 | **21** | **37** | .55 | .14 | .26 | .42 |
| RE(X) | OW | **28** | −3 | **12** | **26** | .31 | .13 | .19 | .30 |
| RE(X) | CL | 1 | **11** | **14** | **22** | .17 | .18 | .21 | .30 |

*Note.* $J$ = number of clusters; $n_j$ = cluster size; FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates $X$ and $Z$; RE(X) = random-effects propensity scores with covariate $X$; IPW = inverse probability weighting; IPW-T = inverse probability weighting with trimming; OW = overlap weights; CL = cluster-normalized IPW; CAL1 = calibration weights of Kim et al. (2017); CAL2 = calibration weights of Yang (2018). Relative biases smaller than −5 or larger than 5 are printed in bold.

used for estimating propensity scores, i.e., RE(X). However, even the estimates based on the correctly-specified multilevel logistic model, i.e., RE(XZ), were slightly biased, particularly in conditions with small cluster sizes. The two estimators based on the calibration weights (CAL1 and CAL2) provided approximately unbiased estimates of the ATE, with the exception that CAL1 was slightly positively biased in conditions with small cluster sizes. Finally, the estimator with cluster-normalized (CL) weights, which averages the cluster-specific estimates, produced strongly biased estimates of the ATE whenever level-1 confounding was present, even in the scenarios with 30 units per cluster.

In terms of RMSE, we found for both the FE and the RE propensity scores that the estimators based on IPW weights resulted in more variable estimates of the ATE, particularly in conditions with confounding at both levels. Consistent with results for single-level data, trimming (i.e., IPW-T) units with extreme

**Table 2.** Simulation study 1: Relative bias and relative RMSE as a function of cluster size and proportion treated for a large number of groups (J = 100).

| | | | Bias | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Weight | $n_j$ | 10 | 20 | 30 | 50 | 10 | 20 | 30 | 50 |
| *10% Treated* | | | | | | | | | | |
| FE | IPW | | 2 | 4 | 0 | 4 | .79 | .65 | .58 | .52 |
| FE | IPW-T | | 0 | 3 | 0 | 0 | .57 | .38 | .32 | .25 |
| FE | OW | | −1 | 2 | −1 | 0 | .43 | .30 | .23 | .17 |
| | CAL1 | | **40** | **19** | **6** | 1 | .67 | .47 | .37 | .30 |
| | CAL2 | | **24** | **11** | 2 | 0 | .59 | .43 | .36 | .30 |
| RE(XZ) | IPW | | **−17** | **−25** | **−32** | **−27** | .79 | .80 | .84 | .76 |
| RE(XZ) | IPW-T | | −2 | 0 | **−6** | **−6** | .54 | .38 | .32 | .26 |
| RE(XZ) | OW | | −5 | −5 | **−7** | **−6** | .44 | .30 | .24 | .18 |
| RE(X) | IPW | | **99** | **71** | **56** | **49** | 1.13 | .87 | .73 | .64 |
| RE(X) | IPW-T | | **90** | **54** | **34** | **19** | 1.03 | .65 | .45 | .31 |
| RE(X) | OW | | **84** | **53** | **38** | **26** | .95 | .61 | .44 | .31 |
| *20% Treated* | | | | | | | | | | |
| FE | IPW | | 0 | 2 | 4 | 0 | .64 | .54 | .46 | .41 |
| FE | IPW-T | | 1 | 1 | 1 | 0 | .43 | .30 | .25 | .19 |
| FE | OW | | 1 | 1 | 0 | 0 | .33 | .23 | .18 | .14 |
| | CAL1 | | **27** | 3 | 1 | 0 | .50 | .31 | .27 | .22 |
| | CAL2 | | 3 | 0 | 1 | 0 | .40 | .29 | .25 | .21 |
| RE(XZ) | IPW | | **−9** | **−11** | **−9** | **−12** | .63 | .67 | .55 | .57 |
| RE(XZ) | IPW-T | | 1 | 2 | −1 | −3 | .42 | .29 | .24 | .19 |
| RE(XZ) | OW | | **−6** | **−7** | **−6** | −4 | .34 | .24 | .19 | .15 |
| RE(X) | IPW | | **100** | **72** | **60** | **47** | 1.09 | .81 | .70 | .57 |
| RE(X) | IPW-T | | **92** | **54** | **36** | **21** | 1.00 | .61 | .43 | .28 |
| RE(X) | OW | | **75** | **44** | **31** | **20** | .82 | .49 | .36 | .25 |
| *50% Treated* | | | | | | | | | | |
| FE | IPW | | 0 | 0 | 1 | 1 | .54 | .48 | .37 | .35 |
| FE | IPW-T | | 2 | 0 | 0 | 0 | .36 | .25 | .20 | .16 |
| FE | OW | | 0 | 0 | 0 | 0 | .28 | .19 | .15 | .12 |
| | CAL1 | | **16** | 1 | 0 | 0 | .37 | .23 | .20 | .17 |
| | CAL2 | | 1 | 0 | 0 | 0 | .30 | .21 | .18 | .15 |
| RE(XZ) | IPW | | 5 | **10** | **10** | **11** | .47 | .46 | .38 | .35 |
| RE(XZ) | IPW-T | | **6** | 4 | 2 | 1 | .34 | .24 | .20 | .15 |
| RE(XZ) | OW | | **−8** | **−7** | −5 | −4 | .29 | .21 | .16 | .12 |
| RE(X) | IPW | | **102** | **74** | **61** | **48** | 1.07 | .81 | .67 | .54 |
| RE(X) | IPW-T | | **95** | **55** | **37** | **21** | .99 | .60 | .41 | .26 |
| RE(X) | OW | | **67** | **37** | **26** | **16** | .73 | .42 | .30 | .20 |

*Note.* $J$ = number of clusters; $n_j$ = cluster size; FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates $X$ and $Z$; RE(X) = random-effects propensity scores with covariate $X$; IPW = inverse probability weighting; IPW-T = inverse probability weighting with trimming; OW = overlap weights; CAL1 = calibration weights of Kim et al. (2017); CAL2 = calibration weights of Yang (2018). Relative biases smaller than −5 or larger than 5 are printed in bold.

weights provided more stable estimates of the ATE. The overlap weights (OW) produced the most accurate estimates of the ATE in terms of RMSE. However, the two calibration estimators performed very similar to OW and were only outperformed in conditions with strong confounding at level-2. In line with the results for bias, the estimates with the CL weights were not very accurate in terms of RMSE.

We also investigated whether our results generalize to scenarios with treated-to-control ratios other than 1:1 (see Table 2). In an additional simulation, we generated data with confounding at both levels ($\alpha_X = \beta_X = .5$, $\alpha_Z = \beta_Z = 1$), various treated-to-control ratios (10%, 20% and 50% treated per cluster on average), and different cluster sizes (10, 20, 30, and 50 units per cluster). Under these conditions, OW, IPW-T, and the

two calibration weights (CAL1 and CAL2) outperformed the other estimators in terms of bias and RMSE, with the exception that in conditions with a very low proportion of treated units per cluster and a small cluster size CAL1 and CAL2 were positively biased.

Furthermore, in many research designs less than 50 clusters are included at level 2. Therefore, we conducted additional simulations in which we investigated the performance of the different approaches for smaller numbers of clusters. We evaluated for a selected condition of the main simulation (confounding at both levels and a stronger effect of the confounder at level 2; i.e., $\alpha_X = \beta_X = .5$ and $\alpha_Z = \beta_Z = 1$), the bias and RMSE of the different approaches as a function of the cluster sizes ($n_j = 10$, 30, and 50), and the number of clusters ($J = 20$, 30, 50, and 100). Figure 2 shows the bias and Figure 3 shows the RMSE as a function of the cluster sizes, and the number of clusters for four selected estimators that performed favorably in the main simulation (FE-IPW, RE-IPW, FE-OW, and CAL2; the full results for all estimators are presented in Supplement S6). Overall, the results show the main conclusions about the performance of the different estimators can be generalized to conditions with a smaller number of clusters. As can be seen, the estimator that is based on the overlap weights that are obtained from an FE propensity score model (FE-OW), and the estimator that is based on the calibration weights (CAL2) clearly outperformed the two estimators that are based on IPW weights that are obtained from an FE propensity score model (FE-IPW) or a RE propensity score model (RE-IPW) in terms of RMSE.[4]

### Summary and discussion

The main findings of the simulation can be summarized as follows. First, the estimators based on the FE propensity scores were able to adjust for confounders at level 2 and yielded suitable estimates of the ATE, as did the estimators based on the RE propensity scores from the correctly-specified multilevel model RE(XZ). Second, we confirmed the variance findings of the literature regarding IPW weights, which provided unstable estimates of the ATE in challenging data constellations (i.e., strong confounding). Third, the performance of the propensity score weighting estimators in terms of RMSE was nevertheless improved by trimming units with extreme weights (IPW-T) or downweighting units at the tails of the propensity score distribution (OW). Moreover, because in this study, the treatment effect is the same everywhere on the support of the propensity score, the variance gains of discarding or downweighting the troublesome tails come at no cost in terms of bias. Fourth, though the well-known variance advantage of RE over FE estimators is present in the simulation results, the variance difference practically vanishes when either of the weight-stabilization procedures is applied (IPW-T or OW). In the most demanding condition ($\alpha_X = \beta_X = .5$, $\alpha_Z = \beta_Z = 1$, $n_j = 10$), trimming, and downweighting units at the tails of the FE propensity score distribution yielded RMSE reductions of around 30% and 50%, respectively. Fifth, the calibration estimators performed best overall, although CAL1 showed some bias in scenarios with small clusters (i.e., $n_j = 10$) and strong confounding. Sixth, the misspecified RE(X) model produced weighting estimators that were substantially biased, especially in conditions with strong confounding at level 2, and the estimator with cluster-normalized weights was not able to control for the effects of a measured level-1 covariate. We, therefore, decided to leave out the RE(X) model and the cluster-normalized weights in the simulation studies of the next sections. Finally, we note that rare treatments (i.e., 10% treated units) can strain the calibration estimators, which required moderate or large cluster sizes (i.e., $n_j \geq 30$) for obtaining accurate estimates of the treatment effect.

