Propensity Score Weighting Procedures for Causal Inference with Clustered Data

Inaugural-Dissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschafts- und Sozialwissenschaften der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Master of Science

Alvaro Fuentes Higuera
aus Mexiko Stadt

Kiel, 2022

Gedruckt mit Genehmigung der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel

Dekan:
Prof. Dr. Christian Martin

Erstberichterstattender:
Prof. Dr. Kai Carstensen

Zweitberichterstattender:
Prof. Dr. Uwe Jensen

Drittberichterstattender:
Prof. Dr. Oliver Lüdtke

Tag der Abgabe der Arbeit:
5. Dezember 2022

Tag der mündlichen Prüfung:
15. Mai 2023

Contents

List of Abbreviations

ATE Average treatment effect
BRR Balanced repeated replication
CAL Calibration estimator
CATE Conditional average treatment effect
CL Clustered estimator
FE Fixed effects
ICC Intraclass correlation
IPW Inverse probability weighting
IPW-T Trimmed inverse probability weighting
L1 Level one
L2 Level two
OW Overlap weights
PAM Partitioning around medoids
PISA Programme for International Student Assessment
PS Propensity score
RE Random effects
RMSE Root mean square error

List of Figures

List of Tables

Introduction

This dissertation was written during my tenure as a research associate at the Leibniz Institute for Science and Mathematics Education. A typical, cross-sectional database generated by educational research projects —from local district studies to international large-scale assessments— has students as observations, and contains information on the students' academic performance, socioeconomic background, and noncognitive traits. Importantly, sampled students that attend the same school tend to be more alike in many of these characteristics than sampled students attending different schools. In statistical terms we say that the students in such databases are "clustered" in schools and that these databases have a multilevel structure, where students represent the unit level ("level one") and schools the cluster level ("level two"). Databases may of course be structured with more than two levels, with, for instance, students clustered into classrooms, classrooms into schools, schools into districts, districts into states, and states into countries. Regardless of the number of levels, a multilevel structure signals possible dependency among units, that is, a violation of the independent sampling many statistical algorithms assume.

In the three articles that make up this dissertation, we studied the behavior of treatment effect estimators that take into account the multilevel structure of the data. The estimators we studied accomplish this by modeling the relationship between treatment assignment and confounders at both the unit and cluster levels. More specifically, they estimate a function of the sample called the propensity score, which is a summary of the confounding information. The literature on propensity score procedures offers a variety of ways to use the propensity score to control for confounding, but here we focus on propensity score weighting estimators, which weight the sample using weights constructed from the propensity score prior to computing a treatment effect.

The first article, presented in Chapter 1, compares the performance of several variants of the traditional fixed-effects and random-effects estimators from the multilevel literature, and of two more recent estimators that are based on calibration procedures. Through Monte Carlo simulation, we tested the ability of these estimators to determine the effect of a binary treatment under many conditions, including conditions where there were few or many clusters, where clusters were small or large, where the treatment prevalence was balanced or imbalanced, and where the effect of confounders varied from cluster to cluster (i.e., random slopes). I refer to this article as Fuentes et al. (2021) in other sections of the text.

The article of Chapter 2 describes the application of these estimation methods to the case of a multicategorical treatment. Although the procedures themselves are very similar, multilevel multicategorical treatment effect estimation does present some unique choices (e.g., whether to apply multicategorical estimation at all, rather than binary) and challenges (e.g., extreme prevalences due to many categories coexisting in small clusters; high-leverage and/or outlying units at level-two). Again through Monte Carlo simulation, we study the performance of the estimators in this setting. I refer to this article as Fuentes & Lüdtke (2022) in other sections of the text.

Finally, the article of Chapter 3 asks whether the performance of the fixed-effects estimator and of one of the calibration estimators can be improved through a procedure that has been shown to improve the performance of the random-effects estimator in the literature. The procedure, called partial pooling, groups together similar clusters prior to estimation. In Monte Carlo simulations, we check whether partial pooling can help these estimators deal with conditions that proved difficult in the previous two studies; specifically, few and small clusters, and the presence of random slopes. The article explores this possibility for a multicategorical treatment effect, which again offered interesting challenges for implementation, but the findings are directly applicable to binary treatments. This article is Fuentes (2022) in the reference list.

Chapter 1

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

Co-authored with Oliver Lüdtke and Alexander Robitzsch

Motivation

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.

1.1 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 $Y_{ij} = Y_{ij}(1)T_{ij} + Y_{ij}(0)(1 - T_{ij})$.

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

$$\tau = E\left[Y_{ij}(1) - Y_{ij}(0)\right] = \mu_1 - \mu_0 \tag{1.1}$$

where μ_1 and μ_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 $\boldsymbol{X}_{ij}$ and $\boldsymbol{V}_j$ 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 $\boldsymbol{V}_j$ into an observed part $\boldsymbol{Z}_j$ and an unobserved part $\boldsymbol{W}_j$, i.e., $\boldsymbol{V}_j = (\boldsymbol{Z}_j, \boldsymbol{W}_j)$. If contextual effects of level-1 covariates are present, the observed part $\boldsymbol{Z}_j$ 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):

$$Y_{ij}(1), Y_{ij}(0) \perp T_{ij} | \boldsymbol{X}_{ij}, \boldsymbol{V}_j \tag{1.2}$$

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).[a] Because the ignorability assumption in Equation (1.2) also involves the unobserved cluster-level variables $\boldsymbol{W}_j$, Yang (2018) used the term latent ignorability. Note that the vector of observed cluster-level variables $\boldsymbol{Z}_j$ 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 $(Y_{ij}, T_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{V}_j)$. If the ignorability assumption in Equation (1.2) holds, the ATE can be identified as follows:

$$\tau = E\left[E(Y_{ij} | \boldsymbol{X}_{ij}, \boldsymbol{V}_j, T_{ij} = 1) - E(Y_{ij} | \boldsymbol{X}_{ij}, \boldsymbol{V}_j, T_{ij} = 0)\right] \tag{1.3}$$

[a] 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., $0 < P(T_{ij} = 1 | \boldsymbol{X}_{ij} \boldsymbol{V}_j) < 1$. This assumption implies that there exists sufficient overlap in the covariate distributions between the treatment and control groups.

The expected values of the potential outcomes in the treatment and control conditions (i.e., μ_1 and μ_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.

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., $\boldsymbol{X}_{ij} = \boldsymbol{x}$ and $\boldsymbol{V}_j = \boldsymbol{v}$):

$$\begin{aligned}\tau_{CATE}(\boldsymbol{x}, \boldsymbol{v}) = E(Y_{ij}|\boldsymbol{X}_{ij} = \boldsymbol{x}, \boldsymbol{V}_j = \boldsymbol{v}, T_{ij} = 1) \\ - E(Y_{ij}|\boldsymbol{X}_{ij} = \boldsymbol{x}, \boldsymbol{V}_j = \boldsymbol{v}, T_{ij} = 0)\end{aligned} \tag{1.4}$$

If $\tau_{CATE}(\boldsymbol{x}, \boldsymbol{v})$ 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 (1.3) is obtained by averaging the CATE across the covariate distribution, i.e., $E\left[\tau_{CATE}(\boldsymbol{x}, \boldsymbol{v})\right] = \tau$.

1.2 Propensity scores

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

$$\pi_{ij} = \pi_{ij}(\boldsymbol{X}_{ij}, \boldsymbol{V}_j) = P(T_{ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{V}_j) \tag{1.5}$$

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:

$$Y_{ij}(1), Y_{ij}(0) \perp T_{ij}|\pi_{ij}(\boldsymbol{X}_{ij}, \boldsymbol{V}_j) \tag{1.6}$$

In practice, the propensity scores π_{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):

$$g(\pi_{ij}) = g(P(T_{ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{Z}_j, \boldsymbol{W}_j)) = \gamma_0 + \boldsymbol{X}_{ij}\boldsymbol{\gamma_X} + \boldsymbol{Z}_j\boldsymbol{\gamma_Z} + \boldsymbol{W}_j\boldsymbol{\gamma_W} + U_{0j} \quad (1.7)$$

where γ_0 is the intercept, $\boldsymbol{\gamma_X}$ are the effects of the covariates at the individual level, $\boldsymbol{\gamma_Z}$ and $\boldsymbol{\gamma_W}$ 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: $Cov(\boldsymbol{W}_j, U_{0j}) = 0$, $Cov(\boldsymbol{X}_{ij}, U_{0j}) = 0$, and $Cov(\boldsymbol{Z}_j, U_{0j}) = 0$. Note that $\boldsymbol{W}_j$ 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:

$$g(\pi_{ij}) = \gamma_{0,FE} + \boldsymbol{X}_{ij}\boldsymbol{\gamma}_{\boldsymbol{X},FE} + U_{0j,FE} \quad (1.8)$$

Here, $U_{0j,FE}$ 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 $\hat{\boldsymbol{\gamma}}_{\boldsymbol{X},FE}$ and $\hat{U}_{0j,FE}$ are then used to compute predicted probabilities of treatment $\hat{\pi}_{ij,FE} = g^{-1}(\hat{\gamma}_{0,FE} + \boldsymbol{X}_{ij}\hat{\boldsymbol{\gamma}}_{\boldsymbol{X},FE} + \hat{U}_{0j,FE})$. 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:

$$g(\pi_{ij}) = \gamma_{0,RE} + \boldsymbol{X}_{ij}\boldsymbol{\gamma}_{\boldsymbol{X},RE} + \boldsymbol{Z}_j\boldsymbol{\gamma}_{\boldsymbol{Z},RE} + U_{0j,RE} \tag{1.9}$$

The random intercepts $U_{0j,RE}$ are assumed to be normally distributed. Because the unobserved cluster-level variables $\boldsymbol{W}_j$ are not included in the model, the random intercepts $U_{0j,RE}$ will, in general, be correlated with the level-1 and level-2 variables, i.e., $Cov(\boldsymbol{X}_{ij}, U_{0j,RE}) \neq 0$, and $Cov(\boldsymbol{Z}_j, U_{0j,RE}) \neq 0$. 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 $\hat{\pi}_{ij,RE}$ that converge against $\hat{\pi}_{ij,FE}$. We now discuss how the estimated probabilities are used to construct weighting estimators of the treatment effect.

1.3 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):

$$\hat{\tau} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij}T_{ij}Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij}T_{ij}} - \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij}(1-T_{ij})Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij}(1-T_{ij})} \tag{1.10}$$

These are weighted averages of the outcome among the treated and non-treated units, where the first term is an estimate of μ_1, and the second term estimates μ_0 (see Equation (1.1)). 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:

$$\begin{aligned} E[\hat{\omega}_{ij}T_{ij}\boldsymbol{X}_{ij}] = E[\hat{\omega}_{ij}(1-T_{ij})\boldsymbol{X}_{ij}] = E[\boldsymbol{X}_{ij}] \\ E[\hat{\omega}_{ij}T_{ij}\boldsymbol{V}_j] = E[\hat{\omega}_{ij}(1-T_{ij})\boldsymbol{V}_j] = E[\boldsymbol{V}_j] \end{aligned} \tag{1.11}$$

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, $\hat{\tau}_{IPW}$, the weights are defined as:

$$\hat{\omega}_{ij,IPW} = \begin{cases} 1/\hat{\pi}_{ij} & for \quad T_{ij} = 1 \\ 1/(1-\hat{\pi}_{ij}) & for \quad T_{ij} = 0 \end{cases} \tag{1.12}$$

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 (1.12) 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)[b] by choosing a cutoff value c and setting all weights larger than the cutoff value to zero:

$$\hat{\omega}_{ij,IPW-T} = \hat{\omega}_{ij,IPW} \mathbf{1}_{\{\hat{\omega}_{ij,IPW} < c\}}, \tag{1.13}$$

where $\mathbf{1}_{\{\}}$ denotes the indicator function. The trimmed weight $\hat{\omega}_{ij,IPW-T}$ equals $\hat{\omega}_{ij,IPW}$ if $\hat{\omega}_{ij,IPW}$ is smaller than c, and it is 0 otherwise. Trimmed weights are then used to obtain an IPW trimming estimator $\hat{\tau}_{IPW-T}$ for the ATE as in Equation (1.10). Importantly, trimming implies a redefinition of the causal estimand (i.e., ATE in Equation (1.1)), since $\hat{\tau}_{IPW-T}$ 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,

[b] 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.

leaving it up to the analyst to choose an appropriate cutoff. Still, this is not a problem under homogeneous 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):

$$\hat{\omega}_{ij,OW} = \begin{cases} 1 - \hat{\pi}_{ij} & for \quad T_{ij} = 1 \\ \hat{\pi}_{ij} & for \quad T_{ij} = 0 \end{cases} \tag{1.14}$$

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, $\hat{\tau}_{OW}$, 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 $\hat{\tau}_{OW}$ has two desirable features: weighting by overlap weights achieves an exact balance of the covariates between treatment and control groups, and $\hat{\tau}_{OW}$ 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.

1.3.1 Calibration estimator

Importantly, if the ignorability assumption holds (see Equation (1.2)), any set of weights that yields a pseudopopulation where the balance conditions of Equation (1.11) 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 $\hat{\omega}_{ij,CAL}$ must fulfill sample analogs of the balancing conditions for the level-1 covariates

$$\begin{aligned}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}\boldsymbol{X}_{ij} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}(1-T_{ij})\boldsymbol{X}_{ij} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\boldsymbol{X}_{ij}\end{aligned} \tag{1.15}$$

and the level-2 covariates

$$\begin{aligned}\sum_{j=1}^{J}\boldsymbol{V}_j\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij} &= \sum_{j=1}^{J}\boldsymbol{V}_j\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}(1-T_{ij}) \\ &= \sum_{j=1}^{J}n_j\boldsymbol{V}_j\end{aligned} \tag{1.16}$$

Since not all variables in $\boldsymbol{V}_j$ are observed, the empirical balancing condition cannot be directly evaluated. Instead, a sufficient condition for Equation (1.16) is:

$$\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij} = \sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}(1-T_{ij}) = \sum_{i=1}^{n_j}1 = n_j \quad (j=1,...,J) \tag{1.17}$$

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 $\hat{\tau}_{CAL}$ is obtained by inserting the weights $\hat{\omega}_{ij,CAL}$ into Equation (1.10). We now show that the calibration estimator provides unbiased estimates of the ATE if the ignorability assumption is fulfilled.

1.3.2 Unbiasedness of the calibration estimators

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

$$Y_{ij}(t) = \beta_{0,t} + \boldsymbol{X}_{ij}\boldsymbol{\beta}_{\boldsymbol{X},t} + \boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{V},t} + U_{j,t} + e_{ij,t} \quad (t=0,1) \tag{1.18}$$

where $E[U_{j,t}] = E[e_{ij,t}] = 0$. The random intercept is allowed to be correlated with $\boldsymbol{X}$ and $\boldsymbol{V}$, while residuals $e_{ij,t}$ are uncorrelated with these covariates. If the data-generating model in Equation (1.18) holds, it can be shown that the calibration estimator $\hat{\tau}_{CAL}$ provides an unbiased estimate of the ATE. Let us denote by $\boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\mu}_{\boldsymbol{V}}$, the expected values of $\boldsymbol{X}$ and $\boldsymbol{V}$, respectively. The ATE is then obtained

by inserting Equation (1.18) into Equation (1.1) and taking expectations

$$\begin{aligned}\tau &= E[Y_{ij}(1) - Y_{ij}(0)] \\ &= \beta_{0,1} - \beta_{0,0} + \boldsymbol{\mu}_{\boldsymbol{X}}(\boldsymbol{\beta}_{\boldsymbol{X},1} - \boldsymbol{\beta}_{\boldsymbol{X},0}) + \boldsymbol{\mu}_{\boldsymbol{V}}(\boldsymbol{\beta}_{\boldsymbol{V},1} - \boldsymbol{\beta}_{\boldsymbol{V},0})\end{aligned} \tag{1.19}$$

When the ignorability condition of Equation (1.2) holds, and the balancing conditions of Equations (1.15) and (1.17) are met (also note that $\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij} = N$ directly follows), we obtain for the first term in $\hat{\tau}_{CAL}$:

$$\begin{aligned}&E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}Y_{ij}\right] \\ &= E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}Y_{ij}(1)\right] \\ &= E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}(\beta_{0,1} + \boldsymbol{X}_{ij}\boldsymbol{\beta}_{\boldsymbol{X},1} + \boldsymbol{V}_{j}\boldsymbol{\beta}_{\boldsymbol{V},1} + U_{j,1} + e_{ij,1})\right] \\ &= E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\beta_{0,1} + \boldsymbol{X}_{ij}\boldsymbol{\beta}_{\boldsymbol{X},1} + \boldsymbol{V}_{j}\boldsymbol{\beta}_{\boldsymbol{V},1})\right] \\ &= N\left[\beta_{0,1} + \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\beta}_{\boldsymbol{X},1} + \boldsymbol{\mu}_{\boldsymbol{V}}\boldsymbol{\beta}_{\boldsymbol{V},1}\right].\end{aligned} \tag{1.20}$$

Hence, we arrive at

$$E\left[\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}}\right] = \beta_{0,1} + \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\beta}_{\boldsymbol{X},1} + \boldsymbol{\mu}_{\boldsymbol{V}}\boldsymbol{\beta}_{\boldsymbol{V},1} \tag{1.21}$$

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

$$E\left[\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}(1-T_{ij})Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}(1-T_{ij})}\right] = \beta_{0,0} + \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\beta}_{\boldsymbol{X},0} + \boldsymbol{\mu}_{\boldsymbol{V}}\boldsymbol{\beta}_{\boldsymbol{V},0} \tag{1.22}$$

Now, by subtracting Equation (1.22) from Equation (1.21), the ATE in Equation (1.19) 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 (1.15)).

1.3.3 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:

$$\hat{\omega}_{ij,CAL1} = \begin{cases} 1 + n_{0j} \frac{exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}\}}{\sum_{h=1}^{n_j} T_{hj} exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}\}} & for \quad T_{ij} = 1 \\ 1 + n_{1j} \frac{exp\{-\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}\}}{\sum_{h=1}^{n_j} (1-T_{hj}) exp\{-\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}\}} & for \quad T_{ij} = 0 \end{cases} \tag{1.23}$$

where n_{1j} and n_{0j} are the number of treated and control units in the jth 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 $\hat{\omega}_{ij,CAL1}$ 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 ω^*_{ij} (e.g., weights constructed from the propensities of an initial working model) to arrive at the calibration weights:

$$\hat{\omega}_{ij,CAL2} = \begin{cases} n_j \frac{\omega^*_{ij} exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}_1}\}}{\sum_{h=1}^{n_j} \omega^*_{hj} T_{hj} exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}_1}\}} & for \quad T_{ij} = 1 \\ n_j \frac{\omega^*_{ij} exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}_0}\}}{\sum_{h=1}^{n_j} \omega^*_{hj} (1-T_{hj}) exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}_0}\}} & for \quad T_{ij} = 0 \end{cases} \tag{1.24}$$

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 (1.15) and (1.17) (see Appendix B).

In simulations, Kim et al. (2017) found their estimator $\hat{\tau}_{CAL1}$ to be superior to the $\hat{\tau}_{IPW}$ based on random-effects propensities in terms of both bias and variance in all conditions studied. Yang (2018) pit her estimator $\hat{\tau}_{CAL2}$ against $\hat{\tau}_{IPW}$ 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 $\hat{\tau}_{IPW}$ based on a fixed-effects propensity score.

1.3.4 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:

$$\hat{\omega}_{ij,CL} = \begin{cases} n_j \frac{\omega_{ij,IPW}}{\sum_{h=1}^{n_j} T_{hj}\omega_{hj,IPW}} & for \quad T_{ij} = 1 \\ n_j \frac{\omega_{ij,IPW}}{\sum_{h=1}^{n_j} (1-T_{hj})\omega_{hj,IPW}} & for \quad T_{ij} = 0 \end{cases} \tag{1.25}$$

One major limitation of the clustered estimator $\hat{\tau}_{CL}$ is that only the level-2 balance condition (see Equation (1.17)) 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 $\hat{\tau}_{CL}$ 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.

1.4 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 (1.25), 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.

1.4.1 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 uncorrelated (see Figure 1.1). Treatment assignment follows the multilevel logistic random-intercept model:

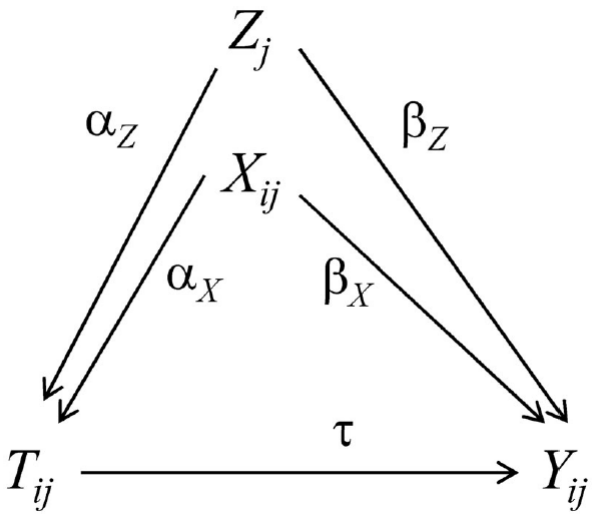
$$T_{ij}^{*} = \alpha_0 + \alpha_X X_{ij} + \alpha_Z Z_j + u_{0j} + \epsilon_{ij} \tag{1.26}$$

where an individual was assigned to treatment $T_{ij} = 1$ if $T_{ij}^{*} > 0$; α_X and α_Z are the regression coefficients; $u_{0j} \sim N(0, \sigma_u^2)$ is the residual at level 2, and $\epsilon_{ij} \sim Logistic(0, 1)$ is the residual at level 1. The intraclass correlation of X (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 $R_{L1}^2 = \left[\alpha_X^2(1 - ICC_X)\right]/Var_{total}$, and at level 2 by $R_{L2}^2 = \left[\alpha_X^2 ICC_X + \alpha_Z^2\right]/Var_{total}$, where $Var_{total} = \alpha_X^2 + \alpha_Z^2 + \sigma_u^2 + \pi^2/3$ is the total variation of the treatment indicator (Snijders & Bosker, 2012; see Rights & Sterba, 2019).
The outcome follows a multilevel random-intercept model:

$$Y_{ij} = \beta_0 + \tau T_{ij} + \beta_X X_{ij} + \beta_Z Z_j + v_{0j} + e_{ij}, \tag{1.27}$$

where β_X and β_Z are regression coefficients, and v_{0j} and e_{ij} are normally distributed residuals at level 2 and level 1, respectively. The treatment effect τ (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

Figure 1.1: Schematic description of the data-generating model of Simulation Study 1.



($\alpha_X = \beta_X = .5$ and $\alpha_Z = \beta_Z = .5$, implying $R^2_{L1} = .04$ and $R^2_{L2} = .07$); 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 $R^2_{L1} = .04$ and $R^2_{L2} = .20$).
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) $\times$ 2 (number of clusters) $\times$ 4 (number of observations per cluster) = 32 conditions, 1,000 simulated datasets were generated. For each simulated dataset, 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 (1.8)). In addition, we implemented two variants of the random-effects (RE) approach by specifying two different multilevel logistic regression models: 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 *lme*4 package (using the *glmer* function).
The propensity score predictions from these models were used to construct IPW

weights (see Equation (1.12)), trimmed IPW weights (IPW-T; see Equation (1.13)), overlap weights (OW; see Equation (1.14)), and cluster-normalized weights (CL; see Equation (1.25)). 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).[c] Thus, 3 (propensity score model) $\times$ 4 (type of weights) = 12 different estimators of the ATE were computed by substituting the various weights into Equation (1.10). Additionally, we implemented the two calibration estimators that were proposed by Kim et al. (2017; see Equation (1.23)) and Yang (2018; see Equation (1.24)). 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.

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.

1.4.2 Results

Table 1.1 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 used for estimating propensity

[c] 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.

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.

Table 1.1: 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	0.29	0.33	0.41	0.51
FE	IPW-T	0	1	3	-1	0.29	0.27	0.31	0.34
FE	OW	0	1	2	-1	0.27	0.23	0.25	0.25
FE	CL	0	**14**	**19**	**19**	0.28	0.31	0.35	0.37
	CAL1	0	3	**9**	**15**	0.28	0.26	0.32	0.36
	CAL2	1	1	3	-1	0.28	0.24	0.27	0.28
RE(XZ)	IPW	-5	**10**	**13**	4	0.31	0.26	0.33	0.49
RE(XZ)	IPW-T	-2	**8**	**9**	5	0.29	0.25	0.3	0.32
RE(XZ)	OW	-5	**-4**	**-4**	**-8**	0.28	0.23	0.25	0.27
RE(XZ)	CL	0	**31**	**36**	**37**	0.28	0.39	0.45	0.47
RE(X)	IPW	**80**	**10**	**57**	**102**	0.85	0.26	0.63	1.07
RE(X)	IPW-T	**80**	**8**	**53**	**94**	0.85	0.25	0.59	0.99
RE(X)	OW	**68**	**-4**	**36**	**67**	0.73	0.23	0.43	0.72
RE(X)	CL	0	**31**	**37**	**39**	0.28	0.39	0.45	0.48
J=100,	n_j=30								
FE	IPW	1	0	0	2	0.18	0.18	0.24	0.36
FE	IPW-T	1	0	0	1	0.17	0.15	0.17	0.21
FE	OW	1	0	0	0	0.15	0.13	0.14	0.15
FE	CL	1	**6**	**8**	**16**	0.18	0.17	0.19	0.27
	CAL1	1	1	-1	0	0.18	0.14	0.16	0.2
	CAL2	1	0	-1	0	0.18	0.13	0.15	0.18
RE(XZ)	IPW	2	**6**	**11**	**8**	0.23	0.16	0.22	0.41
RE(XZ)	IPW-T	**10**	3	3	3	0.19	0.14	0.17	0.2
RE(XZ)	OW	-2	-3	-4	-5	0.15	0.13	0.15	0.16
RE(XZ)	CL	1	**11**	**14**	**22**	0.18	0.18	0.21	0.3
RE(X)	IPW	**53**	**6**	**30**	**62**	0.55	0.16	0.35	0.68
RE(X)	IPW-T	**53**	**3**	**21**	**37**	0.55	0.14	0.26	0.42
RE(X)	OW	**28**	**-3**	**12**	**26**	0.31	0.13	0.19	0.3
RE(X)	CL	1	**11**	**14**	**22**	0.17	0.18	0.21	0.3

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.