## Simulation study 2: heterogeneous treatment effects and cluster-level endogeneity

In Study 2, we explore the impact of introducing treatment-covariate interactions, which, in the single-level literature, is a well-known issue when working with estimators that modify the target population (Li et al., 2013). Since trimmed and overlap weights disregard or downweight the tails of the propensity score distribution, one cannot hope to recover an ATE when the treatment effect changes across the support of the propensity score. Nevertheless, heterogeneous

[4]With real educational or organizational data, cluster sizes usually differ across clusters. To evaluate the robustness of the different estimators in the case of unbalanced cluster sizes, we conducted an additional, restricted simulation for a subset of the conditions of Simulation Study 1. We fixed the number of clusters ($J = 100$) and manipulated the confounding at level 1 and level 2. In the unbalanced condition, the cluster sizes were uniformly distributed across {5, 6, … ,15} with an average cluster size of 10. In the balanced condition, the cluster sizes were constant ($n = 10$). Overall, the results show that for some approaches (OW and IPW weights, and CAL2), the RMSE slightly increased in conditions with unbalanced clusters. However, all differences in RMSE (unbalanced vs. balanced condition) were below 5%, and the conclusions about the performance of the different approaches did not change with unbalanced cluster sizes (see for the results Supplement S7)."

treatment effects are common in psychological and educational research (see, e.g., Morgan & Winship, 2014). Additionally, by manipulating the correlation of the intercepts of the treatment and outcome data-generating models, this study further explores the behavior of estimators that are based on the RE propensities when some of the level-2 confounding information is unobserved. Such scenarios of "omitted context" (Arpino & Mealli, 2011) arise in practice whenever researchers fail to gather data on all relevant level-2 covariates. Previous studies suggest that the IPW estimator with FE propensities should also be able to handle the effects of unobserved level-2 confounders in this setting (Arpino & Mealli, 2011).

## Method

We adopted a simulation design of Kim et al. (2017) and specified the following data-generating mechanism for treatment assignment:

$$T_{ij}^{*} = \alpha_0 + \alpha_X X_{ij} + \alpha_Z Z_j + u_{0j} + \varepsilon_{ij}, \qquad (28)$$

where, again, the level-1 covariate $X_{ij}$ and the level-2 covariate $Z_j$ were specified to be independent, though now $Z_j \sim$ Unif(0, 1), while $X_{ij}$ follows a standard normal distribution with an $ICC_X$ of zero. The regression coefficients were set to $\alpha_0 = -1$, $\alpha_X = 0.7$, and $\alpha_Z = -0.8$, and the variance of the level-2 residual was set to 1. This implies explained variations at level 1 and level 2 of $R^2_{L1} = .10$, and $R^2_{L2} = .01$, respectively.

The outcome equation allows for heterogeneity of the treatment effect:

$$Y_{ij} = \beta_0 + (\tau_0 + \tau_1 X_{ij} + \tau_2 Z_j + \tau_3 v_{1j}) T_{ij} + \beta_X X_{ij} + \beta_Z Z_j + v_{0j} + e_{ij} \qquad (29)$$

Here, in line with Kim et al. (2017), we set the regression coefficients to $\beta_0 = .3$, $\beta_X = 1.3$, and $\beta_Z = -0.5$. The variance of the residual at level 2 was fixed to 1, and the variance of the residual at level 1 to 0.4. The random slope was set equal to the random intercept (i.e., $v_{1j} = v_{0j}$). In addition, we set $\tau_0$ =1.25 and manipulated the heterogeneity by setting the effect of the interactions to $\tau_1 = 0$, $\tau_2 = 0$, and $\tau_3 = 0$ (i.e., no treatment effect heterogeneity) or $\tau_1 = 1$, $\tau_2 = -0.5$, and $\tau_3 = 0.8$ (i.e., treatment effect heterogeneity). The true ATE is given by $\tau_0 + 0.5\tau_2$. Furthermore, we manipulated the endogeneity at level 2 by setting the correlation of the cluster-level residuals, $u_{0j}$ and $v_{0j}$, to either zero or one. Note that perfectly correlated residuals at level 2 imply omitted context variables that influence assignment to treatment and the outcome (Arpino & Mealli, 2011). The number of clusters was set to $J = 50$ and 100, and the number of units per cluster was set to $n_j$ = 10, 20, 30, and 50. The R code for the data-generating model is provided in Supplement S8.

For each of the 2 (treatment effect homogeneity vs. heterogeneity) × 2 (exogeneity vs. endogeneity at level 2) × 2 (number of clusters) × 4 (number of observations per cluster) = 32 conditions, 1,000 simulated data sets were generated. Eight different estimates of the treatment effect were computed: three estimates using the IPW, IPW-T, and OW weights based on FE propensity scores; three estimates that used the same weighting approaches (IPW, IPW-T, OW), but were based on propensity scores obtained from a multilevel logistic model including the covariates $X$ and $Z$ (i.e., RE(XZ)); and the two calibration estimators (i.e., CAL1 and CAL2). The implementation of the eight estimators was identical to Study 1. Again, we evaluated their performance through relative bias and RMSE statistics.

## Results

Table 3 presents the relative bias and RMSE of the different estimators with a large number of groups ($J = 100$; see Supplement S9 for the full results). Consistent with the results from Study 1, the weights constructed with FE propensities yielded approximately unbiased estimates of the ATE under conditions with level-2 endogeneity. However, when the treatment effect was heterogeneous, both the IPW-T weights that discard units with extreme weights and the OW weights that focus more on units in the middle range of the propensity score distribution produced strongly biased estimates of the ATE. The bias of IPW-T and OW was independent of the cluster size and of the model used to estimate propensity scores, i.e., FE or RE(XZ). The estimators based on propensity scores from a multilevel logistic regression, i.e., RE(XZ), produced biased estimates, particularly in conditions with level-2 endogeneity and treatment effect heterogeneity. This finding was expected since, due to shrinkage, the RE(XZ) model does a poor job of capturing the omitted group-level confounding introduced by the perfectly correlated random intercepts of the treatment assignment model and the outcome model. The two calibration estimators performed favorably under level-2 endogeneity and treatment effect heterogeneity, with the exception that both were positively biased in conditions with a small cluster size ($n_j$ = 10). Again, consistent with Study 2,

**Table 3.** Simulation study 2: Relative bias and relative RMSE as a function of level 2 endogeneity, treatment effect heterogeneity and cluster size for a large number of groups (J = 100).

| | | Bias | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Exo | | Endo | | Exo | | Endo | |
| Model | Weights | Hom | Het | Hom | Het | Hom | Het | Hom | Het |
| $J = 100$, $n_j = 10$ | | | | | | | | | |
| FE | IPW | 0 | 2 | 0 | **17** | .19 | .32 | .20 | .38 |
| FE | IPW-T | 1 | **12** | 0 | **32** | .16 | .27 | .16 | .41 |
| FE | OW | 0 | **23** | 0 | **52** | .14 | .30 | .14 | .55 |
| | CAL1 | 2 | **17** | 2 | **32** | .16 | .29 | .17 | .39 |
| | CAL2 | 1 | 3 | 0 | **19** | .16 | .22 | .17 | .28 |
| RE(XZ) | IPW | 4 | **9** | **33** | **81** | .16 | .25 | .37 | .84 |
| RE(XZ) | IPW-T | 3 | **9** | **32** | **81** | .15 | .24 | .35 | .84 |
| RE(XZ) | OW | −2 | **20** | **27** | **98** | .14 | .29 | .31 | 1.00 |
| $J = 100$, $n_j = 30$ | | | | | | | | | |
| FE | IPW | −1 | 0 | 0 | 5 | .13 | .21 | .13 | .23 |
| FE | IPW-T | 0 | **12** | 0 | **29** | .09 | .20 | .09 | .32 |
| FE | OW | 0 | **24** | 0 | **52** | .08 | .28 | .08 | .53 |
| | CAL1 | 0 | 5 | 0 | **9** | .10 | .16 | .10 | .17 |
| | CAL2 | 0 | 0 | 0 | 5 | .09 | .15 | .09 | .15 |
| RE(XZ) | IPW | 2 | 4 | **19** | **43** | .10 | .18 | .22 | .47 |
| RE(XZ) | IPW-T | −1 | **6** | **14** | **47** | .09 | .17 | .17 | .49 |
| RE(XZ) | OW | −2 | **22** | **12** | **72** | .08 | .26 | .14 | .73 |
| $J = 100$, $n_j = 50$ | | | | | | | | | |
| FE | IPW | 0 | 1 | 0 | 2 | .10 | .17 | .10 | .20 |
| FE | IPW-T | 0 | **13** | 0 | **28** | .07 | .18 | .07 | .31 |
| FE | OW | 0 | **25** | 0 | **53** | .06 | .28 | .06 | .53 |
| | CAL1 | 0 | 3 | 0 | 5 | .08 | .13 | .08 | .14 |
| | CAL2 | 0 | 0 | 0 | 2 | .07 | .12 | .07 | .13 |
| RE(XZ) | IPW | 2 | 4 | **14** | **31** | .08 | .15 | .16 | .35 |
| RE(XZ) | IPW-T | −1 | **8** | **9** | **38** | .07 | .15 | .11 | .40 |
| RE(XZ) | OW | −1 | **24** | **8** | **65** | .06 | .27 | .10 | .66 |

*Note.* Endo = endogeneity; Exo = exogeneity; Hom = treatment effect homogeneity; Het = treatment effect heterogeneity; $J$ = number of clusters; $n_j$ = cluster size; FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates $X$ and $Z$; IPW = inverse probability weighting; IPW-T = inverse probability weighting with trimming; OW = overlap weights; CAL1 = calibration weights of Kim et al. (2017); CAL2 = calibration weights of Yang (2018). Relative biases smaller than −5 or larger than 5 are printed in bold.

the estimator CAL2 slightly outperformed CAL1. The RMSE results closely paralleled the relative bias results, with the calibration estimators outperforming the others in scenarios of treatment effect heterogeneity.