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 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 1.2). 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 1.2 shows the bias and Figure 1.3 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.[d]

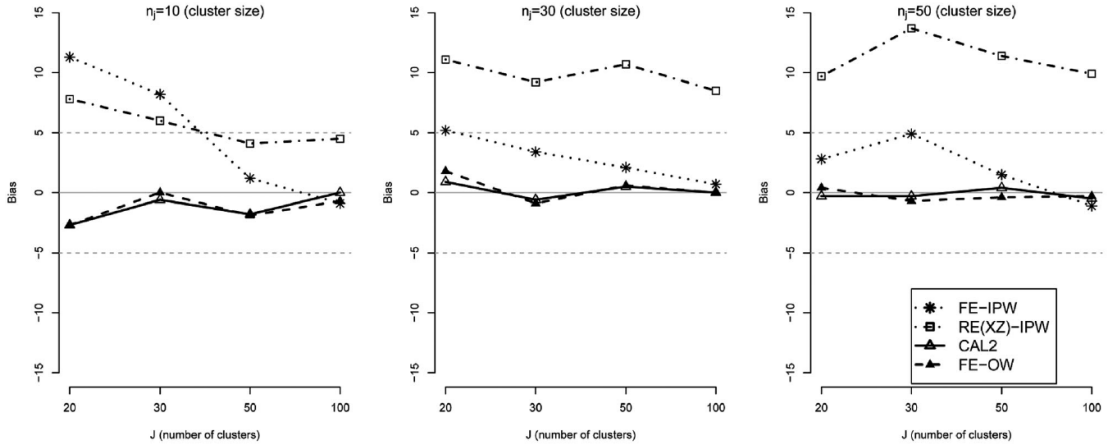
[d]With real educational or organizational data, cluster sizes usually differ across clusters. To evaluate the robustness of the different estimators in the case of unbalanced cluster sizes, we conducted an additional, restricted simulation for a subset of the conditions of Simulation Study 1. We fixed 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 ($n_j = 10$). Overall, the results show that for some approaches (OW and IPW weights,

Table 1.2: 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	0.79	0.65	0.58	0.52
FE	IPW-T		0	3	0	0	0.57	0.38	0.32	0.25
FE	OW		-1	2	-1	0	0.43	0.3	0.23	0.17
	CAL1		**40**	**19**	6	1	0.67	0.47	0.37	0.3
	CAL2		**24**	**11**	2	0	0.59	0.43	0.36	0.3
RE(XZ)	IPW		**-17**	**-25**	**-32**	**-27**	0.79	0.8	0.84	0.76
RE(XZ)	IPW-T		-2	0	**-6**	**-6**	0.54	0.38	0.32	0.26
RE(XZ)	OW		-5	-5	**-7**	**-6**	0.44	0.3	0.24	0.18
RE(X)	IPW		**99**	**71**	**56**	**49**	1.13	0.87	0.73	0.64
RE(X)	IPW-T		**90**	**54**	**34**	**19**	1.03	0.65	0.45	0.31
RE(X)	OW		**84**	**53**	**38**	**26**	0.95	0.61	0.44	0.31
20% Treated										
FE	IPW		0	2	4	0	0.64	0.54	0.46	0.41
FE	IPW-T		1	1	1	0	0.43	0.3	0.25	0.19
FE	OW		1	1	0	0	0.33	0.23	0.18	0.14
	CAL1		**27**	3	1	0	0.5	0.31	0.27	0.22
	CAL2		3	0	1	0	0.4	0.29	0.25	0.21
RE(XZ)	IPW		**-9**	**-11**	**-9**	**-12**	0.63	0.67	0.55	0.57
RE(XZ)	IPW-T		1	2	-1	-3	0.42	0.29	0.24	0.19
RE(XZ)	OW		**-6**	**-7**	**-6**	-4	0.34	0.24	0.19	0.15
RE(X)	IPW		**100**	**72**	**60**	**47**	1.09	0.81	0.7	0.57
RE(X)	IPW-T		**92**	**54**	**36**	**21**	1	0.61	0.43	0.28
RE(X)	OW		**75**	**44**	**31**	**20**	0.82	0.49	0.36	0.25
50% Treated										
FE	IPW		0	0	1	1	0.54	0.48	0.37	0.35
FE	IPW-T		2	0	0	0	0.36	0.25	0.2	0.16
FE	OW		0	0	0	0	0.28	0.19	0.15	0.12
	CAL1		**16**	1	0	0	0.37	0.23	0.2	0.17
	CAL2		1	0	0	0	0.3	0.21	0.18	0.15
RE(XZ)	IPW		5	**10**	**10**	**11**	0.47	0.46	0.38	0.35
RE(XZ)	IPW-T		**6**	4	2	1	0.34	0.24	0.2	0.15
RE(XZ)	OW		**-8**	**-7**	-5	**-4**	0.29	0.21	0.16	0.12
RE(X)	IPW		**102**	**74**	**61**	**48**	1.07	0.81	0.67	0.54
RE(X)	IPW-T		**95**	**55**	**37**	**21**	0.99	0.6	0.41	0.26
RE(X)	OW		**67**	**37**	**26**	**16**	0.73	0.42	0.3	0.2

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.

Figure 1.2: 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).



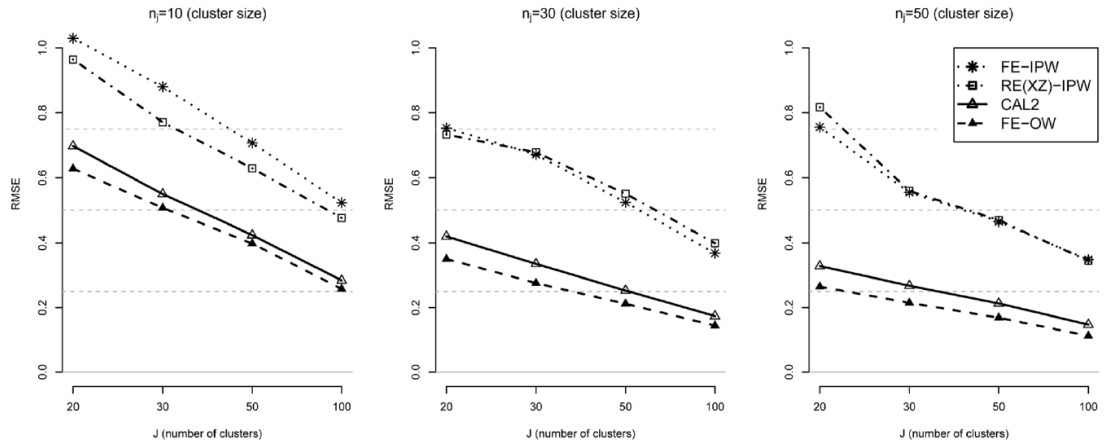
Note: FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates X and Z; IPW = inverse probability weighting; OW = overlap weights; CAL2 = calibration weights of Yang (2018).

1.4.3 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

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).

Figure 1.3: 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).



Note: 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).

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., $n_j = 30$) for obtaining accurate estimates of the treatment effect.

1.5 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 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).

1.5.1 Method

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

$$T_{ij}^{*} = \alpha_0 + \alpha_X X_{ij} + \alpha_Z Z_j + u_{0j} + \epsilon_{ij}, \tag{1.28}$$

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:

$$Y_{ij} = \beta_0 + (\tau_0 + \tau_1 X_{ij} + \tau_2 Z_j + \tau_3 v_{1j}) T_{ij} + \beta_X X_{ij} + \beta_Z Z_j + v_{0j} + e_{ij} \tag{1.29}$$

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.

1.5.2 Results

Table 1.3 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, 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.

Table 1.3: Simulation study 2: Relative bias and relative RMSE as a function of level 2 endogeneity, treatment effect hetero geneity 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**	0.19	0.32	0.2	0.38
FE	IPW-T	1	**12**	0	**32**	0.16	0.27	0.16	0.41
FE	OW	0	**23**	0	**52**	0.14	0.3	0.14	0.55
	CAL1	2	**17**	2	**32**	0.16	0.29	0.17	0.39
	CAL2	1	3	0	**19**	0.16	0.22	0.17	0.28
RE(XZ)	IPW	4	**9**	**33**	**81**	0.16	0.25	0.37	0.84
RE(XZ)	IPW-T	3	**9**	**32**	**81**	0.15	0.24	0.35	0.84
RE(XZ)	OW	-2	**20**	**27**	**98**	0.14	0.29	0.31	1
J=100,	n_j=30								
FE	IPW	-1	0	0	5	0.13	0.21	0.13	0.23
FE	IPW-T	0	**12**	0	**29**	0.09	0.2	0.09	0.32
FE	OW	0	**24**	0	**52**	0.08	0.28	0.08	0.53
	CAL1	0	5	0	9	0.1	0.16	0.1	0.17
	CAL2	0	0	0	5	0.09	0.15	0.09	0.15
RE(XZ)	IPW	2	4	**19**	**43**	0.1	0.18	0.22	0.47
RE(XZ)	IPW-T	-1	**6**	**14**	**47**	0.09	0.17	0.17	0.49
RE(XZ)	OW	-2	**22**	**12**	**72**	0.08	0.26	0.14	0.73
J=100,	n_j=50								
FE	IPW	0	1	0	2	0.1	0.17	0.1	0.2
FE	IPW-T	0	**13**	0	**28**	0.07	0.18	0.07	0.31
FE	OW	0	**25**	0	**53**	0.06	0.28	0.06	0.53
	CAL1	0	3	0	5	0.08	0.13	0.08	0.14
	CAL2	0	0	0	2	0.07	0.12	0.07	0.13
RE(XZ)	IPW	2	4	**14**	**31**	0.08	0.15	0.16	0.35
RE(XZ)	IPW-T	-1	**8**	**9**	**38**	0.07	0.15	0.11	0.4
RE(XZ)	OW	-1	**24**	**8**	**65**	0.06	0.27	0.1	0.66

Note. Endo = endogeneity; Exo = exogeneity; Hom = treatment effect homogeneity; Het = treatment effect heterogeneity; J = number of clus ters; 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)

1.5.3 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.

1.6 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 (1.7)). 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:

$$g(\pi_{ij}) = \gamma_0 + \boldsymbol{X}_{ij}\boldsymbol{\gamma}_{\boldsymbol{X}} + \boldsymbol{V}_j\boldsymbol{\gamma}_{\boldsymbol{V}} + \boldsymbol{X}_{ij}\boldsymbol{V}_j\boldsymbol{\gamma}_{\boldsymbol{XV}} + U_{0j} + \boldsymbol{X}_{ij}\boldsymbol{U}_{1j} \tag{1.30}$$

where U_{1j} is a vector of cluster-specific effects that allow the effects of $\boldsymbol{X}_{ij}$ to vary across clusters, and $\boldsymbol{X}_{ij}\boldsymbol{V}_j$ are the cross-level interactions between the level-1 covariates $\boldsymbol{X}_{ij}$ and the observed and potentially unobserved level-2 covariates $\boldsymbol{V}_j$. The model for the potential outcomes can also be extended to include covariate-by-cluster interactions (see Equation (1.18)):

$$\begin{aligned} Y_{ij}(t) = \beta_{0,t} + \boldsymbol{X}_{ij}\boldsymbol{\beta}_{\boldsymbol{X},t} + \boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{V},t} + \boldsymbol{X}_{ij}\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{XV},t} \\ + U_{0j,t} + \boldsymbol{X}_{ij}\boldsymbol{U}_{1j,t} + e_{ij,t} \quad (t = 0, 1), \end{aligned} \tag{1.31}$$

where $E[\boldsymbol{U}_{1j,t}] = \boldsymbol{0}$, and $E[U_{0j,t}] = E[e_{ij,t}] = 0$. The random effects $U_{0j,t}$ and $\boldsymbol{U}_{1j,t}$ 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:

$$\begin{aligned}\sum_{i=1}^{n_j} \hat{\omega}_{ij,CAL} T_{ij} \boldsymbol{X}_{ij} &= \sum_{i=1}^{n_j} \hat{\omega}_{ij,CAL} (1 - T_{ij}) \boldsymbol{X}_{ij} \\ &= \sum_{i=1}^{n_j} \boldsymbol{X}_{ij} \quad (j = 1, ..., J)\end{aligned} \tag{1.32}$$

$$\sum_{i=1}^{n_j} \hat{\omega}_{ij,CAL} T_{ij} = \sum_{i=1}^{n_j} \hat{\omega}_{ij,CAL} (1 - T_{ij}) = \sum_{i=1}^{n_j} 1 = n_j \quad (j = 1, ..., J) \tag{1.33}$$

In Equation (1.32), 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 (1.33) is identical to the random-intercept scenario (see Equation (1.17)) 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 (1.32) and (1.33) are fulfilled. However, large cluster sizes are likely to be needed to obtain stable estimates of the ATE with calibration 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 (1.8). 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 (1.9) 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.[e] 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.

1.7 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., $n_j = 30$) are needed to provide stable estimates of the ATE when covariate-by-cluster interactions are included.

1.7.1 Method

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

$$T_{ij}^{*} = \alpha_0 + \alpha_X X_{ij} + \alpha_Z Z_j + \alpha_{XZ} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} + \epsilon_{ij} \tag{1.34}$$

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 ϵ_{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 $Var(u_{1j}) = f_{slo} Var(u_{0j})$, and investigated two conditions: with no slope

[e]However, it is still important that all relevant covariates are measured so that the ignorability assumption is met (see Equation (1.2)).

variation (i.e., $f_{slo} = 0$), and with half of the variation of the random intercept (i.e., $f_{slo} = 0.5$).
The equation for the outcome was a multilevel model with a random slope for the covariate X:

$$Y_{ij} = \beta_0 + \tau T_{ij} + \beta_X X_{ij} + \beta_Z Z_j + \beta_{XZ} X_{ij} Z_j + v_{0j} + v_{1j} X_{ij} + e_{ij} \tag{1.35}$$

where v_{0j}, v_{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 τ (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 $Var(v_{1j}) = f_{slo} Var(v_{0j})$, 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.[f] 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) $\times$ 2 (effect of Z and XZ vs. no effect of Z and XZ) $\times$ 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

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

included random slopes (REC(XZ)). The propensity scores were then used to compute IPW, IPW-T, and OW weights. Thus, 4 (propensity score models) $\times$ 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.