### *Summary and discussion*

The main takeaways of this study are as follows. First, the IPW weights with FE propensities were able to deal with treatment effect heterogeneity, and control for unobserved level-2 confounding information. Note, however, that the combination of these two difficulties still resulted in bias in scenarios with small clusters (i.e., $n_j < 30$). Second, the IPW weights with RE propensities can also recover the ATE under treatment effect heterogeneity but is strongly biased under omitted context. Third, IPW-T and OW weights could not recover the ATE when the treatment effect is heterogeneous, since they focus on a different target distribution and only estimate the treatment effect for a subset of the support of the propensity score. Finally, as long as clusters were not too small, the estimators based on the calibration weights can estimate without bias under both heterogeneity and omitted context, and they achieved lower variance than the IPW weights with FE propensities in most conditions.

## Extension to models with covariate-by-cluster interactions

Our previous simulation studies assumed that the effects of the covariates in the treatment assignment were constant across clusters. More specifically, we assumed that a multilevel logistic random-intercept model describes the data-generating mechanism for treatment assignment (see Equation (7)). However, it is typically more realistic to expect that the effects of level-1 covariates on the probability of treatment assignment vary across clusters. A more general data-generating model for the treatment is given by the multilevel logistic model with random slopes:

$$g(\pi_{ij}) = \gamma_0 + \mathbf{X}_{ij}\boldsymbol{\gamma}_{\mathbf{X}} + \mathbf{V}_j\boldsymbol{\gamma}_{\mathbf{Z}} + \mathbf{X}_{ij}\mathbf{V}_j\boldsymbol{\gamma}_{\mathbf{XV}} + U_{0j} + \mathbf{X}_{ij}\mathbf{U}_{1j} \tag{30}$$

where $\mathbf{U}_{1j}$ is a vector of cluster-specific effects that allow the effects of $\mathbf{X}_{ij}$ to vary across clusters, and $\mathbf{X}_{ij}\mathbf{V}_j$ are the cross-level interactions between the level-1 covariates $\mathbf{X}_{ij}$ and the observed and potentially unobserved level-2 covariates $\mathbf{V}_j$. The model for the potential outcomes can also be extended to include covariate-by-cluster interactions (see Equation (18)):

$$\begin{aligned} Y_{ij}(t) = \beta_{0,t} + \mathbf{X}_{ij}\boldsymbol{\beta}_{\mathbf{X},t} + \mathbf{V}_j\boldsymbol{\beta}_{\mathbf{V},t} + \mathbf{X}_{ij}\mathbf{V}_j\boldsymbol{\beta}_{\mathbf{XV},t} + U_{0j,t} \\ + \mathbf{X}_{ij}\mathbf{U}_{1j,t} + e_{ij,t}(t = 0, 1), \end{aligned} \tag{31}$$

where $\mathrm{E}(\mathbf{U}_{1j,t}) = \mathbf{0}$, and $\mathrm{E}(U_{0j,t}) = \mathrm{E}(e_{ij,t}) = 0$. The random effects $U_{0j,t}$ and $\mathbf{U}_{1j,t}$ are allowed to be correlated with observed and unobserved covariates, while the residuals $e_{ij,t}$ have to be uncorrelated with covariates. When the effects of the level-1 covariates vary across clusters in the data-generating model for the treatment assignment as well as the outcome, it can be shown that the following balancing conditions need to be fulfilled for the calibration estimators:

$$\begin{aligned} \sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} T_{ij}\mathbf{X}_{ij} &= \sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}}(1 - T_{ij})\mathbf{X}_{ij} \\ &= \sum_{i=1}^{n_j} \mathbf{X}_{ij} \quad (j = 1, \ldots, J) \end{aligned} \tag{32}$$

$$\sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} T_{ij} = \sum_{i=1}^{n_j} \hat{\omega}_{ij,\mathrm{CAL}} (1 - T_{ij}) = \sum_{i=1}^{n_j} 1 = n_j \quad (j = 1, \ldots, J) \quad (33)$$

In Equation (32), it is specified that the distribution of the level-1 covariates is balanced within each cluster. This guarantees that the effects of the level-1 covariates, which vary by cluster in both the treatment and the outcome equations, do not distort estimates of the ATE. Equation (33) is identical to the random-intercept scenario (see Equation (17)) and expects that both the within-cluster sum of weights for treated units and the within-cluster sum of weights for the controls equal the cluster size. In Appendix C, it is shown that the calibration estimators produce unbiased estimates of the ATE when the ignorability assumption holds and the Equations (32) and (33) are fulfilled. However, large cluster sizes are likely to be needed to obtain stable estimates of the ATE with calibrations estimators, particularly when the number of covariates is not small. To the best of our knowledge, the calibration estimators have not been studied in scenarios with covariate effects that vary across clusters.

As was already mentioned, when estimating the propensity scores from multilevel data, researchers can select between an FE approach and a RE approach. In the case of covariate-by-cluster interactions, the FE approach is extended by including covariate-by-cluster interaction terms in the logistic regression model in Equation (8). However, estimating separate slopes for each cluster requires that the clusters be quite large, particularly with a larger number of level-1 covariates. Alternatively, an RE model can be specified by extending the multilevel logistic model in Equation (9) to include random slopes and cross-level interactions for the level-1 covariates. By adding assumptions about the distribution of the random slopes (i.e., random effects are normally distributed), the RE approach is less "data-hungry" than FE. However, as shown in Study 1, the RE approach requires that all level-2 confounders be measured.

It should be emphasized that the balancing of covariate distributions within clusters is only needed when the covariate-by-cluster interactions (i.e., all interactions of a covariate and cluster-indicator variables) are present in the data-generating mechanism of both the treatment assignment and the outcome. The reasoning here is that potential confounders of a treatment effect have to be associated with both the treatment and the outcome. Thus, researchers can ignore covariate-by-cluster interactions when modeling the treatment assignment, if the covariate effects are constant in the outcome model.[5] This would also explain why some previous simulation research found that ignoring variation of covariate effects across clusters in the propensity score model did not substantially bias estimates of the treatment effect (e.g., Leite et al., 2015). In the next section, we evaluate propensity score weighting approaches when random slopes are present in the treatment as well as the outcome model.

## Simulation study 3: random slopes in treatment and outcome model

Study 3 has two aims. First, we evaluate the performance of the different weighting methods (IPW, IPW-T, and OW weights) in the more general case of level-1 covariate effects that vary across groups. Importantly, we allow the level-1 covariate effects to vary in the treatment assignment model as well as in the outcome model. As previously pointed out, the random slopes should only have a confounding effect on the estimates of the ATE when they are present in both the treatment and outcome data-generating mechanisms. Second, we test the performance of the two calibration estimators in scenarios with random slopes for covariates. We expect that at least moderate cluster sizes (i.e., $n_j \geq 30$) are needed to provide stable estimates of the ATE when covariate-by-cluster interactions are included.

### Method

We specified the following data-generating equation for treatment assignment:

$$T_{ij}^{*} = \alpha_0 + \alpha_X X_{ij} + \alpha_Z Z_j + \alpha_{XZ} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} + \varepsilon_{ij} \quad (34)$$

where $X$ and $Z$ are two independent, standard normal covariates at level 1 and level 2, respectively. The ICC of $X$ was set .20. The level-2 residuals $u_{0j}$ and $u_{1j}$ are bivariate normally distributed with mean zero, and $\varepsilon_{ij}$ follows a logistic distribution. The residual ICC of the treatment indicator was fixed to .2. The random slopes and intercepts are perfectly correlated. We manipulated the magnitude of the slope variation in the treatment equation by setting $\mathrm{Var}(u_{1j}) = f_{slo}\mathrm{Var}(u_{0j})$, and investigated two conditions: with no slope variation (i.e., $f_{slo} = 0$), and with half of the variation of the random intercept (i.e., $f_{slo} = 0.5$).

[5]However, it is still important that all relevant covariates are measured so that the ignorability assumption is met (see Equation (2)).

The equation for the outcome was a multilevel model with a random slope for the covariate $X$:

$$Y_{ij} = \beta_0 + \tau T_{ij} + \beta_X X_{ij} + \beta_Z Z_j + \beta_{XZ} X_{ij} Z_j + \nu_{0j} + \nu_{1j} X_{ij} + e_{ij} \quad (35)$$

where $\nu_{0j}$, $\nu_{1j}$ and $e_{ij}$ are the normally distributed residuals at level 2 and level 1. The residual ICC of the outcome was fixed to .2. The treatment effect $\tau$ (ATE) was set to .30, and regression intercepts in the treatment and outcome equations were set to zero. Again, we manipulated the magnitude of the slope variation by setting $\text{Var}(\nu_{1j}) = f_{slo}\text{Var}(\nu_{0j})$, and investigated the two conditions $f_{slo} = 0$ and $f_{slo} = 0.5$. The random slopes and the random intercept were assumed to be uncorrelated. We set the effect of the level-1 covariate $X$ to be equal in the treatment and outcome equations ($\alpha_X = \beta_X = .5$), and manipulated the effects of the level-2 covariate and the cross-level interaction between $X$ and $Z$ in two conditions: zero ($\alpha_Z = \beta_Z = 0$, and $\alpha_{XZ} = \beta_{XZ} = 0$), and .5 ($\alpha_Z = \beta_Z = .5$, and $\alpha_{XZ} = \beta_{XZ} = .5$). In the scenarios with random slope variation, this resulted in the following explained variation for the treatment assignment model: $R^2_{L1} = .11$ and $R^2_{L2} = .03$, when the effects of $Z$ and $XZ$ were assumed to be zero, and $R^2_{L1} = .14$, and $R^2_{L2} = .08$, when the effects of $Z$ and $XZ$ were assumed to be .5.[6] We set the number of clusters to $J = 100$ and manipulated the number of units per cluster $n_j = 20$, 30, 50, and 100.