1.7.2 Results

In Table 1.4, 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) are 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 1.4), 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 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

Table 1.4: 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**	0.67	0.55	0.38	0.23
FEC	IPW-T		**32**	**24**	**15**	**9**	0.37	0.29	0.19	0.12
FEC	OW		**25**	**19**	**11**	**7**	0.3	0.23	0.15	0.1
	CALC1		**24**	**16**	**7**	**4**	0.33	0.24	0.16	0.1
	CALC2		**24**	**16**	**7**	**3**	0.33	0.25	0.16	0.11
REC(XZ)	IPW		**16**	**14**	**10**	**7**	0.27	0.25	0.18	0.14
REC(XZ)	IPW-T		6	4	1	1	0.21	0.16	0.12	0.09
REC(XZ)	OW		-5	-3	-3	-1	0.18	0.14	0.11	0.07
FE	IPW		2	1	-1	1	0.28	0.25	0.19	0.13
FE	IPW-T		0	1	0	0	0.22	0.16	0.12	0.09
FE	OW		0	1	-1	0	0.17	0.13	0.1	0.07
	CAL1		0	1	-1	0	0.2	0.16	0.13	0.09
	CAL2		0	1	-1	1	0.19	0.15	0.11	0.08
RE(XZ)	IPW		**13**	**13**	**9**	**7**	0.26	0.23	0.18	0.13
RE(XZ)	IPW-T		**6**	5	1	1	0.21	0.16	0.12	0.09
RE(XZ)	OW		-5	-3	-3	-1	0.18	0.14	0.11	0.07
Data generated with random slopes										
FEC	IPW		**103**	**86**	**67**	**43**	1.05	0.89	0.7	0.47
FEC	IPW-T		**40**	**26**	**16**	**8**	0.44	0.3	0.21	0.12
FEC	OW		**30**	**20**	**13**	**7**	0.34	0.24	0.17	0.1
	CALC1		**54**	**38**	**24**	**12**	0.6	0.44	0.3	0.18
	CALC2		**39**	**27**	**17**	**8**	0.47	0.35	0.25	0.16
REC(XZ)	IPW		**47**	**43**	**37**	**31**	0.59	0.54	0.54	0.38
REC(XZ)	IPW-T		4	2	2	1	0.21	0.17	0.13	0.09
REC(XZ)	OW		**-8**	**-7**	-3	-1	0.2	0.16	0.12	0.08
FE	IPW		**80**	**80**	**81**	**80**	0.89	0.88	0.88	0.85
FE	IPW-T		**84**	**84**	**86**	**83**	0.89	0.88	0.89	0.86
FE	OW		**77**	**76**	**78**	**76**	0.81	0.8	0.81	0.78
	CAL1		**87**	**89**	**92**	**92**	0.91	0.92	0.95	0.94
	CAL2		**86**	**86**	**88**	**86**	0.9	0.9	0.91	0.89
RE(XZ)	IPW		**96**	**94**	**92**	**86**	1.02	1	0.97	0.9
RE(XZ)	IPW-T		**93**	**91**	**91**	**86**	0.98	0.95	0.94	0.88
RE(XZ)	OW		**76**	**76**	**78**	**76**	0.8	0.79	0.81	0.78

Note. n_j = cluster size; FE = fixed-effects propensity scores; RE(XZ) = random-effects propensity scores with covariates X and Z; FEC = fixedeffects 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-T¼inverse 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.

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 1.5, 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).[g] 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.

1.7.3 Summary and discussion

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

[g]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.

Table 1.5: 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**	0.7	0.56	0.41
FEC	IPW-T	**16**	**10**	5	0.2	0.16	0.1
FEC	OW	**12**	**8**	4	0.2	0.13	0.08
	CAL1	**22**	**15**	**9**	0.3	0.22	0.15
	CAL2	**22**	**14**	**8**	0.3	0.22	0.16
REC(XZ)	IPW	**36**	**33**	**26**	0.5	0.42	0.43
REC(XZ)	IPW-T	**6**	3	2	0.2	0.13	0.09
REC(XZ)	OW	**-4**	-3	-1	0.2	0.11	0.07
FE	IPW	**-10**	**-6**	-4	0.3	0.22	0.17
FE	IPW-T	0	-1	0	0.2	0.13	0.11
FE	OW	0	0	0	0.1	0.1	0.07
	CAL1	0	-1	0	0.2	0.13	0.09
	CAL2	0	-1	0	0.2	0.12	0.08
RE(XZ)	IPW	1	2	1	0.2	0.18	0.15
RE(XZ)	IPW-T	4	2	1	0.2	0.13	1
RE(XZ)	OW	-4	-2	-1	0.1	0.1	0.07
Only random slopes in outcome model							
FEC	IPW	**62**	**46**	**27**	0.7	0.5	0.31
FEC	IPW-T	**26**	**16**	**8**	0.3	0.21	0.13
FEC	OW	**20**	**13**	**7**	0.2	0.17	0.1
	CAL1	**22**	**15**	**6**	0.3	0.24	0.14
	CAL2	**22**	**14**	**5**	0.3	0.23	0.14
REC(XZ)	IPW	**16**	**16**	**10**	0.3	0.27	0.21
REC(XZ)	IPW-T	4	2	1	0.2	0.15	0.1
REC(XZ)	OW	-4	-2	-1	0.2	0.12	0.08
FE	IPW	2	4	0	0.4	0.3	0.24
FE	IPW-T	2	1	0	0.2	0.16	0.11
FE	OW	1	1	0	0.2	0.12	0.09
	CAL1	0	1	0	0.2	0.19	0.14
	CAL2	1	1	0	0.2	0.17	0.12
RE(XZ)	IPW	**14**	**14**	**9**	0.4	0.29	0.22
RE(XZ)	IPW-T	5	3	1	0.2	0.16	0.11
RE(XZ)	OW	-3	-2	-1	0.2	0.13	0.09

Note. n_j = cluster size; FE = fixed-effects propensity scores; RE(XZ) = ran dom-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, 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., $\alpha_{XZ} = \beta_{XZ} = .5s$, and the slope variation, i.e., $Var(u_{1j}) = Var(v_{1j}) = .5s$, in the treatment as well as the outcome model by setting $s = 0, .2, .4, .6, .8$, and 1. Figure 1.4 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.[h]

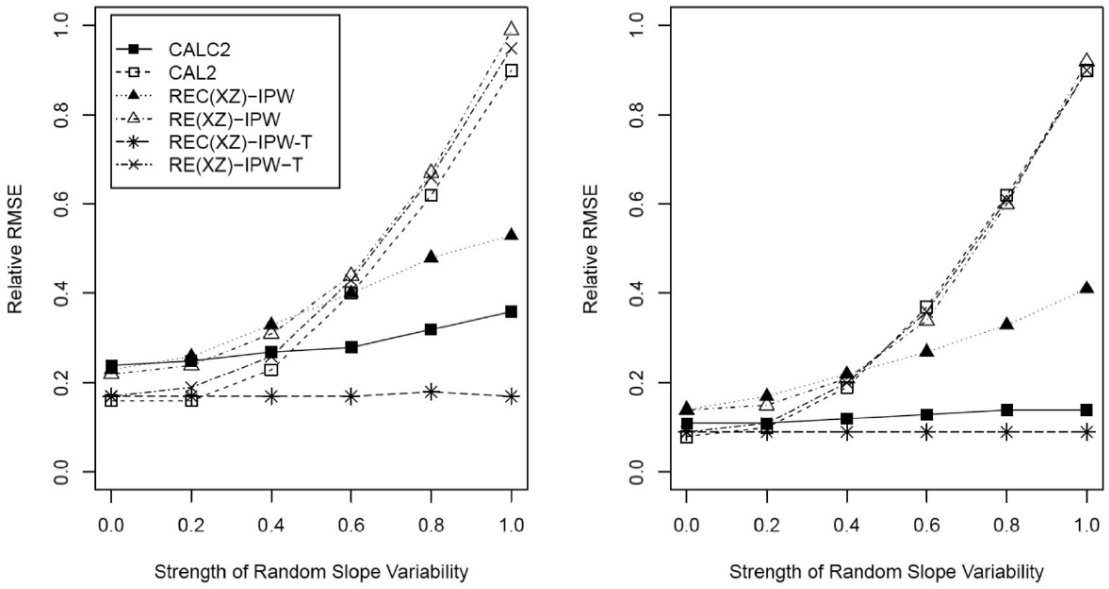
1.8 Inclusion of survey weights

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}

[h]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).

Figure 1.4: 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).



Note: 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.

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 (1.12)). However, sampling weights have to be included in the weighted treatment effect estimate:

$$\hat{\tau} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij}T_{ij}Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij}T_{ij}} - \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij}(1-T_{ij})Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij}(1-T_{ij})} \tag{1.36}$$

In the computation of calibration weights, the balancing conditions now also include sampling weights. Equations (1.15) and (1.17) are modified to

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij,CAL}T_{ij}\boldsymbol{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij,CAL}(1-T_{ij})\boldsymbol{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\boldsymbol{X}_{ij} \tag{1.37}$$

$$\sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij,CAL}T_{ij} = \sum_{i=1}^{n_j} w_{ij}\hat{\omega}_{ij,CAL}(1-T_{ij}) = \sum_{i=1}^{n_j} w_{ij} \quad (j=1,...,J) \tag{1.38}$$

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

$$\hat{\omega}_{ij,CAL2} = \begin{cases} (\sum_{h=1}^{n_j} w_{hj})\frac{w_{ij}\omega^*_{ij}exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}_1\}}{\sum_{h=1}^{n_j} w_{hj}\omega^*_{hj}T_{hj}exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}_1\}} & for \quad T_{ij}=1 \\ (\sum_{h=1}^{n_j} w_{hj})\frac{w_{ij}\omega^*_{ij}exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}_0\}}{\sum_{h=1}^{n_j} w_{hj}\omega^*_{hj}(1-T_{hj})exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}_0\}} & for \quad T_{ij}=0 \end{cases} \tag{1.39}$$

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

1.9 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 non immigrant 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 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 (1.8)) only included the main effects of the level-1 covariates; the model for the RE propensities (see Equation (1.9)) 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 1.6). 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 1.7 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.

Table 1.6: 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	–25.6	5	–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	–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	–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
	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).

1.10 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

Table 1.7: Differences of the estimators under M3b

Method	Weight	Naïve		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***	-4.2	2.9	3.4				
	IPW-T	-41.3***	-12.5*	-5.4	-4.9*	-8.3**			
	OW	-38.9***	-10.2*	-3	-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	4.5	1	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.

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 upweight 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 it 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).

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 $\hat{\omega}_{ij,CAL1}$ (see Equation (1.23)) in the calibration estimator of Kim et al. (2017). Kim et al. (2017) derived the estimating equations for the calibration weights $\hat{\omega}_{ij,CAL1} = \omega_{ij,CAL1}(\hat{\boldsymbol{\lambda}})$ 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 (1.15) in Kim et al., 2017) for $\hat{\boldsymbol{\lambda}}$ can be rewritten as (set $\phi_1 = -\hat{\boldsymbol{\lambda}}$ in the notation of Kim

et al. 2017, and also use Equations (1.13) and (1.14) in Kim et al., 2017)

$$\begin{aligned}&\sum_{j=1}^{J}\sum_{i=1}^{n_j} T_{ij}\left\{1+n_{0j}\frac{exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}\}}{\sum_{h=1}^{n_j} T_{hj}exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}\}}\right\}\\&=\sum_{j=1}^{J}\sum_{i=1}^{n_j}(1-T_{ij})\left\{1+n_{1j}\frac{exp\{-\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}\}}{\sum_{h=1}^{n_j}(1-T_{hj})exp\{-\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}\}}\right\}\end{aligned} \tag{1.40}$$

Calibration weights are then computed as

$$\hat{\omega}_{ij,CAL1}=\begin{cases}1+n_{0j}\frac{exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}\}}{\sum_{h=1}^{n_j} T_{hj}exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}\}} & for \quad T_{ij}=1\\ 1+n_{1j}\frac{exp\{-\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}\}}{\sum_{h=1}^{n_j}(1-T_{hj})exp\{-\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}\}} & for \quad T_{ij}=0\end{cases} \tag{1.41}$$

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 $\hat{\omega}_{ij,CAL2}$ (see Equation (1.24)) is obtained. Yang (2018) starts from an initial vector of weights ω^*_{ij}. Calibration weights $\hat{\omega}_{ij,CAL2}$ are constructed by minimizing the Kullback-Leibler information

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\omega_{ij,CAL2}log\frac{\omega_{ij,CAL2}}{\omega^*_{ij}} \tag{1.42}$$

subject to calibration conditions defined by Equations (1.15) and (1.17) as side conditions. Using the Lagrange multipliers technique (Yang, 2018), the calibration weights $\hat{\omega}_{ij,CAL2}=\omega_{ij,CAL2}(\hat{\boldsymbol{\lambda}}_0,\hat{\boldsymbol{\lambda}}_1)$ are given as (see Equation (1.10) in Yang, 2018)

$$\hat{\omega}_{ij,CAL2}=\begin{cases}n_j\frac{\omega^*_{ij}exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}_1}\}}{\sum_{h=1}^{n_j}\omega^*_{hj}T_{hj}exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}_1}\}} & for \quad T_{ij}=1\\ n_j\frac{\omega^*_{ij}exp\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}_0}\}}{\sum_{h=1}^{n_j}\omega^*_{hj}(1-T_{hj})exp\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}_0}\}} & for \quad T_{ij}=0\end{cases} \tag{1.43}$$

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 (1.11) in Yang, 2018)

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\omega_{ij,CAL2}(\hat{\boldsymbol{\lambda}}_0,\hat{\boldsymbol{\lambda}}_1)T_{ij}\boldsymbol{X}_{ij}=\sum_{j=1}^{J}\sum_{i=1}^{n_j}\boldsymbol{X}_{ij} \tag{1.44}$$

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\omega_{ij,CAL2}(\hat{\boldsymbol{\lambda}}_0,\hat{\boldsymbol{\lambda}}_1)(1-T_{ij})\boldsymbol{X}_{ij}=\sum_{j=1}^{J}\sum_{i=1}^{n_j}\boldsymbol{X}_{ij} \tag{1.45}$$

For p covariates $\boldsymbol{X}_{ij}$, the vectors $\hat{\boldsymbol{\lambda}}_0$ and $\hat{\boldsymbol{\lambda}}_1$ are both of length p, and there are $2p$ nonlinear equations in Equations (1.44) and (1.45). Note that Equations (1.44) and (1.45) can be independently solved for $\hat{\boldsymbol{\lambda}}_0$ and $\hat{\boldsymbol{\lambda}}_1$ because $\omega_{ij,CAL2}(\hat{\boldsymbol{\lambda}}_0,\hat{\boldsymbol{\lambda}}_1)$ is only a function of $\hat{\boldsymbol{\lambda}}_1$ in Equation (1.44) and $\omega_{ij,CAL2}(\hat{\boldsymbol{\lambda}}_0,\hat{\boldsymbol{\lambda}}_1)$ is only a function of $\hat{\boldsymbol{\lambda}}_0$ in Equation (1.45).

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 (1.31), the population ATE $\tau = E(\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{Y_{ij}(1)-Y_{ij}(0)\})$ is given as

$$\begin{aligned}\tau = {} & \beta_{0,1}-\beta_{0,0}+E(\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\boldsymbol{X}_{ij})(\boldsymbol{\beta}_{\boldsymbol{X},1}-\boldsymbol{\beta}_{\boldsymbol{X},0})\\ & +E(\frac{1}{N}\sum_{j=1}^{J}n_j\boldsymbol{V}_j)(\boldsymbol{\beta}_{\boldsymbol{V},1}-\boldsymbol{\beta}_{\boldsymbol{V},0})\\ & +E(\frac{1}{N}\sum_{j=1}^{J}\boldsymbol{V}_j\sum_{i=1}^{n_j}\boldsymbol{X}_{ij})(\boldsymbol{\beta}_{\boldsymbol{XV},1}-\boldsymbol{\beta}_{\boldsymbol{XV},0})\\ & +E(\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\boldsymbol{X}_{ij}(\boldsymbol{U}_{1j,1}-\boldsymbol{U}_{1j,0}))\end{aligned} \tag{1.46}$$

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

$$
\begin{aligned}
&E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}Y_{ij}\right] \\
&= E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}Y_{ij}(1)\right] \\
&= E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}(\beta_{0,1}+\boldsymbol{X}_{ij}\boldsymbol{\beta}_{\boldsymbol{X},1}+\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{V},1}+\boldsymbol{X}_{ij}\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{XV},1}+U_{0j,1}+\boldsymbol{X}_{ij}\boldsymbol{U}_{1j,1}+e_{ij,1})\right] \\
&= E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}T_{ij}(\beta_{0,1}+\boldsymbol{X}_{ij}\boldsymbol{\beta}_{\boldsymbol{X},1}+\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{V},1}+\boldsymbol{X}_{ij}\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{XV},1}+\boldsymbol{X}_{ij}\boldsymbol{U}_{1j,1})\right] \\
&= N\beta_{0,1}+E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\boldsymbol{X}_{ij}\boldsymbol{\beta}_{\boldsymbol{X},1}+\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{V},1}+\boldsymbol{X}_{ij}\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{XV},1}+\boldsymbol{X}_{ij}\boldsymbol{U}_{1j,1})\right]
\end{aligned}
\tag{1.47}
$$

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

$$
\begin{aligned}
&E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}(1-T_{ij})Y_{ij}\right] \\
&= E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{ij,CAL}(1-T_{ij})Y_{ij}(0)\right] \\
&= N\beta_{0,0}+E\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\boldsymbol{X}_{ij}\boldsymbol{\beta}_{\boldsymbol{X},0}+\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{V},0}+\boldsymbol{X}_{ij}\boldsymbol{V}_j\boldsymbol{\beta}_{\boldsymbol{XV},0}+\boldsymbol{X}_{ij}\boldsymbol{U}_{1j,0})\right]
\end{aligned}
\tag{1.48}
$$

Hence, by using Equations (1.47) and (1.48), the expected value of $\hat{\tau}_{CAL}$ equals the population quantity τ that is given by Equation (1.46).

Chapter 2

Multiple Treatment Effect Estimation with Propensity Score Weighting for Two-Level Data

Co-authored with Oliver Lüdtke

Motivation

Multicategorical variables found in or constructed from observational datasets are often the target of studies comparing average outcomes across categories. For instance, researchers may want to estimate and compare the effects of various treatments on some clinical outcome (e.g., Dingena Spreeuwenberg et al., 2010), of various educational strategies on some developmental outcome (e.g., Gupta & Simonsen, 2010), or of various policies on some behavior (e.g., Gibbs et al., 2019). To generalize to populations of interest, treatment effect estimation must be controlled by balancing the distribution of confounders in the various categories (Austin, 2011; Rosenbaum & Rubin, 1983). Guidance for such controlled treatment effect estimation with binary treatments, or with multicategorical treatments that have been collapsed to binary form, is abundant in the literature (e.g., Morgan & Winship, 2014). Research on multicategorical treatment effect estimation with i.i.d. samples is also available (e.g., McCaffrey et al., 2012; Yang et al., 2016; Lopez & Gutman, 2017). Nevertheless, the case of individual-level, multicategorical treatment effects in a context of clustered data, such as students clustered in schools or patients clustered in hospitals, has received little attention (e.g., Hu et al., 2022).

Propensity score weighting (Rosenbaum & Rubin, 1983; Imbens, 2000) is a relatively-simple, flexible and robust method to achieve balance in confounders.

The ease with which survey weights can be combined with the balancing weights has long made this family of methods attractive (e.g., Leite et al., 2015), as the complex structure of survey data can be a key challenge for analysts (Liou & Hung, 2014). But recent literature has also shown that two-level propensity score weighting is reasonably robust to misspecification (Li et al., 2013; Fuentes et al., 2021) and can be further stabilized to address its main weaknesses (Li, et al., 2018; Kim et al., 2017; Yang, 2018). Nevertheless, the application of these methods to multicategorical treatment effect estimation with clustered data has yet to be demonstrated in the literature.

This article treats the case of multicategorical treatment effect estimation with two-level data explicitly. We implement multicategorical versions of the traditional logistic random-effects and logistic fixed-effects estimators of the propensity score (Hong & Raudenbush, 2006), as well as the multicategorical extension of a more recent, calibration estimator (Yang, 2018), and address two potential pitfalls for their application in a multicategorical context.

First, as has been pointed out elsewhere (e.g., Li & Li, 2018), treatment effect estimation with more than two categories is particularly susceptible to overlap issues, because the difficulty of finding comparable units typically grows with the number of categories. Clustered data settings compound this problem, since the overlap condition is even less likely to be fulfilled within each of the cluster samples. While estimators are available that allow for sharing of information across clusters, such that clusters with few comparable units may benefit from the information of others, observational samples often include clusters where not every category is present (e.g., due to a strong selection mechanism, a small cluster size or a large number of categories). Our simulation studies show that such clusters should be flagged as potential high-leverage and/or outlying level-two units, tested, and possibly removed from the sample to reduce bias.

Second, under strong selection, small cluster sizes and many categories, some categories are likely to be rare within clusters. Comparisons with rare categories are prone to bias in the two-category literature (e.g., Fuentes et al., 2021), and our simulation studies show there are limits to the number of categories that can be accurately compared given typical cluster sizes (e.g., school or classroom sample sizes).

The next section introduces the theory of propensity score weighting for multicategorical settings (Yang et al., 2016; Imbens, 2000) and outlines the estimation problem at hand. Section 2 presents three simulation studies where, under various multicategorical conditions, we test the robustness of the traditional fixed-effects and random-effects propensity score weighting estimators (Hong & Raudenbush, 2006) and the calibration estimator of Yang (2018).

Section 3 implements these estimation strategies in a worked example, using student background variables from the Programme for International Student Assessment. Section 4 concludes.

2.1 Propensity score weighting theory

Consider the mock sample in Table 2.1, which contains $M = 9$ units grouped into $J = 3$ clusters, with units indexed $i = 1, 2, 3$ and clusters $j = 1, 2, 3$. The units in this sample have been exposed to one of three conditions: treatment 1 ($A_{ij} = 1$) or treatment 2 ($A_{ij} = 2$) or, say, no treatment at all ($A_{ij} = 0$). Taking the units with $A_{ij} = 0$ as a baseline group, an analyst may be interested in estimating the effect of treatments 1 and 2 on an outcome Y_{ij}, and may further suspect that X_{ij}, a variable at the unit level ("level one" or L1), and V_j, a variable at the cluster level ("level two" or L2), are confounders for these treatment effects. Under the potential outcomes framework (Imbens & Rubin, 2015), each unit has three potential outcomes: $Y_{ij}(0)$, $Y_{ij}(1)$ and $Y_{ij}(2)$. These are the values of Y_{ij} that are or would have been observed if the unit received or would have received treatments 0, 1 and 2, respectively. The observed outcome is therefore:

$$Y_{ij} = Y_{ij}(0)A_{0,ij} + Y_{ij}(1)A_{1,ij} + Y_{ij}(2)A_{2,ij} \tag{2.1}$$

where $A_{0,ij}$, $A_{1,ij}$ and $A_{2,ij}$ are binary indicators constructed from the categorical A_{ij} (see Table 2.1). Consider now the population quantities we would like to estimate:

$$ATE_1 = E\left[Y_{ij}(1) - Y_{ij}(0)\right] = \mu_1 - \mu_0 \tag{2.2}$$

$$ATE_2 = E\left[Y_{ij}(2) - Y_{ij}(0)\right] = \mu_2 - \mu_0 \tag{2.3}$$

where the second equality makes clear that estimators of the expected difference can be constructed from estimators of the potential outcome expectations μ_0, μ_1 and μ_2. The mock sample of Table 2.1 illustrates the main obstacle for estimating these expected potential outcomes. As Equation (2.1) indicated, the columns $Y_{ij}(0)$, $Y_{ij}(1)$ and $Y_{ij}(2)$ of the table are only partially observed, since units are only observed in their own treatment status, and our task is then to estimate μ_0, μ_1 and μ_2 despite the missing data. If the data are "missing at random" (Rosenbaum & Rubin, 1983), that is, if their missingness is unrelated to the potential outcomes themselves given a set of level-one covariates $\boldsymbol{X}_{ij}$ and level-two covariates $\boldsymbol{V}_j$:

$$A_{a,ij} \perp Y_{ij}(a)|\boldsymbol{X}_{ij}, \boldsymbol{V}_j \quad for \quad a \in (0,1,2) \tag{2.4}$$

then we can hope to recover the average potential outcomes from the observed Y_{ij}. Equation (2.4) is known as the weak unconfoundedness assumption (Imbens, 2000; Yang et al., 2016). Propensity score weighting accomplishes this by rebalancing the distribution of the values we do observe to match the distribution of the full columns (Li et al., 2018), conditional on relevant observed covariates and, as will be detailed later, conditional on unobserved cluster-level information (Arpino & Mealli, 2011; Yang, 2018). The generalized propensity score weighting scheme of Imbens (2000) for a multicategorical comparison proceeds as follows.

Table 2.1: An artificial sample with three clusters of three units

j	**i**	**A**	**A0**	**A1**	**A2**	**Y**	**Y(0)**	**Y(1)**	**Y(2)**	**X**	**V**
1	1	2	0	0	1	515			515	2.5	1.5
1	2	0	1	0	0	541	541			3	1.5
1	3	1	0	1	0	510		510		2.1	1.5
2	1	0	1	0	0	537	537			3.5	0.7
2	2	0	1	0	0	545	545			2.7	0.7
2	3	1	0	1	0	512		512		2.3	0.7
3	1	2	0	0	1	530			530	2.8	1.2
3	2	0	1	0	0	520	520			2.8	1.2
3	3	0	1	0	0	526	526			3.2	1.2

Continuing with our three-category example, a propensity score $p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)$ is the probability that unit i in cluster j received treatment $a \in (0,1,2)$, conditional on some unit-level and cluster-level characteristics $(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)$:

$$p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j) = p(A_{a,ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{V}_j) \quad for \quad a \in (0,1,2) \tag{2.5}$$

Note that the scores p_0, p_1 and p_2 for our example are simply regressions of the indicator variables A_0, A_1 and A_2 of Table 2.1 on the covariates (X_{ij},V_j). Indeed, they are a summary of the confounding information in (X_{ij},V_j) and, under weak unconfoundedness, the potential outcome expectations we need can be recovered from the observed outcome Y_{ij} by conditioning on these propensity scores, rather than on the full list of relevant covariates (i.e., unlike in Equation (2.4)):

$$\mu_a = E[E[Y_{ij}|A_{a,ij} = 1, p_{a,ij}]] \quad for \quad a \in (0,1,2) \tag{2.6}$$

In words, for individuals that received treatment a, the expectations of the outcome Y_{ij} within various levels of $p_{a,ij}$ can be averaged to obtain μ_a. As an estimation method, this strategy of averaging conditional averages is known as

propensity score stratification or subclassification (Rosenbaum & Rubin, 1983). Alternatively, the conditioning can be accomplished through weighting with the inverse propensity score since, for each category $a \in (0, 1, 2)$:

$$\begin{aligned} E\left[\frac{A_{a,ij}Y_{ij}}{p_{a,ij}}\right] &= E\left[\frac{A_{a,ij}Y_{ij}(a)}{p_{a,ij}}\right] = E\left[E\left[\frac{A_{a,ij}Y_{ij}(a)}{p_{a,ij}}|\boldsymbol{X}_{ij}\boldsymbol{V}_j\right]\right] \\ &= E\left[\frac{E[A_{a,ij}|\boldsymbol{X}_{ij}\boldsymbol{V}_j]E[Y_{ij}(a)|\boldsymbol{X}_{ij}\boldsymbol{V}_j]}{p_{a,ij}}\right] \\ &= E\left[\frac{p_{a,ij}E[Y_{ij}(a)|\boldsymbol{X}_{ij}\boldsymbol{V}_j]}{p_{a,ij}}\right] \\ &= E\left[E[Y_{ij}(a)|\boldsymbol{X}_{ij}\boldsymbol{V}_j]\right] = E[Y_{ij}(a)] = \mu_a \end{aligned} \tag{2.7}$$

Our propensity score weighting estimators for the expected potential outcomes will then be of the form:

$$\hat{\mu}_a = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{a,ij}A_{a,ij}Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{a,ij}A_{a,ij}} \tag{2.8}$$

where $\hat{\omega}_{a,ij} = 1/\hat{p}_{a,ij}$ are the inverse-probability weights constructed from estimated propensity scores and the denominator of Equation 8 normalizes these weights to sum to one. To estimate the ATEs of Equations (2.2) and (2.3), we can then compute the differences $\hat{\mu}_1 - \hat{\mu}_0$ and $\hat{\mu}_2 - \hat{\mu}_0$, respectively. The asymptotic properties of this estimator were derived by Cattaneo (2010), though in a single-level context.