For each of the 2 (no random slope variation vs. random slope variation) × 2 (effect of $Z$ and $XZ$ vs. no effect of $Z$ and $XZ$) × 4 (number of units per cluster) = 16 conditions, 1,000 simulated data sets were generated. For each simulated data set, propensity scores were estimated with four different models. We specified two variants of a logistic regression model. In the fixed-effects clustered (FEC) approach, we included the level-1 covariate $X$, a set of $J$–1 cluster indicators, and $J$–1 interaction terms between $X$ and the $J$–1 cluster indicators. We also studied the FE approach from the previous two simulations, which only included $X$ and the $J$–1 cluster indicators. Besides, we implemented two variants of the random-effects approach. We specified a multilevel model that included both covariates (i.e., $X$ and $Z$) and their cross-level interaction (i.e., $XZ$), but no random slopes for $X$. This multilevel random-intercept model was labeled RE(XZ). The second random-effects model is a multilevel model that included random slopes (REC(XZ)). The propensity scores were then used to compute IPW, IPW-T, and OW weights. Thus, 4 (propensity score models) × 3 (type of weights) = 12 different estimators of the ATE were calculated. Finally, we implemented the calibration estimators in two variants: one version ignored covariate-by-cluster interactions and was identical to the estimators that we used in the previous simulations (CAL1 and CAL2). The second variant included interaction terms between the level-1 covariate $X$ and the $J$–1 cluster indicators in the design matrix (CALC1 and CALC2). In total, 16 estimators of the ATE were compared (the R code for the data-generating and analysis models is provided in Supplements S10 and S11). Again, we computed the relative bias and the relative RMSE to evaluate the quality of the parameter estimates.

## Results

In Table 4, bias and RMSE results for the different weighting estimators when the data were generated without random slopes (upper panel) and with random slopes (lower panel) is shown. In the case of no random slope variation, the results of the previous simulation studies are confirmed: all the FE approaches produced unbiased estimates of the ATE, as do the RE approaches with the more stable IPW-T and OW weights, and the CAL1 and CAL2 procedures. In contrast, the weighting estimators that wrongly assumed cluster-specific effects for the level-1 covariate in the propensity score model (i.e., FEC and REC(XZ)), were substantially biased, mainly when clusters are small. The FEC approach with IPW weights yielded particularly unstable estimates of the ATE with smaller cluster sizes, indicating that the data did not provide enough information to estimate cluster-specific covariate effects. This result suggests that the random-effects approach RE(XZ) is preferable with smaller cluster sizes. However, the difference was less pronounced when the more stable IPW-T and OW weights were used.

When the data were generated with random slopes in the treatment and outcome equations (lower panel in Table 4), all the methods that ignore the varying effect of the level-1 covariate were substantially biased, regardless of the cluster size. The methods that allowed for covariate-by-cluster interactions in the propensity score model needed large cluster sizes to achieve an acceptable performance. This was also true for the two calibration

[6] When random slopes are included in the multilevel model, the explained variation is calculated as follows: $R^2_{L1} = [\alpha^2_X(1 - ICC_X) + \alpha^2_{XZ}(1 - ICC_X) + (1 - ICC_X)Var(u_{1j})]/\text{Var}_{\text{total}}$, and $R^2_{L2} = [\alpha^2_X ICC_X + \alpha^2_Z + \alpha^2_{XZ} ICC_X + ICC_X Var(u_{1j})]/\text{Var}_{\text{total}}$, where $\text{Var}_{\text{total}} = \alpha^2_X + \alpha^2_Z + \alpha^2_{XZ} + \text{Var}(u_{0j}) + \text{Var}(u_{1j}) + \pi^2/3$ (Snijders & Bosker, 2012).

**Table 4.** Simulation study 3: relative bias and relative RMSE for data generated without and with random slopes and cross-level interactions as a function of cluster size.

| | | | Bias | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Weight | $n_j$ | 20 | 30 | 50 | 100 | 20 | 30 | 50 | 100 |
| *Data generated without random slopes* | | | | | | | | | | |
| FEC | IPW | | **64** | **52** | **35** | **21** | .67 | .55 | .38 | .23 |
| FEC | IPW-T | | **32** | **24** | **15** | **9** | .37 | .29 | .19 | .12 |
| FEC | OW | | **25** | **19** | **11** | **7** | .30 | .23 | .15 | .10 |
| | CALC1 | | **24** | **16** | **7** | 4 | .33 | .24 | .16 | .10 |
| | CALC2 | | **24** | **16** | **7** | 3 | .33 | .25 | .16 | .11 |
| REC(XZ) | IPW | | **16** | **14** | **10** | **7** | .27 | .25 | .18 | .14 |
| REC(XZ) | IPW-T | | **6** | 4 | 1 | 1 | .21 | .16 | .12 | .09 |
| REC(XZ) | OW | | −5 | −3 | −3 | −1 | .18 | .14 | .11 | .07 |
| FE | IPW | | 2 | 1 | −1 | 1 | .28 | .25 | .19 | .13 |
| FE | IPW-T | | 0 | 1 | 0 | 0 | .22 | .16 | .12 | .09 |
| FE | OW | | 0 | 1 | −1 | 0 | .17 | .13 | .10 | .07 |
| | CAL1 | | 0 | 1 | −1 | 0 | .20 | .16 | .13 | .09 |
| | CAL2 | | 0 | 1 | −1 | 1 | .19 | .15 | .11 | .08 |
| RE(XZ) | IPW | | **13** | **13** | **9** | **7** | .26 | .23 | .18 | .13 |
| RE(XZ) | IPW-T | | **6** | 5 | 1 | 1 | .21 | .16 | .12 | .09 |
| RE(XZ) | OW | | −5 | −3 | −3 | −1 | .18 | .14 | .11 | .07 |
| *Data generated with random slopes* | | | | | | | | | | |
| FEC | IPW | | **103** | **86** | **67** | **43** | 1.05 | .89 | .70 | .47 |
| FEC | IPW-T | | **40** | **26** | **16** | **8** | .44 | .30 | .21 | .12 |
| FEC | OW | | **30** | **20** | **13** | **7** | .34 | .24 | .17 | .10 |
| | CALC1 | | **54** | **38** | **24** | **12** | .60 | .44 | .30 | .18 |
| | CALC2 | | **39** | **27** | **17** | **8** | .47 | .35 | .25 | .16 |
| REC(XZ) | IPW | | **47** | **43** | **37** | **31** | .59 | .54 | .54 | .38 |
| REC(XZ) | IPW-T | | 4 | 2 | 2 | 1 | .21 | .17 | .13 | .09 |
| REC(XZ) | OW | | **−8** | **−7** | −3 | −1 | .20 | .16 | .12 | .08 |
| FE | IPW | | **80** | **80** | **81** | **80** | .89 | .88 | .88 | .85 |
| FE | IPW-T | | **84** | **84** | **86** | **83** | .89 | .88 | .89 | .86 |
| FE | OW | | **77** | **76** | **78** | **76** | .81 | .80 | .81 | .78 |
| | CAL1 | | **87** | **89** | **92** | **92** | .91 | .92 | .95 | .94 |
| | CAL2 | | **86** | **86** | **88** | **86** | .90 | .90 | .91 | .89 |
| RE(XZ) | IPW | | **96** | **94** | **92** | **86** | 1.02 | 1.00 | .97 | .90 |
| RE(XZ) | IPW-T | | **93** | **91** | **91** | **86** | .98 | .95 | .94 | .88 |
| RE(XZ) | OW | | **76** | **76** | **78** | **76** | .80 | .79 | .81 | .78 |

*Note.* $n_j$ = cluster size; FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates *X* and *Z*; FEC = fixed-effects propensity scores with fixed effects for intercepts and slopes; REC(XZ) = random-effects propensity scores with random slopes and interaction effect; IPW = inverse probability weighting; IPW-T = inverse probability weighting with trimming; OW = overlap weights; CAL1 = calibration weights of Kim et al. (2017); CAL2 = calibration weights of Yang (2018); CALC1 = calibration weights of Kim et al. (2017) with covariate-by-cluster interactions; CALC2 = calibration weights of Yang (2018) with covariate-by-cluster interactions. Relative biases smaller than −5 or larger than 5 are printed in bold.

**Table 5.** Simulation study 3: relative bias and relative RMSE for conditions in which random slopes were only generated in the propensity score model and conditions in which random slopes were only generated in the outcome model.

| | | | Bias | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|
| Model | Weight | $n_j$ | 30 | 50 | 100 | 30 | 50 | 100 |
| *Only random slopes in PS model* | | | | | | | | |
| FEC | IPW | | **66** | **53** | **38** | .69 | .56 | .41 |
| FEC | IPW-T | | **16** | **10** | 5 | .23 | .16 | .10 |
| FEC | OW | | **12** | **8** | 4 | .19 | .13 | .08 |
| | CALC1 | | **22** | **15** | **9** | .30 | .22 | .15 |
| | CALC2 | | **22** | **14** | **8** | .30 | .22 | .16 |
| REC(XZ) | IPW | | **36** | **33** | **26** | .46 | .42 | .43 |
| REC(XZ) | IPW-T | | **6** | 3 | 2 | .17 | .13 | .09 |
| REC(XZ) | OW | | −4 | −3 | −1 | .15 | .11 | .07 |
| FE | IPW | | **−10** | **−6** | −4 | .30 | .22 | .17 |
| FE | IPW-T | | 0 | −1 | 0 | .18 | .13 | .11 |
| FE | OW | | 0 | 0 | 0 | .14 | .10 | .07 |
| | CALC1 | | 0 | −1 | 0 | .16 | .13 | .09 |
| | CALC2 | | 0 | −1 | 0 | .15 | .12 | .08 |
| RE(XZ) | IPW | | 1 | 2 | 1 | .23 | .18 | .15 |
| RE(XZ) | IPW-T | | 4 | 2 | 1 | .18 | .13 | .11 |
| RE(XZ) | OW | | −4 | −2 | −1 | .14 | .10 | .07 |
| *Only random slopes in outcome model* | | | | | | | | |
| FEC | IPW | | **62** | **46** | **27** | .66 | .50 | .31 |
| FEC | IPW-T | | **26** | **16** | **8** | .31 | .21 | .13 |
| FEC | OW | | **20** | **13** | **7** | .24 | .17 | .10 |
| | CALC1 | | **22** | **15** | **6** | .31 | .24 | .14 |
| | CALC2 | | **22** | **14** | 5 | .31 | .23 | .14 |
| REC(XZ) | IPW | | **16** | **16** | **10** | .34 | .27 | .21 |
| REC(XZ) | IPW-T | | 4 | 2 | 1 | .20 | .15 | .10 |
| REC(XZ) | OW | | −4 | −2 | −1 | .16 | .12 | .08 |
| FE | IPW | | 2 | 4 | 0 | .37 | .30 | .24 |
| FE | IPW-T | | 2 | 1 | 0 | .21 | .16 | .11 |
| FE | OW | | 1 | 1 | 0 | .16 | .12 | .09 |
| | CALC1 | | 0 | 1 | 0 | .23 | .19 | .14 |
| | CALC1 | | 1 | 1 | 0 | .20 | .17 | .12 |
| RE(XZ) | IPW | | **14** | **14** | **9** | .35 | .29 | .22 |
| RE(XZ) | IPW-T | | 5 | 3 | 1 | .21 | .16 | .11 |
| RE(XZ) | OW | | −3 | −2 | −1 | .16 | .13 | .09 |