2.1.1 Positivity, overlap and target populations

A necessary condition for all estimation methods that rely on propensity scores is that the scores must be strictly positive (Rosenbaum & Rubin, 1983). For the purposes of our example, this is:

$$0 < p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j) \quad for \quad a \in (0, 1, 2) \tag{2.9}$$

Interpreting this condition is straightforward in treatment effect studies: the impact of a treatment is typically only interesting and meaningful when units could have instead been exposed to some alternative condition (Schafer & Kang, 2008). However, in the context of a controlled descriptive comparison (Li et al., 2013), the condition under study is not manipulable, that is, units were not "assigned" to their category and could never switch or be switched from one category to another. Common examples include comparisons across demographic groups, such as groups with different migration background or age. In such

contexts, the positivity condition simply implies that it is possible (i.e., there is a positive probability) to find units of all caracteristics (i.e., all possible covariate values) in every category.

Even if this condition holds in the population, the sample distributions of the covariates among the different categories will typically overlap only in part. Importantly, the algorithms we will use for estimating treatment effects rely on the units in this overlap (i.e., units that have the most in common with units from other categories) and extrapolate to units outside of it, so that a small overlap may lead to bias under misspecification (Schafer & Kang, 2008).

With clustered data, and particularly when categories are rare, the sample may contain clusters where some categories are not present at all. In the mock sample of Table 2.1, for instance, none of the units of cluster 2 received treatment 2, and none of the units of cluster 3 received treatment 1. Faced with this situation, an analyst may reason that these clusters are only missing units from those categories due to sampling variation, and not due to any other consideration that could render $p_2(X_{ij}, V_j) = 0$ for cluster 2 and $p_1(X_{ij}, V_j) = 0$ for cluster 3 in the population (e.g., some form of discrimination that would make it impossible for individuals of certain categories to ever attend certain schools or be treated in certain hospitals). With this in mind, the analyst may then include cluster 2 in their estimation of μ_0 and μ_1, and cluster 3 in their estimation of μ_0 and μ_2.[a]

Unfortunately, as we demonstrate in the simulation studies of the next section, these choices may lead to bias, and clusters that do not fulfill the positivity condition in sample (i.e., clusters where units from some category are absent) should be suspected as possible high-leverage and/or outlying level-two observations, tested as such, and possibly removed prior to estimation.

Because of the above, in some applications, adhering to the sample positivity condition will mean reducing the sample to a subset of clusters which, like cluster 1 of Table 2.1, contain units from every category. This may seem like a waste of observations, since we could instead consider binary comparisons separately: the clusters that contain units from category 0 and category 1 could be used to estimate a binary comparison of those two groups; and the clusters that contain units from category 0 and category 2 could be used for another binary comparison. So what is gained by insisting on a multicategorical analysis?

First, it is important to understand that these are fundamentally different exercises, aimed at different estimands (Imbens, 2000). Instead of the ATEs of Equations (2.2) and (2.3), a series of binary comparisons would estimate:

[a]Similar considerations may drive the analyist to include cluster 2, which has no category-2 units, in their estimation of μ_2 (and cluster 3, with no category-1 units, in their estimation of μ_1). But this is well-known to be inappropriate in the multilevel literature and we do not discuss it further (see, e.g., Li, Zaslavsky & Landrum, 2013; Arpino & Mealli, 2011).

$$ATE_{1,binary} = E[Y_{ij}(1) - Y_{ij}(0)|A_{ij} \neq 2] \tag{2.10}$$

$$ATE_{1,binary} = E[Y_{ij}(2) - Y_{ij}(0)|A_{ij} \neq 1] \tag{2.11}$$

where A_{ij} is the three-category indicator (see Table 2.1). Consider the estimand of Equation (2.10). Clearly, if the distributions of $Y_{ij}(1)$ and $Y_{ij}(0)$ among category-2 units are different from their distributions among the other two categories, the estimand of Equation (2.10) will differ from the estimand of Equation (2.2). And a similar statement can be made about the distributions of $Y_{ij}(2)$ and $Y_{ij}(0)$ among category-1 units and Equations (2.11) and (2.3). The question is then one of target populations: are we interested in estimating an average treatment effect for all kinds of individuals in the population, or would it suffice to make binary contrasts that do not generalize to individuals that received other treatments? The answer is clear for controlled descriptive comparisons, where units could not have been in any category other than their own and generalizations outside of pairwise comparison are meaningless.[b] The choice for treatment effect studies, on the other hand, will require more careful consideration.

Lastly, we note that multicategorical studies may carry a higher risk of misspecification, since the number of functional forms to specify is higher than in a series of binary comparisons. For instance, a single propensity score is involved in the estimation of the binary treatment effect $ATE_{1,binary}$ of Equation (2.10), but two propensity scores are required to estimate the treatment effect ATE_1 of Equation (2.2): one to estimate μ_0 and one to estimate μ_1. Whether this additional risk is justified will again depend on the particular application.

2.1.2 Estimation of the propensity score

Many algorithms that can retrieve the conditional expectation of Equation (2.5) are available in the literature. Here we focus on the traditional logistic random-effects and logistic fixed-effects methodologies (Hong & Raudenbush, 2006), as well as a more recent strategy that relies on calibration and has been shown to be robust in many scenarios (Yang, 2018; Fuentes et al., 2021). A fundamental concern that is addressed by the fixed-effects and calibration estimators of the propensity score is the possibility that not all relevant

[b] What would it mean, for example, to generalize the difference in academic achievement among two demographic groups to a third one? In contrast, the effect of some intervention A versus no intervention is of interest for a subpopulation that received some intervention B but could have received A instead.

cluster-level information is present in the observed covariates, although these estimators do assume that all individual-level confounders are observed (Arpino & Mealli, 2011). Unbiasedness of the random-effects estimator, in contrast, typically requires that all cluster-level confounders be observed as well (Arpino & Mealli, 2011; Fuentes et al., 2021). Formally, the full list of cluster-level covariates $\boldsymbol{V}_j$ that render the potential outcomes independent of the category indicators in Equation (2.4), may in fact be $\boldsymbol{V}_j = (\boldsymbol{Z}_j, \boldsymbol{W}_j)$, with $\boldsymbol{Z}_j$ the covariates we observe and $\boldsymbol{W}_j$ the covariates we do not observe. The logistic random-effects estimator of the propensity score captures this unobserved information with a random intercept:

$$logit(p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)) = \gamma_{0,RE} + \boldsymbol{X}_{ij}\gamma_{X,RE} + \boldsymbol{Z}_j\gamma_{Z,RE} + U_{0j} \tag{2.12}$$

In this model, the cluster-specific intercepts U_{0j} represent unobserved level-two variation and are assumed Normal, while the slopes $\gamma_{X,RE}$ and $\gamma_{Z,RE}$ and the fixed intercept $\gamma_{0,RE}$ are typical logistic regression parameters. The model can also be extended to consider cross-level interactions, $X_{ij}Z_j\gamma_{XZ,RE}$, and random slopes, $X_{ij}U_{xj}$ (Snijders & Bosker, 2012). Importantly, unless clusters are large, the cluster-specific intercepts U_{0j} cannot be trusted to capture all relevant cluster-level information on their own, and therefore any observed cluster-level confounders Z_j should also be included in the regression (Ebbes et al., 2004). In contrast, a logistic fixed-effects estimator models both observed and unobserved level-two information into an indicator variable for each cluster:

$$logit(p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)) = \gamma_{0,FE} + \boldsymbol{X}_{ij}\gamma_{X,FE} + \sum_{j=1}^{J-1} I_j\gamma_j \tag{2.13}$$

The indicator variables I_j are equal to 1 if unit i belongs to cluster j, and equal to 0 otherwise. Note that only $J-1$ indicator variables are included to avoid the dummy-variable trap (Snijders & Bosker, 2012). Additionally, interactions between covariates and the cluster indicators (i.e., $X_{ij}\sum_{c=1}^{J-1} I_c\gamma_{cx}$) could be included to account for cluster-specific slopes.

Like the fixed-effects estimator, the calibration strategy of Yang (2018) does not take into account the observed cluster-level variables. To adjust for level-two confounding, it relies exclusively on the within-cluster frequencies of the different categories. But, in contrast to the maximum-likelihoood algorithms typically used to estimate the logit models of Equations (2.12) and (2.13), the estimating equations of Yang (2018) are aimed at achieving perfect balance for the unit-level covariates across clusters, and perfect balance for the treatment indicator within clusters: in the resulting weighted sample, the means of all covariates $\boldsymbol{X}_{ij}$ are

identical across categories, and the weighted sums of units from each category are identical within clusters.[c] The calibration estimator achieves this with the weights:

$$\hat{\omega}_{a,ij} = n_j \frac{\omega^*_{ij} exp\left\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}\right\}}{\sum_{h=1}^{n_j} A_{a,hj}\omega^*_{hj} exp\left\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}\right\}} \tag{2.14}$$

where an initial vector of weights ω^*_{ij} must be provided (e.g., fixed-effects propensity score weights, or simply a constant vector), and the algorithm computes the calibration parameter $\hat{\boldsymbol{\lambda}}$ by minimizing a probability distance between the initial vector of weights and the vector of weights $\hat{\omega}_{a,ij}$ that fulfills the balancing equations. More specifically, $\hat{\boldsymbol{\lambda}}$ are Lagrange Multipliers from the minimization of

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{a,ij} ln \frac{\hat{\omega}_{a,ij}}{\omega^*_{ij}} \tag{2.15}$$

which is the Kullback-Leibler distance between ω^*_{ij} and $\hat{\omega}_{a,ij}$; subject to the level-one balance condition:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{a,ij} A_{a,ij} \boldsymbol{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{a,ij} (1 - A_{a,ij}) \boldsymbol{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \boldsymbol{X}_{ij} \tag{2.16}$$

and the level-two balance condition:

$$\sum_{i=1}^{n_j} \hat{\omega}_{a,ij} A_{a,ij} = \sum_{i=1}^{n_j} \hat{\omega}_{a,ij} (1 - A_{a,ij}) = n_j \quad for \quad j = 1, ..., J \tag{2.17}$$

The resulting vector $\hat{\omega}_{a,ij}$ of Equation (2.14) is a set of weights that make the observed outcome of units from category a proportional to the potential outcome $Y_{ij}(a)$, such that $\sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{\omega}_{a,ij} A_{a,ij} Y_{ij}$ is an unbiased estimator of the population mean μ_a (see Equation (2.7)).

2.2 Simulation studies

To study the performance of these estimators under conditions typical to behavioral research, we conducted three Monte Carlo simulation experiments. First, we simulated a three-category process with random intercepts, and challenged the estimators with various strength-of-confounding conditions under different cluster sizes and counts. Under that basic setup, we also demonstrate

[c] In a maximum-likelihood logit, the estimating equations aim at balancing the derivative of the propensity score, rather than individual covariates. For an insightful discussion, see Imai & Ratkovic (2014).

the importance of the within-cluster, sample positivity condition, by comparing estimates computed with the "full" subsamples (i.e., estimating each μ_a using all clusters that contain at least one category-a unit and at least one unit from some other category) against estimates computed with the "reduced" subsample (i.e., only clusters that contain units from every category). In a second study, again with three categories and random intercepts, we tested the robustness of these estimators to less favorable treatment prevalences. Finally, in the third study, we returned to a balanced treated-to-control ratio and tested the estimators' robustness under a growing number of categories.

2.2.1 Simulation 1: strength of confounding

A three-category indicator A_{ij} was generated, such that each category $a \in (0, 1, 2)$ had a propensity score:

$$p_a(\boldsymbol{X}_{ij}, Z_j) = \frac{exp\left\{\alpha_{0a} + X_{1,ij}\alpha_{a,X1} + X_{2,ij}\alpha_{a,X2} + Z_j\alpha_{a,Z} + U_{A,j}\right\}}{\sum_{k=0}^{2} exp\left\{\alpha_{0k} + X_{1,ij}\alpha_{k,X1} + X_{2,ij}\alpha_{k,X2} + Z_j\alpha_{k,Z} + U_{A,j}\right\}} \quad (2.18)$$

where $X_{1,ij}$ and $X_{2,ij}$ are standard Normal, level-one covariates; Z_j is a standard Normal, level-two covariate; and $U_{A,j}$ is a Normal level-two residual, with variance set to achieve a residual ICC of 0.2.