*Note.* $n_j$ = cluster size; FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates *X* and *Z*; FEC = fixed-effects propensity scores with fixed effects for intercepts and slopes; REC(XZ) = random-effects propensity scores with random slopes and interaction effect; IPW = inverse probability weighting; IPW-T = inverse probability weighting with trimming; OW = overlap weights; CAL1 = calibration weights of Kim et al. (2017); CAL2 = calibration weights of Yang (2018); CALC1 = calibration weights of Kim et al. (2017) with covariate-by-cluster interactions; CALC2 = calibration weights of Yang (2018) with covariate-by-cluster interactions. Relative biases smaller than −5 or larger than 5 are printed in bold.

estimators, which were strongly biased unless cluster sizes were large ($n_j > 50$).

A key consideration that has not been sufficiently emphasized in previous research is whether random slopes are present in both the treatment and outcome equations or only in one of these. The performance of the methods that ignore variation in the level-1 covariate slopes only deteriorates if random slopes are present in the population models for the treatment as well as the outcome. This is clearly illustrated in Table 5, which presents results for conditions in which random slope variation was only included in the data-generating model for the treatment (upper panel) or the outcome (lower panel).[7] The methods without covariate-by-cluster interactions in the estimation of the propensity model (i.e., FE and RE(XZ)), and the calibration procedures CAL1 and CAL2 produced approximately unbiased estimates of the ATE.

## Summary and discussion

The main findings of the simulation can be summarized as follows. First, when covariate random slopes

[7]For these simulations, the same data generating parameters were used as in the main study, with the only exception that no slope variation was simulated in either the treatment or the outcome model.

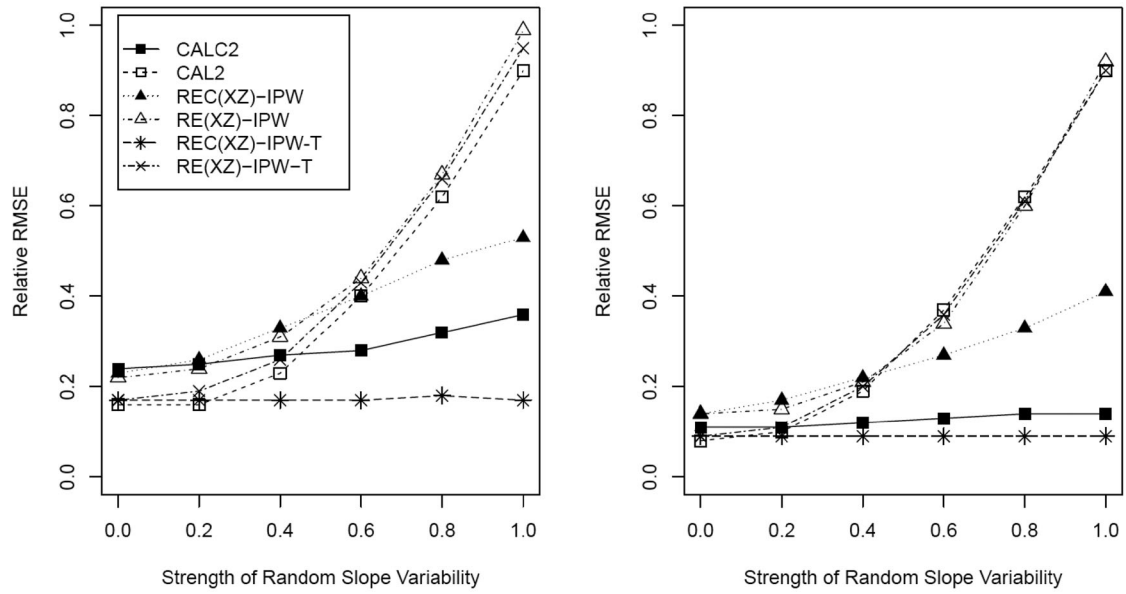


**Figure 4.** Relative RMSE of the different estimators of the treatment effect as a function of the strength of random slope variability for moderate cluster sizes $n_j = 30$ (left panel), and large cluster sizes $n_j = 100$ (right panel). RE(XZ) = random-effects propensity scores with covariates $X$ and $Z$; REC(XZ) = random-effects propensity scores with random slopes and interaction effect; IPW = inverse probability weighting; IPW-T = inverse probability weighting with trimming; CAL2 = calibration weights of Yang (2018); CALC2 = calibration weights of Yang (2018) with covariate-by-cluster interactions.

were not present in the data-generating mechanism, most of the methods that assumed random slopes showed severe bias unless clusters were large. The only exceptions were the weighting estimators computed with RE propensities and stabilized through either IPW-T or OW weights, which were able to recover the treatment effect even in the $n_j = 20$ condition. In this case, the distributional assumption made in the RE model was advantageous, since it shrinks the slopes closer to the truth of no variation. Second, when random slopes are present in both the population treatment and outcome equations, all estimates that do not account for them are severely biased, regardless of cluster size. Third, again under a data-generating model with random slopes, among the estimators that assume varying slopes, only the weighting estimators of the RE approach with IPW-T or OW weights were able to recover the ATE across all cluster size conditions. The calibration estimators and the FE approach with IPW-T and OW weights could only recover the ATE with acceptable bias in conditions with large cluster sizes (i.e., $n_j > 50$). However, in this simulation study we assumed no unobserved level-2 confounders (i.e., the level-2 covariate $Z$ was observed) and homogeneous treatment effects (e.g., no treatment-covariate interactions).The presence of unobserved confounders at level 2 would result in biased estimates of the treatment effect for the RE approaches. In addition, as demonstrated in Simulation Study 2, under heterogeneous treatment effects the IPW-T and OW weights would not recover the ATE as they focus on a different target population (i.e., subpopulation that had nontrivial probabilities for both being among treated and controls).

To further investigate how slope variation affects the performance of the different estimators, we conducted an additional simulation, in which we manipulated the magnitude of the slope variation. More specifically, we varied the strength of the cross-level interaction, i.e., $\alpha_{XZ} = \beta_{XZ} = .5{\cdot}s$, and the slope variation, i.e., $\mathrm{Var}(u_{1j}) = \mathrm{Var}(v_{1j}) = .5{\cdot}s$, in the treatment as well as the outcome model by setting $s = 0$, .2, .4, .6, .8, and 1. Figure 4 shows the performance of a subset of the weighting estimators in terms of RMSE as a function of the simulated slope variation and the cluster size (left panel: $n_j = 30$, right panel: $n_j = 100$). As the figure shows, strong random slope variation ($s > .5$) needs to be present in order to deteriorate the estimates of the calibration estimator CALC2 that takes into account covariate-by-cluster-interactions (see Supplement S12 for detailed results). Since REC(XZ) assumes all relevant level-2 covariates are measured, the CALC2 estimator that only requires that all level-1 confounders be

measured is attractive, particularly in data constellations with only moderate slope variation.[8]

## Inclusion of survey weights

Complex survey data, like the Programme for International Student Assessment (PISA; OECD, 2018), employ a complex, stratified cluster sampling. In these studies, level-1 units and level-2 units are typically accommodated with sampling weights at the respective levels. These sampling weights also have to be included in the estimation of treatment effects (see Leite, 2016). Each cluster $j$ possesses a level-2 sampling weight $w_j$ that reflects the probability that the cluster is sampled in the study. Each student $i$ within a cluster $j$ possesses a level-1 sampling weight $w_{i|j}$. Moreover, for analyses at the total population level, students also receive a total sampling weight $w_{ij}$. Stapleton (2013) provides an accessible review of using different weights in international large-scale assessment studies. In the following, we show how the propensity score weighting estimators have to be modified for accommodating sampling weights (Dong et al., 2020; Ridgeway et al., 2015).

First, sampling weights have to be included in the propensity score model. As the fixed-effects logistic model is a single-level model, total sampling weights $w_{ij}$ have to be used. In the random-effects model, level-2 sampling weights $w_j$ and level-1 sampling weights $w_{i|j}$ must be applied. The predicted probabilities $\hat{\pi}_{ij}$ are then used to calculate weights $\hat{\omega}_{ij}$ like in the case without sampling weights (see Equation (12)). However, sampling weights have to be included in the weighted treatment effect estimate:

$$\hat{\tau} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij}T_{ij}Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij}T_{ij}} - \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij}(1-T_{ij})Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij}(1-T_{ij})} \tag{36}$$

In the computation of calibration weights, the balancing conditions now also include sampling weights. Equations (15) and (17) are modified to

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij,\mathrm{CAL}}T_{ij}\mathbf{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij,\mathrm{CAL}}(1-T_{ij})\mathbf{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\mathbf{X}_{ij} \tag{37}$$

$$\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij,\mathrm{CAL}}T_{ij} = \sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij,\mathrm{CAL}}\left(1-T_{ij}\right) = \sum_{i=1}^{n_j} w_{ij} \quad (j=1,\ldots,J) \tag{38}$$

In addition, sampling weights are also included in the definition of the calibration weights. The calibration weights of Yang (2018) are given as

$$\hat{\omega}_{ij,\mathrm{CAL2}} = \begin{cases} \left(\sum_{h=1}^{n_j} w_{hj}\right)\dfrac{w_{ij}\omega_{ij}^{*}\exp\left\{\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}_1\right\}}{\sum_{h=1}^{n_j} w_{hj}\omega_{hj}^{*}T_{hj}\exp\left\{\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}_1\right\}} & \text{for } T_{ij}=1 \\ \left(\sum_{h=1}^{n_j} w_{hj}\right)\dfrac{w_{ij}\omega_{ij}^{*}\exp\left\{\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}_0\right\}}{\sum_{h=1}^{n_j} w_{hj}\omega_{hj}^{*}(1-T_{hj})\exp\left\{\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}_0\right\}} & \text{for } T_{ij}=0 \end{cases}, \tag{39}$$

The calibration weights of Kim et al. (2017) are similarly modified.

## Example: effect of migration background on reading outcomes

In this section, we apply the various propensity weighting estimators to data from the German sample of the 2015 PISA study. We are interested in the effect of a student's migration background on his or her reading score. Our binary treatment variable (immig) pools together students who are first or second generation immigrants (immig = 1), to compare their reading performance with that of students who did not report having such backgrounds (immig = 0). As immigrant status is not manipulable (Holland, 1986), the main goal was to make a controlled descriptive comparison (see Li et al., 2013) between immigrant and nonimmigrant students' reading scores. To this end, we controlled for a small set of level-1 covariates representing the student's socioeconomic background-home possessions, index of highest parental occupational status, and index of highest parental education in years of schooling-as these are likely to have different distributions among immigrants and natives, and are also strongly associated with educational outcomes (OECD, 2018). We additionally consider the school-level aggregates of these three socioeconomic variables (i.e., cluster means of the level-1 covariates), since the

[8]However, it needs to be pointed out that we assumed the treatment effect is constant (i.e., no treatment effect heterogeneity) in the data-generating model of Simulation Study 3. It can be expected that including heterogeneous treatment effects would even further increase the cluster sizes that are needed to produce stable estimates with the calibration estimators because treatment effect heterogeneity can be expected to introduce further uncertainty in the estimation of the ATE. Thus, it is an important topic for future research to develop more stable versions of the calibration estimators, and investigate their performance under scenarios with heterogeneous treatment effects as well covariate-by-cluster interactions (Kranker et al., 2020; Soriano et al., 2021).

**Table 6.** Point estimates and standard errors for the effect of migration background on reading scores in the German sample of PISA 2015.

| | | M1 | | M2a | | M2b | | M3a | | M3b | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Weight | $\hat{\tau}$ | S.E. | $\hat{\tau}$ | S.E. | $\hat{\tau}$ | S.E. | $\hat{\tau}$ | S.E. | $\hat{\tau}$ | S.E. |
| FE | IPW | −26.7 | 4.5 | −25.4 | 5.0 | −25.6 | 5.0 | −24.8 | 6.1 | −25.2 | 6.1 |
| | IPW-T | −23.7 | 4.1 | −18.5 | 3.8 | −17.8 | 3.5 | −18.2 | 4.0 | −18.1 | 4.6 |
| | OW | −19.5 | 2.9 | −17.8 | 2.8 | −17.8 | 2.7 | −17.4 | 2.8 | −17.6 | 2.7 |
| RE | IPW | −24.8 | 4.1 | −21.5 | 4.3 | −21.3 | 4.3 | −21.4 | 4.7 | −21.0 | 4.6 |
| | IPW-T | −18.1 | 3.8 | −11.5 | 3.9 | −11.5 | 3.9 | −12.5 | 4.0 | −12.7 | 4.1 |
| | OW | −17.8 | 3.5 | −15.4 | 3.6 | −15.5 | 3.6 | −15.0 | 3.7 | −15.1 | 3.6 |
| | CAL1 | −21.4 | 4.2 | −19.8 | 4.6 | −19.5 | 4.6 | −22.2 | 5.2 | −22.1 | 5.0 |
| | CAL2 | −22.6 | 4.2 | −21.1 | 4.4 | −20.7 | 4.3 | −22.7 | 4.8 | −22.4 | 4.6 |

*Note.* FE = fixed-effects propensity scores; RE = random-effects propensity scores; IPW = inverse probability weighting; IPW-T = inverse probability weighting with trimming; OW = overlap weights; CAL1 = calibration weights of Kim et al. (2017); CAL2 = calibration weights of Yang (2018). M1 = main effects only; M2a = main effects, all two-way interactions of level-1 variables, all squares of level-1 variables; M3a = M2a plus all cross-level interactions; M2b and M3b include main effects and only the higher-order terms deemed important by a likelihood-ratio criterion (Imbens & Rubin, 2015).

**Table 7.** Differences of the estimators under M3b.

| Method | Weight | Naive | FE | | | RE | | | CAL1 |
|---|---|---|---|---|---|---|---|---|---|
| | | | IPW | IPW-T | OW | IPW | IPW-T | OW | |
| FE | IPW | −28.7*** | | | | | | | |
| | IPW-T | −35.9*** | −7.2 | | | | | | |
| | OW | −36.4*** | −7.7 | −0.5 | | | | | |
| RE | IPW | −33.0*** | −4.2 | 2.9 | 3.4 | | | | |
| | IPW-T | −41.3*** | −12.5* | −5.4 | −4.9* | −8.3** | | | |
| | OW | −38.9*** | −10.2* | −3.0 | −2.5 | −5.9* | 2.4 | | |
| | CAL1 | −31.5*** | −2.8 | 4.4 | 4.9 | 1.4 | 9.7** | 7.4* | |
| | CAL2 | −31.9*** | −3.2 | 4.0 | 4.5 | 1.0 | 9.4** | 7.0* | −0.4 |

*Note.* Differences are estimator of the column minus estimator of the row. Naive = unadjusted mean difference; FE = fixed-effects propensity scores; RE = random-effects propensity scores; IPW = inverse probability weighting; IPW-T = inverse probability weighting with trimming; OW = overlap weights; CAL1 = calibration weights of Kim et al. (2017); CAL2 = calibration weights of Yang (2018). $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

composition of schools has also been shown to have an effect on academic performance (OECD, 2018).

The German sample of PISA 2015 consists of 6,504 students from 256 schools, where an average of 25.4 students per school was tested (standard deviation of 6.0), including an average of 3.8 immigrant students per school (standard deviation of 3.9). After listwise deletion of cases with missing data on at least one covariate and the removal of 56 schools in which no students with a migration background were present, the sample reduced to 4,188 students nested in 199 schools. Multiple imputation could have been used to deal with incomplete covariate data (Leyrat et al., 2019; see also Cham & West, 2016). On average, schools in this reduced sample have 3.7 immigrant students (range = 1 to 15) and 17.4 native students (range = 1 to 28).

We applied the propensity score weighting approaches with different covariate specifications. In the first model M1, we only controlled for the main effects of the covariates, that is: the model for the FE propensities (see Equation (8)) only included the main effects of the level-1 covariates; the model for the RE propensities (see Equation (9)) only included the main effects of the level-1 and level-2 covariates; and in the procedure for computing the calibration weights, only the main effects of the level-1 covariates were included. In model M2a, we additionally considered all squares and two-way interactions of the level-1 covariates. In model M3a, we also included all cross-level interactions, that is, all interactions of the three level-1 covariates with the three level-2 aggregates (i.e., school means). Finally, model M2b and M3b correspond to specifications where higher-order terms (i.e., squares and interactions of level-1 covariates, and cross-level interactions) were only included if they were deemed significant by a likelihood-ratio procedure recommended in Imbens and Rubin (2015). For the German PISA sample, this procedure determined that the squares of all level-1 covariates should be included, as well as the interactions of home possessions and parental education with their respective level-2 aggregates, but none of the other cross-level interactions and none of the level-1 two-way interactions. The analysis used the ten plausible values for the reading score, as well as the sampling weights (i.e., school weights and cluster-normalized student weights for the RE propensity score model, and total student weights for the FE propensity score model and the ATE weighting estimator) of the PISA dataset, as outlined in the previous section on survey weights. Standard errors were calculated using the balanced repeated replication (BRR) weights (see OECD, 2009).

The main results are as follows (see Table 6). First, across all covariate specifications, the estimates obtained by weighting with FE propensities see migrants at a larger disadvantage than do the estimates from RE propensity scores, though not all the differences are statistically significant (see Table 7 for statistical significance results on the differences of M3b). This pattern suggests that the estimates based on the FE propensity scores are controlling for unobserved level-2 confounders that the RE propensity scores overlook. Second, although the differences are not all statistically significant, estimates based on IPW are larger in absolute value than estimates that use the IPW-T and OW weights. This could indicate that the effect of migration background is heterogeneous, that is, that the reading achievement gap between migrants and natives is different when comparing students at the low, mid, or high socioeconomic ranges. Lastly,

the estimates produced by the calibration weights remain relatively stable across covariate specifications and lie, for the most part, somewhere between IPW and IPW-T with the FE approach. However, most of the differences were not statistically significant.