We manipulated the slopes of Equation (2.18) to study nine strength-of-confounding scenarios, defined as follows. First, note that with only one covariate X_{ij} (rather than two) at level one, the total variance of each of the treatment indicators $A_{0,ij}$, $A_{1,ij}$ and $A_{2,ij}$ has the form $Var_{Total} = \alpha_X^2 + \alpha_Z^2 + \sigma_{U_A}^2 + \pi^2/3$ (Snijders & Bosker, 2012). Setting $\alpha_Z = 0$, we computed the three values of α_X that would give at level 1 an $R_{L1}^2 = \alpha_X^2/Var_{Total}$ of 0.05, 0.1 and 0.15 (these are $\alpha_X = 0.4752, 0.6904$, and 0.8701 respectively). Then, setting $\alpha_X = 0$, we computed the three values of α_Z that would give at level 2 an $R_{L2}^2 = \alpha_Z^2/Var_{Total}$ of 0.05, 0.1 and 0.2 (these are $\alpha_Z = 0.4752, 0.6904$, and 1.036 respectively). Crossing the three α_X conditions and the three α_Z conditions gives nine different scenarios. In each condition, the slopes of $X_{1,ij}$ and $X_{2,ij}$ for the baseline category were both set to $-\alpha_X/2$, while for category 1 they were set to $\alpha_X/4$ and $3\alpha_X/4$, and for category 2 to $3\alpha_X/4$ and $\alpha_X/4$. The slope of Z_j was set to $-\alpha_Z$ for the baseline category and to α_Z for categories 1 and 2. The intercepts α_{01} and α_{02} of Equation (2.18) were set to zero, while the intercept of the baseline category, α_{00}, was varied to keep the average probability of the three categories balanced at around 33% under all nine

confounding scenarios.[d]

Three potential outcomes were then generated as:

$$Y_{0,ij} = X_{1,ij}\beta_{X1} + X_{2,ij}\beta_{X2} + Z_j\beta_Z + U_{Y,j} + \epsilon_{ij} \tag{2.19}$$

$$Y_{1,ij} = A_{1,ij}\delta_1 + X_{1,ij}\beta_{X1} + X_{2,ij}\beta_{X2} + Z_j\beta_Z + U_{Y,j} + \epsilon_{ij} \tag{2.20}$$

$$Y_{2,ij} = A_{2,ij}\delta_2 + X_{1,ij}\beta_{X1} + X_{2,ij}\beta_{X2} + Z_j\beta_Z + U_{Y,j} + \epsilon_{ij} \tag{2.21}$$

where $A_{1,ij}$ and $A_{2,ij}$ are binary indicators constructed from the factor variable A_{ij} for categories 1 and 2, respectively; $\delta_1 = 0.5$ and $\delta_2 = 0.3$ are the treatment effects under study; the confounder slopes are set to $\beta_{X1} = \beta_{X2} = \beta_Z = 0.25$; the level-one residual ϵ_{ij} is Normal with mean zero and variance 0.8; and the level-two residual $U_{Y,j}$ is Normal with mean zero and variance 0.2. The "observed" outcome for the simulation is then:

$$Y_{ij} = Y_{0,ij}A_{0,ij} + Y_{1,ij}A_{1,ij} + Y_{2,ij}A_{2,ij} \tag{2.22}$$

where, under our parametrizations, the confounders explain the outcome with approximately an $R^2_{L1} = 0.1$ and $R^2_{L2} = 0.05$, while the total explained variation (i.e., at both levels) of $A_{1,ij}$ and $A_{2,ij}$ for the outcome equation is around 0.045 and 0.016, respectively.[e]

For each of two cluster count conditions ($N = 100$ or 30), and two cluster size conditions ($n_j = 30$ or 10), under each of the nine confounding scenarios, we simulated 1,000 samples and estimated δ_1 and δ_2 with each sample.[f] Tables 2.2 and 2.3 present the bias and RMSE results for three different estimators of δ_1: random-effects propensity score weighting, fixed-effects propensity score weighting and calibration estimation.[g] Note that the estimates here were computed using only the "reduced" subsample of clusters that fulfilled the within-cluster, sample positivity condition (i.e., clusters that contained units from all three categories). The main results are as follows.

In terms of bias, the calibration estimator performs well even under strong confounding with 30 units per cluster, while the random-effects and fixed-effects

[d] With an $R^2_{L1} = .05$ and $R^2_{L2} = .05, .1, .2$, we set $\alpha_{00} = -.16, -.28, -.49$, respectively. With an $R^2_{L1} = .1$ and $R^2_{L2} = .05, .1, .2$, we set $\alpha_{00} = -.21, -.30, -.51$, respectively. And with an $R^2_{L1} = .15$ and $R^2_{L2} = .05, .1, .2$, we set $\alpha_{00} = -.24, -.32, -.52$, respectively.

[e] The explained variations for the outcome equation were computed through simulation, given the parametrizations described in this section.

[f] The results for δ_1 follow. See Tables 2.8 and 2.9 in the Appendix for the results of δ_2.

[g] The table presents relative bias, computed over $R = 1{,}000$ replications as $100 \times \frac{1}{R}\sum_{r=1}^{R}(\hat{\delta}_r - \delta)/\delta$. RMSE is computed as $\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\delta - \hat{\delta}_r)^2}$.

Table 2.2: Simulation Study 1: Relative bias (%) for estimation of δ_1, as a function of the number of clusters, the cluster size and the strength of confounding at level one and level two

	N = 30								N = 100							
	$n_j = 10$				$n_j = 30$				$n_j = 10$				$n_j = 30$			
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Random Effects	0.05	0.354	2.828	4.03	0.05	2.048	4.242	7.63	0.05	1.208	3.219	4.19	0.05	3.032	4.553	5.51
	0.1	-0.34	2.088	6.77	0.1	3.092	5.13	6.98	0.1	1.389	3.115	3.51	0.1	3.82	5.089	5.88
	0.2	1.751	4.255	7.27	0.2	2.753	4.043	8.55	0.2	-1.54	2.21	5.21	0.2	1.62	5.601	5.45
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Fixed Effects	0.05	-0.16	4.577	4.59	0.05	1.504	4.205	6.77	0.05	1.625	3.312	2.38	0.05	2.513	4.362	4.47
	0.1	1.593	4.507	8.13	0.1	3.388	5.153	6.83	0.1	1.939	2.882	2.77	0.1	4.042	5.253	5.99
	0.2	3.226	7.084	9.77	0.2	5.05	6.755	10.7	0.2	1.449	4.092	6	0.2	4.416	7.95	8.32
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Calibration Estimator	0.05	-1.19	2.076	4.55	0.05	-0.5	-0.25	0.28	0.05	0.142	-0.77	1.33	0.05	-0.05	-0.25	0.14
	0.1	0.399	2.611	6.41	0.1	0.06	-0.47	-1.3	0.1	-0.12	0.378	2.45	0.1	0.15	-0.25	0.21
	0.2	-0.53	3.78	7.66	0.2	-0.26	-1	1.35	0.2	-1.06	1.378	6.77	0.2	-0.03	0.696	0.26

Note. R^2 L1 and R^2 L2 are approximate explained variances for the treatment equations, at level one and level two, respectively. N is the number of clusters and n_j the cluster size.

estimators are more easily affected. Nevertheless, under small clusters, the bias behavior of the calibration estimator is closer here to the behavior of the other two estimators than in studies from the binary-treatment literature (Yang, 2018; Fuentes et al., 2021) where, under a treated-to-control ratio around 1, the calibration estimator strongly dominates; this is likely due to the fact that three categories with an evenly-split average probability of 33% are rarer than two categories split at 50%, and the calibration estimator has been observed to struggle under low prevalences (Fuentes et al., 2021). We explore this further in Simulation 2 with rare treatments and in Simulation 3 with more categories.
In terms of RMSE, the calibration estimator dominates across all conditions, followed in most conditions by the random-effects estimator, which typically achieves lower variance than the fixed-effects estimator thanks to the distributional assumption it imposes on cluster intercepts (Snijders & Bosker, 2012).

Cluster quality and the positivity condition in the simulation

As was mentioned before, the estimates of Table 2.2 were computed using the "reduced" subsample of clusters that contain units from every category. We now turn to our discussion of "reduced" versus "full" subsamples, and how the bias of using the latter is due to the quality of the clusters that are included in

Table 2.3: Simulation Study 1: RMSE for estimation of δ_1, as a function of the number of clusters, the cluster size and the strength of confounding at level 1 and level 2

	N = 30								N = 100							
	$n_j = 10$				$n_j = 30$				$n_j = 10$				$n_j = 30$			
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Random Effects	0.05	0.177	0.202	0.22	0.05	0.096	0.111	0.13	0.05	0.091	0.101	0.13	0.05	0.056	0.065	0.08
	0.1	0.191	0.211	0.23	0.1	0.117	0.137	0.15	0.1	0.098	0.119	0.13	0.1	0.067	0.081	0.09
	0.2	0.211	0.249	0.26	0.2	0.154	0.166	0.2	0.2	0.125	0.144	0.15	0.2	0.097	0.109	0.12
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Fixed Effects	0.05	0.189	0.216	0.24	0.05	0.1	0.12	0.14	0.05	0.097	0.116	0.16	0.05	0.055	0.066	0.09
	0.1	0.193	0.227	0.26	0.1	0.115	0.14	0.15	0.1	0.103	0.13	0.16	0.1	0.066	0.079	0.1
	0.2	0.217	0.257	0.27	0.2	0.13	0.147	0.18	0.2	0.119	0.15	0.17	0.2	0.075	0.094	0.11
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Calibration Estimator	0.05	0.161	0.171	0.18	0.05	0.084	0.09	0.09	0.05	0.086	0.093	0.1	0.05	0.046	0.046	0.05
	0.1	0.175	0.18	0.2	0.1	0.091	0.098	0.1	0.1	0.088	0.097	0.1	0.1	0.051	0.055	0.05
	0.2	0.197	0.206	0.21	0.2	0.109	0.11	0.12	0.2	0.107	0.111	0.12	0.2	0.058	0.06	0.06

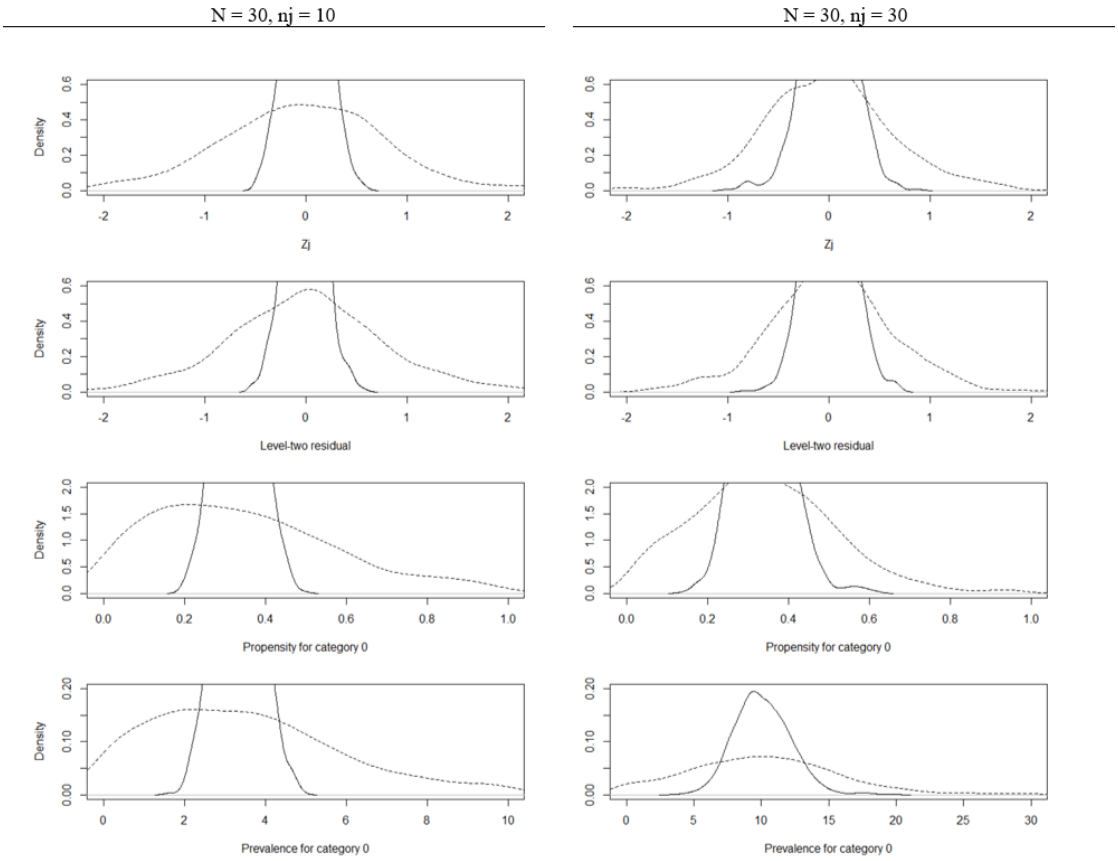
Note. R^2 L1 and R^2 L2 are approximate explained variances for the treatment equations, at level one and level two, respectively. N is the number of clusters and n_j the cluster size.

estimation; specifically, due to the presence of high-leverage and outlying clusters (Anguinis et al., 2013; Langford & Lewis, 1998), and likely also due to within-cluster prevalences, particularly in the case of the calibration estimator (Fuentes et al., 2021).