## Concluding remarks

This paper examined several propensity score weighting approaches and their ability to estimate the effect of a binary level-1 treatment with multilevel, observational data. We confirmed previous findings from the literature that propensity score weighting based on a propensity score model with fixed effects outperforms a model with random effects (Arpino & Mealli, 2011; Li et al., 2013). Furthermore, in contrast to the random-effects model, the fixed-effects model automatically controls for the effects of unmeasured level-2 confounders. We also found that the IPW estimator with a correctly specified propensity score model provided unbiased but highly variable estimates, particularly in the case of small clusters or strong confounding. We confirmed that trimming IPW weights has the potential to reduce the variance of the estimates, though bias can be introduced in the case of treatment effect heterogeneity (Crump et al., 2009). Overlap weights produced the estimates with the smallest variance. However, these estimates can be severely biased as estimates for the ATE because they upweigh observations in the center of the area of overlap. Alternatively, one could argue that the overlap estimator is targeting an estimand that is different from the ATE and that focuses on a population for whom there is equipoise (Zhou et al., 2020). Thus, the application of overlap weights may be particularly attractive when the assessment of treatment effects is most relevant for observations in the area of overlap.

We showed analytically and in the simulation studies that calibration weights produce unbiased estimates of the treatment effect when all relevant level-1 covariates are taken into account. Similar to the weighting estimators based on fixed-effects propensity scores, calibration estimators controlled for unmeasured level-2 confounders and provided estimates with smaller variance than IPW and its trimmed version. However, in the case of random slopes in the propensity score and outcome models, covariate-by-cluster interactions have to be included in the calculation of the calibration weights. Thus, sufficiently large clusters are needed to obtain accurate estimates of the treatment effect. In constellations with small to moderate cluster sizes ($n_j < 30$), the weighting methods based on random-effects propensity scores may provide more accurate estimates, but require that researchers are certain that all level-2 confounders are accounted for. Improving the performance of calibration weights in the presence of random slopes and small cluster sizes is an important topic for future research. Multilevel latent class logit models (Kim et al., 2016) and cluster analysis (Lee et al., 2019) have been proposed to deal with small cluster size issues when estimating treatment effects in scenarios with covariate-by-cluster interactions (see also Rickles & Seltzer, 2014, for a propensity score matching strategy).

In practical applications of propensity score weighting, the selection of covariates and the correct specification of their effects (e.g., interactions and quadratic effects) can be challenging. In the present article, we assumed that all relevant level-1 covariates were observed (i.e., no unmeasured confounders at level 1). Without specific knowledge about the assignment process, this assumption is often hard to justify (Imbens & Rubin, 2015), and it has been argued that in real applications, the estimation of treatment effects should be accompanied by a sensitivity analysis that tests how sensitive the conclusions are to unmeasured confounding (e.g., VanderWeele, 2019). Furthermore, our simulations were limited to only one level-1 and one level-2 covariate with only linear effects. Efficient algorithms would be needed to select relevant covariate effects in the propensity model (McCaffrey et al., 2004; Suk et al., 2019) and to compute calibration weights (Ning et al., 2020) when the set of covariates is large. It would also be interesting to study propensity score weighting approaches for multivalued treatments (e.g., Leite et al., 2019), and continuous treatments (Imai & van Dyk, 2004; Schuler et al., 2016), as well as more complex multilevel structures (e.g., three-level or cross-classified data; Suk et al., 2019).

## Article information

**Conflict of interest disclosures:** Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

**Ethical principles:** The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

**Funding:** This study was not supported.

**Role of the funders/sponsors:** None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

## References


Allison, P. D. (2009). *Quantitative applications in the social sciences: Fixed effects regression models*. SAGE Publications. https://doi.org/10.4135/9781412993869

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780. https://doi.org/10.1016/j.csda.2010.11.008

Aronow, P. M., & Miller, B. T. (2019). *Foundation of agnostic statistics*. Cambridge University Press.

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

Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods*, *21*, 427–445. https://doi.org/10.1037/met0000076

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*, 656–664. https://doi.org/10.1093/aje/kwn164

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199. https://doi.org/10.1093/biomet/asn055

Dong, N., Stuart, E. A., Lenis, D., & Nguyen, T. Q. (2020). Using propensity score analysis of survey data to estimate population average treatment effects: A case study comparing different methods. *Evaluation Review*, *44*, 84–108. https://doi.org/10.1177/0193841X20938497

Ebbes, P., Böckenholt, U., & Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, *58*(2), 161–178. https://doi.org/10.1046/j.0039-0402.2003.00254.x

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46. https://doi.org/10.1093/pan/mpr025

Hansen, B. B., Rosenbaum, P. R., & Small, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association*, *109*(505), 133–144. https://doi.org/10.1080/01621459.2013.863157

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*, 234–249. https://doi.org/10.1037/a0019623

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.2307/2289064

Hong, G. (2015). *Causality in a social world: Moderation, mediation, and spill-over*. Wiley-Blackwell.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association*, *101*(475), 901–910. https://doi.org/10.1198/016214506000000447

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 243–263. https://doi.org/10.1111/rssb.12027

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*(467), 854–866. https://doi.org/10.1198/016214504000001187

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4–29. https://doi.org/10.1162/003465304323023651

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Keele, L., Lenard, M. A., & Page, L. C. (2020). Matching methods for clustered observational studies in education (EdWorkingPaper: 20–235). Retrieved from Annenberg Institute at Brown University, https://doi.org/10.26300/r5hw-g721

Kim, G., Paik, M. C., & Kim, H. (2017). Causal inference with observational data under cluster-specific non-ignorable assignment mechanism. *Computational Statistics & Data Analysis*, *113*, 88–99. https://doi.org/10.1016/j.csda.2016.10.002

Kim, J., & Seltzer, M. (2007). *Causal inference in multilevel settings in which selection process vary across schools*. Working Paper 708, Centre for the Study of Evaluation (CSE). UCLA. Retrieved from http://cresst.org/publications/cresst-publication-3079/

Kim, J.-S., Steiner, P. M., & Lim, W. C. (2016). Mixture modeling strategies for causal inference with multilevel data. In J. R. Harring, L. M. Stapleton, & S. Natasha Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*. (pp. 335–359). IAP - Information Age Publishing, Inc.

Kranker, K., Blue, L., & Forrow, L. V. (2020). Improving effect estimates by limiting the variability in inverse propensity score weights. *The American Statistician*. https://doi.org/10.1080/00031305.2020.1737229

Kreft, I. G., & De Leeuw, J. (1998). *Introducing Statistical Methods: Introducing multilevel modeling*. SAGE Publications. https://doi.org/10.4135/9781849209366

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS One.*, *6*, e18174. https://doi.org/10.1371/journal.pone.0018174

Lee, Y., Nguyen, T. Q., & Stuart, E. A. (2019). Partially pooled propensity score models for average treatment effect estimation with multilevel data. *arXiv preprint arXiv: 1910.05600v1*.

Leite, W. L. (2016). *Practical propensity score methods using R*. Sage Publishing.

Leite, W. L., Aydin, B., & Gurel, S. (2019). A comparison of propensity score weighting methods for evaluating the effects of programs with multiple versions. *The Journal of Experimental Education*, *87*(1), 75–88. https://doi.org/10.1080/00220973.2017.1409179

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & 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. https://doi.org/10.1080/00273171.2014.991018

Leite, W. L., Stapleton, L. M., & Bettini, E. F. (2019). Propensity score analysis of complex survey data with structural equation modeling: A tutorial with Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(3), 448–469. https://doi.org/10.1080/10705511.2018.1522591

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., Resche-Rigon, M., Carpenter, J. R., & Williamson, E. J. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research*, *28*, 3–19. https://doi.org/10.1177/0962280217713032

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

Li, F., Thomas, L. E., & Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, *188*, 250–257. https://doi.org/10.1093/aje/kwy201

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, *32*, 3373–3387. https://doi.org/10.1002/sim.5786

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937–2960. https://doi.org/10.1002/sim.1903

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Mao, H., Li, L., & Greene, T. (2019). Propensity score weighting analysis and treatment effect discovery. *Statistical Methods in Medical Research*, *28*, 2439–2454. https://doi.org/10.1177/0962280218781171

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, *97*, 951–966. https://doi.org/10.1037/a0028380

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluation causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Morgan, S., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107587991.

Ning, Y., Peng, S., & Imai, K. (2020). Robust estimation of causal effects via high-dimensional balancing propensity score. *Biometrika*, *107*(3), 533–554. https://doi.org/10.1093/biomet/asaa020

OECD. (2009). PISA data analysis manual: SPSS (2nd ed.). OECD. https://doi.org/10.1787/9789264056275-en

OECD. (2018). *PISA 2015 results in focus*. OECD. https://doi.org/10.1787/22260919

Page, L., Lenard, M. A., & Keele, L. (2020). The design of clustered observational studies. *AERA Open*, *6*(3), 1–14. https://doi.org/10.1177/2332858420954401

Rickles, J. H., & Seltzer, M. (2014). A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics*, *39*(6), 612–636. https://doi.org/10.3102/1076998614559748

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, *3*(2), 237–249. https://doi.org/10.1515/jci-2014-0039

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*, 309–338. https://doi.org/10.1037/met0000184

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313. https://doi.org/10.1037/a0014268

Schuler, M. S., Chu, W., & Coffman, D. (2016). Propensity score weighting for continuous exposure with multilevel data. *Health Services and Outcomes Research Methodology*, *16*, 271–292. https://doi.org/10.1007/s10742-016-0157-5

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage Publishers.

Soriano, D., Ben-Michael, E., Bickel, P., Feller, A., & Pimentel, S. (2021). Sensitivity analysis for balancing weights. arxiv.org/abs/2102.09052

Stapleton, L. M. (2013). Incorporating sampling weights into single- and multi-level models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 353–388). Chapman Hall/CRC Press.

Steiner, P. M., Kim, J.-S., & Thoemmes, F. J. (2013). Matching strategies for observational multilevel data. In *JSM proceedings* (pp. 5020–5032). American Statistical Association.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21. https://doi.org/10.1214/09-STS313

Suk, Y., Kang, H., & Kim, J. (2019). Random forests approach for causal inference with clustered observational data. PsyArXiv. 16 Sept. https://doi.org/10.31234/osf.io/xgq2k

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*, 514–543. https://doi.org/10.1080/00273171.2011.569395

VanderWeele, T. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*, 211–219. https://doi.org/10.1007/s10654-019-00494-6

Yang, S. (2018). Propensity score weighting for causal inference with clustered data. *Journal of Causal Inference*, *6*(2). https://doi.org/10.1515/jci-2017-0027

Zhou, Y., Matsouaka, R. A., & Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, *29*(12), 3721–3756. https://doi.org/10.1177/0962280220940334

## Appendix A: Estimation equations for the weights of calibration estimator of Kim et al. (2017)

In Appendix A, we further explain the equations for estimating the weights $\hat{\omega}_{ij,\mathrm{CAL1}}$ (see Equation (23)) in the calibration estimator of Kim et al. (2017). Kim et al. (2017) derived the estimating equations for the calibration weights $\hat{\omega}_{ij,\mathrm{CAL1}} = \omega_{ij,\mathrm{CAL1}}(\hat{\boldsymbol{\lambda}})$ that depend on a parameter vector $\hat{\boldsymbol{\lambda}}$. Let $n_{0j} = \sum_{i=1}^{n_j}(1 - T_{ij})$ and $n_{1j} = \sum_{i=1}^{n_j} T_{ij}$ denote the number of control and treated units in cluster $j$, respectively. The estimating equation (see Equation (15) in Kim et al., 2017) for $\hat{\boldsymbol{\lambda}}$ can be rewritten as (set $\boldsymbol{\phi}_1 = -\hat{\boldsymbol{\lambda}}$ in the notation of Kim et al. 2017, and also use Equations (13) and (14) in Kim et al., 2017)

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} T_{ij}\left\{1 + n_{0j}\frac{\exp\left(\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}\right)}{\sum_{h=1}^{n_j} T_{hj}\exp\left(\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}\right)}\right\}$$

$$= \sum_{j=1}^{J}\sum_{i=1}^{n_j}(1 - T_{ij})\left\{1 + n_{1j}\frac{\exp\left(-\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}\right)}{\sum_{h=1}^{n_j}(1 - T_{hj})\exp\left(-\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}\right)}\right\} \tag{A1}$$

Calibration weights are then computed as

$$\hat{\omega}_{ij,\mathrm{CAL1}} = \begin{cases} 1 + n_{0j}\dfrac{\exp\left\{\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}\right\}}{\sum_{h=1}^{n_j} T_{hj}\exp\left\{\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}\right\}} & \text{for } T_{ij} = 1 \\ 1 + n_{1j}\dfrac{\exp\left\{-\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}\right\}}{\sum_{h=1}^{n_j}(1 - T_{hj})\exp\left\{-\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}\right\}} & \text{for } T_{ij} = 0 \end{cases}, \tag{A2}$$

## Appendix B: Estimation equations for the weights of calibration estimator of Yang (2018)

In Appendix B, we show how the estimation equation for the weights $\hat{\omega}_{ij,\mathrm{CAL2}}$ (see Equation (24)) is obtained. Yang (2018) starts from an initial vector of weights $\omega^*_{ij}$. Calibration weights $\omega_{ij,\mathrm{CAL2}}$ are constructed by minimizing the Kullback-Leibler information

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\omega_{ij,\mathrm{CAL2}}\log\frac{\omega_{ij,\mathrm{CAL2}}}{\omega^*_{ij}}, \tag{B1}$$

subject to calibration conditions defined by Equations (15) and (17) as side conditions. Using the Lagrange multipliers technique (Yang, 2018), the calibration weights $\hat{\omega}_{ij,\mathrm{CAL2}} = \omega_{ij,\mathrm{CAL2}}(\hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1)$ are given as (see Equation (10) in Yang, 2018)

$$\hat{\omega}_{ij,\mathrm{CAL2}} = \begin{cases} \dfrac{\omega^*_{ij}\exp\left\{\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}_1\right\}}{\sum_{h=1}^{n_j} T_{hj}\omega^*_{hj,0}\exp\left\{\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}_1\right\}} n_j & \text{for } T_{ij} = 1 \\ \dfrac{\omega^*_{ij}\exp\left\{\mathbf{X}_{ij}\hat{\boldsymbol{\lambda}}_0\right\}}{\sum_{h=1}^{n_j}(1 - T_{hj})\omega^*_{hj,0}\exp\left\{\mathbf{X}_{hj}\hat{\boldsymbol{\lambda}}_0\right\}} n_j & \text{for } T_{ij} = 0 \end{cases}, \tag{B2}$$

where $\hat{\boldsymbol{\lambda}}_0$ and $\hat{\boldsymbol{\lambda}}_1$ are vectors of coefficients of level-1 covariates that fulfill the estimating equations (applying simple algebra to Equation (11) in Yang, 2018)

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,\mathrm{CAL2}}(\hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1)T_{ij}\mathbf{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{X}_{ij} \tag{B3}$$

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,\mathrm{CAL2}}(\hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1)(1 - T_{ij})\mathbf{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{X}_{ij} \tag{B4}$$

For $p$ covariates $\mathbf{X}_{ij}$, the vectors $\hat{\boldsymbol{\lambda}}_0$ and $\hat{\boldsymbol{\lambda}}_1$ are both of length $p$, and there are $2p$ nonlinear equations in Equations (B3) and (B4). Note that Equations (B3) and (B4) can be independently solved for $\hat{\boldsymbol{\lambda}}_0$ and $\hat{\boldsymbol{\lambda}}_1$ because $\omega_{ij,\mathrm{CAL2}}(\hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1)$ is only a function of $\hat{\boldsymbol{\lambda}}_1$ in Equation (B3) and $\omega_{ij,\mathrm{CAL2}}(\hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1)$ is only a function of $\hat{\boldsymbol{\lambda}}_0$ in Equation (B4).

## Appendix C: Unbiasedness of calibration estimators with covariate-by-cluster interactions

We show that unbiased estimates can be obtained for the calibration estimators in condition with random slopes. By using the data generating model defined in Equation (31), the population ATE $\tau = \mathrm{E}\left(\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{Y_{ij}(1) - Y_{ij}(0)\}\right)$ is given as

$$\begin{aligned}\tau = {} & \beta_{0,1} - \beta_{0,0} + \mathrm{E}\left(\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{X}_{ij}\right)\left(\boldsymbol{\beta}_{\mathbf{X},1} - \boldsymbol{\beta}_{\mathbf{X},0}\right) \\ & + \mathrm{E}\left(\frac{1}{N}\sum_{j=1}^{J} n_j\mathbf{V}_j\right)\left(\boldsymbol{\beta}_{\mathbf{V},1} - \boldsymbol{\beta}_{\mathbf{V},0}\right) \\ & + \mathrm{E}\left(\frac{1}{N}\sum_{j=1}^{J}\mathbf{V}_j\sum_{i=1}^{n_j}\mathbf{X}_{ij}\right)\left(\boldsymbol{\beta}_{\mathbf{XV},1} - \boldsymbol{\beta}_{\mathbf{XV},0}\right) \\ & + \mathrm{E}\left(\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{X}_{ij}(\mathbf{U}_{1j,1} - \mathbf{U}_{1j,0})\right) \qquad \text{(C1)}\end{aligned}$$

We now consider the first term in $\hat{\tau}_{\mathrm{CAL}}$ and obtain by using balancing conditions (32) and (33):

$$\begin{aligned}& \mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,\mathrm{CAL}}T_{ij}Y_{ij}\right] \\ & = \mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,\mathrm{CAL}}T_{ij}Y_{ij}(1)\right] \\ & = E\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,\mathrm{CAL}}T_{ij}\left(\beta_{0,1} + \mathbf{X}_{ij}\boldsymbol{\beta}_{\mathbf{X},1}\right.\right. \\ & \qquad \left.\left. + \mathbf{V}_j\boldsymbol{\beta}_{\mathbf{V},1} + \mathbf{X}_{ij}\mathbf{V}_j\boldsymbol{\beta}_{\mathbf{XV},1} + U_{0j,1} + \mathbf{X}_{ij}\mathbf{U}_{1j,1} + e_{ij,1}\right)\right] \\ & = E\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,\mathrm{CAL}}T_{ij}\left(\beta_{0,1} + \mathbf{X}_{ij}\boldsymbol{\beta}_{\mathbf{X},1} + \mathbf{V}_j\boldsymbol{\beta}_{\mathbf{V},1} + \mathbf{X}_{ij}\mathbf{V}_j\boldsymbol{\beta}_{\mathbf{XV},1} + \mathbf{X}_{ij}\mathbf{U}_{1j,1}\right)\right] \\ & = N\beta_{0,1} + E\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\mathbf{X}_{ij}\boldsymbol{\beta}_{\mathbf{X},1} + \mathbf{V}_j\boldsymbol{\beta}_{\mathbf{V},1} + \mathbf{X}_{ij}\mathbf{V}_j\boldsymbol{\beta}_{\mathbf{XV},1} + \mathbf{X}_{ij}\mathbf{U}_{1j,1}\right)\right] \qquad \text{(C2)}\end{aligned}$$

Similarly, we get for the second term in $\hat{\tau}_{\mathrm{CAL}}$ :

$$\begin{aligned}& \mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,\mathrm{CAL}}(1 - T_{ij})Y_{ij}\right] \\ & = \mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,\mathrm{CAL}}(1 - T_{ij})Y_{ij}(0)\right] \\ & = N\beta_{0,0} + \mathrm{E}\left[\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\mathbf{X}_{ij}\boldsymbol{\beta}_{\mathbf{X},0} + \mathbf{V}_j\boldsymbol{\beta}_{\mathbf{V},0} + \mathbf{X}_{ij}\mathbf{V}_j\boldsymbol{\beta}_{\mathbf{XV},0}\right.\right. \\ & \qquad \left.\left. + \mathbf{X}_{ij}\mathbf{U}_{1j,0}\right)\right] \qquad \text{(C3)}\end{aligned}$$

Hence, by using Equations (C2) and (C3), the expected value of $\hat{\tau}_{\mathrm{CAL}}$ equals the population quantity $\tau$ that is given by Equation (C1).