Under our simulation setup, when very small or very large values for Z_j or $U_{A,j}$ are randomly generated, the affected cluster may fall into violation of the within-cluster, sample positivity condition (i.e., it may lack units from category 0 if $Z_j\alpha_{a,Z} + U_{A,j}$ in Equation (2.18) is sufficiently large, such that $p_0(\boldsymbol{X}_{ij}, Z_j)$ is very small; or lack units from categories 1 or 2 if $Z_j\alpha_{a,Z} + U_{A,j}$ is sufficiently small, such that $p_1(\boldsymbol{X}_{ij}, Z_j)$ and $p_2(\boldsymbol{X}_{ij}, Z_j)$ are very small) and be dropped from estimation. Then, since these high-leverage and outlying clusters tend to be dropped, estimation is carried out mostly with clusters that have mild Z_j and $U_{A,j}$ values. To illustrate this, for the 1,000 Monte Carlo replications of the $R^2_{L1} = 0.15$ and $R^2_{L2} = 0.2$ condition, the top panels of Figure 2.1 present the distributions of the average value of the level-two covariate Z_j, and of the average value of the level-two residual $U_{A,j}$. Notice that the ranges of Z_j and $U_{A,j}$ values are narrower (solid lines) among the clusters that fulfill the within-cluster, sample positivity condition than (dashed lines) among clusters that don't fulfill the condition but contain at least one unit from category 0 and at least one unit from some other category (i.e., the additional clusters from the "full" subsample for estimation of μ_0).[h]

[h] Notice also that, because small clusters can more easily fall into violation of positivity, cluster

Figure 2.1: Distribution of the level-two covariate, the level-two residual, the propensity for category 0 and the prevalence for category 0 over 1,000 Monte Carlo iterations with $R^2_{L1} = 0.15$, $R^2_{L2} = 0.2$.



Note: N is the number of clusters and nj the cluster size. The solid line is the distribution among clusters that fulfilled the within-cluster positivity condition; the dashed line the distribution among clusters that did not fulfill the condition, but had at least one unit from category zero and one unit from some other category.

Table 2.4 presents bias results for estimation of δ_1 using the "full" subsamples. Here, μ_0 was computed using all clusters that had at least one unit from category 0 and at least one unit from some other category. Similarly, μ_1 was computed using all clusters that had at least one unit from category 1 and at least one unit from some other category. For conciseness, we only present the results from the most demanding ($N = 30, n_j = 10$) and least demanding ($N = 100, n_j = 30$) conditions. Notice that there is no condition under which using the full subsamples (rather than the reduced subsamples, like in Table 2.2) pays off. Moreover, violating the within-cluster, sample positivity condition can result in severe bias, particularly under strong level-two confounding and when clusters are small. It is worth noting that, although in some of the more demanding scenarios many clusters did not fulfill the positivity condition (e.g., with 30 clusters of 10 units and $R^2_{L1} = 0.05, R^2_{L2} = 0.2$, an average of 18.7 clusters fulfilled the positivity condition in each Monte Carlo run, with as few as 11 and never more than 26 fulfilling this condition), keeping even a relatively small number of these "bad" clusters in the less demanding scenarios still lead to noticeable bias (e.g., with 100 clusters of 30 units and $R^2_{L1} = 0.15, R^2_{L2} = 0.2$, an average of 88.9 clusters fulfilled the positivity condition in each Monte Carlo run, with as few as 78 and never more than 97 fulfilling this condition).

In addition to the small-sample bias from high-leverage and outlying clusters, the calibration estimator is also likely affected by the lower cluster prevalences of the "full" subsamples. Returning to the example of Figure 2.1, the bottom four panels present the average within-cluster true probability of receiving the baseline treatment, $p_0(\boldsymbol{X}_{ij}, Z_j)$, and the average within-cluster prevalence of category 0. Note that, when the "full" subsamples are used, many more clusters with extreme prevalences appear in each sample. As has been shown in the binary-treatment literature (Fuentes et al., 2021), and as will be shown in the next section for multicategorical estimation, the calibration estimator struggles to control for confounding when many clusters of the sample have extreme prevalences.

Cluster quality and the positivity condition in practice

The lesson from this simulation exercise for real data applications is that clusters that do not fulfill the sample positivity condition warrant careful inspection. In the simplest of cases, substantive knowledge of the data may suffice to determine that the clusters in question belong to a different population (e.g., one where

quality is even higher in the conditions with 10 units per cluster than in the conditions with 30 (see Figure 2.1). This is the reason why, counterintuitively, the random-effects and fixed-effects estimators are slightly more biased with 30 units per cluster than with 10 in many cells of Table 2.2.

Table 2.4: Simulation Study 1: Relative bias (%) in the most and least demanding conditions, using the "full" sample

	N = 30, n_j = 10				N = 100, n_j = 30			
		R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Random Effects	0.05	8.37	10.34	11.2	0.05	4.318	5.569	6.23
	0.1	13.54	14.87	18.5	0.1	8.287	8.581	8.86
	0.2	23.36	24.86	25.7	0.2	13.27	15.63	14.4
		R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Fixed Effects	0.05	8.677	12.46	11.7	0.05	3.787	5.384	5.22
	0.1	16.08	18	20.4	0.1	8.524	8.725	8.97
	0.2	24.9	27.94	28.3	0.2	16.1	17.96	17.2
		R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Calibration Estimator	0.05	6.907	8.952	9.9	0.05	1.165	0.705	0.84
	0.1	13.91	13.97	15.7	0.1	4.446	3.072	3.09
	0.2	19.78	21.7	21.7	0.2	11.26	10.37	8.83

Note. R^2 L1 and R^2 L2 are approximate explained variances for the treatment equations, at level one and level two, respectively. N is the number of clusters and n_j the cluster size.

certain treatments are not available), and can thus be dropped safely or modelled separately.

If theory does not support the assumption that these clusters simply follow a different data-generating process, our simulation results suggest they should be inspected as potentially high-leverage and/or outlying. Rather than directly dropping all clusters that violate positivity (as was done, for simplicity, in the simulation), cautious analysts may first estimate propensity scores using the "full" subsamples, and then test whether any of these clusters have undue influence on the propensity estimates. Many such statistics and tests for detection of influential units are available in the multilevel modeling literature, as is guidance for handling influential units once they are detected (e.g., Anguinis et al., 2013; Langford & Lewis, 1998; Longford, 2001; Hosmer et al., 2013).

2.2.2 Simulation 2: Treatment prevalences

In a second simulation, we manipulated the intercept of the baseline category in Equation (2.18) to give it an average probability of 70%, while categories 1 and 2 were each left with an average probability of 15%.[i] Tables 2.5 and 2.6 present the

[i]With an $R^2_{L1} = .05$ and $R^2_{L2} = .05, .1, .2$, we set $\alpha_{00} = 1.8, 1.94, 2.17$, respectively. With an $R^2_{L1} = .1$ and $R^2_{L2} = .05, .1, .2$, we set $\alpha_{00} = 1.88, 2, 2.24$, respectively. And with an $R^2_{L1} = .15$ and

bias and RMSE results of this exercise.
In terms of bias, the calibration estimator now finds it much more difficult to control for confounding when clusters are small, while the other two estimators are only slightly more biased than in the previous simulation with balanced prevalences. As mentioned above, this is consistent with findings from the binary-treatment literature (Fuentes et al., 2021). Still, the calibration estimator outperforms the random-effects and fixed-effects estimators in the conditions with 30 units per cluster.
In terms of RMSE, the calibration estimator continued to dominate, though the difference against the other two estimators is now much narrower under small clusters (i.e., compared to the conditions with $n_j = 10$ of Table 2.3).
The weakness of the calibration estimator under low prevalences may be due, at least in part, to its handing of clusters where some category has only one unit. Consider a sample of 10-unit clusters, and a cluster in that sample where only one unit belongs to category a. For estimation of μ_a, the level-two balance condition of the calibration estimator requires that the weighted sum of category-a units in the cluster be equal to the cluster size (see Equation (2.17)). Then, the only possible weight that can be assigned to the single category-a unit in that cluster is 10. If many such clusters are present in the sample, the calibration estimator will treat their category-a units as identical (i.e., it will assign them all a weight of 10), regardless of their individual differences, as represented by the observed level-one covariates. In contrast, the fixed-effects and random-effects estimators can differentiate (i.e., assign different weights to) such units and thus reflect their individual differences.

2.2.3 Simulation 3: Number of categories

Finally, we simulated scenarios with three, four, five and six categories, to study the bias behavior of these estimators as the number of categories grows. As the bias was likely to grow quickly with the number of categories, we only considered three, relatively mild confounding scenarios: $R^2_{L1} = R^2_{L2} = 0.05$; $R^2_{L1} = 0.05, R^2_{L2} = 0.1$; and $R^2_{L1} = 0.1, R^2_{L2} = 0.05$. In each confounding scenario, the slopes of $X_{1,ij}$ and $X_{2,ij}$ for the baseline category were again both set to $-\alpha_X/2$; again to $\alpha_X/4$ and $3\alpha_X/4$ for category 1; again to $3\alpha_X/4$ and $\alpha_X/4$ for category 2; to $\alpha_X/3$ and $2\alpha_X/3$ for category 3; to $2\alpha_X/3$ and $\alpha_X/3$ for category 4; and to $\alpha_X/5$ and $4\alpha_X/5$ for category 5. The intercept of the baseline category was manipulated to maintain the categories balanced on average at 33% probability (3 categories), 25% probability (4 categories), 20% probability (5

$R^2_{L2} = .05, .1, .2$, we set $\alpha_{00} = 1.97, 2.08, 2.31$, respectively.

Table 2.5: Simulation Study 2: Relative bias (%) for estimation of δ_1, as a function of the number of clusters, the cluster size and the strength of confounding at level 1 and level 2. Rare treatment (15%)

	N = 30								N = 100							
	$n_j = 10$				$n_j = 30$				$n_j = 10$				$n_j = 30$			
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Random Effects	0.05	0.906	6.607	12.4	0.05	3.181	7.35	11.3	0.05	3.475	5.958	7.84	0.05	4.736	7.696	11.7
	0.1	3.016	6.66	12	0.1	7.506	9.089	12.3	0.1	4.218	6.019	8.72	0.1	6.333	9.268	12.8
	0.2	6.609	7.463	12.8	0.2	7.796	12.82	15.8	0.2	4.404	8.699	10.2	0.2	7.908	11.6	15.5
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Fixed Effects	0.05	1.929	5.758	13.9	0.05	3.049	6.643	11.1	0.05	3.448	5.151	7.25	0.05	3.859	7.187	10.8
	0.1	4.78	8.084	11.6	0.1	5.662	8.002	11.9	0.1	5.494	5.888	8.51	0.1	4.76	8.703	11.6
	0.2	6.839	8.996	12.2	0.2	7.527	12.51	14.7	0.2	3.808	7.843	10.4	0.2	6.841	10.5	14.1
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Calibration Estimator	0.05	3.184	10.18	19.2	0.05	-1.39	-0.06	2.24	0.05	2.977	8.443	16.4	0.05	-0.2	0.312	1.19
	0.1	2.21	8.192	13.9	0.1	0.615	0.191	0.2	0.1	2.391	5.215	10.2	0.1	-0.42	0.143	0.5
	0.2	4.098	6.779	10.7	0.2	0.794	1.339	1.67	0.2	0.644	4.495	6.84	0.2	0.345	-0.04	0.12

Note. R^2 L1 and R^2 L2 are approximate explained variances for the treatment equation, at level one and level two, respectively. N is the number of clusters and n_j the cluster size.

Table 2.6: Simulation Study 2: RMSE for estimation of δ_1, as a function of the number of clusters, the cluster size and the strength of confounding at level 1 and level 2. Rare treatment (15%)

	N = 30								N = 100							
	$n_j = 10$				$n_j = 30$				$n_j = 10$				$n_j = 30$			
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Random Effects	0.05	0.21	0.231	0.25	0.05	0.114	0.126	0.14	0.05	0.112	0.124	0.13	0.05	0.067	0.077	0.09
	0.1	0.224	0.232	0.26	0.1	0.12	0.134	0.15	0.1	0.117	0.132	0.14	0.1	0.074	0.087	0.1
	0.2	0.252	0.261	0.29	0.2	0.146	0.151	0.17	0.2	0.128	0.138	0.15	0.2	0.084	0.096	0.11
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Fixed Effects	0.05	0.216	0.24	0.28	0.05	0.121	0.133	0.15	0.05	0.121	0.14	0.16	0.05	0.068	0.081	0.1
	0.1	0.228	0.248	0.28	0.1	0.125	0.14	0.16	0.1	0.121	0.141	0.16	0.1	0.075	0.089	0.1
	0.2	0.266	0.275	0.31	0.2	0.14	0.156	0.18	0.2	0.136	0.155	0.17	0.2	0.082	0.098	0.11
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Calibration Estimator	0.05	0.208	0.212	0.24	0.05	0.107	0.117	0.13	0.05	0.117	0.128	0.15	0.05	0.059	0.063	0.07
	0.1	0.218	0.217	0.24	0.1	0.111	0.12	0.13	0.1	0.112	0.127	0.14	0.1	0.065	0.067	0.07
	0.2	0.236	0.247	0.27	0.2	0.127	0.124	0.14	0.2	0.122	0.128	0.14	0.2	0.065	0.07	0.07

Note. R^2 L1 and R^2 L2 are approximate explained variances for the treatment equations, at level one and level two, respectively. N is the number of clusters and n_j the cluster size.

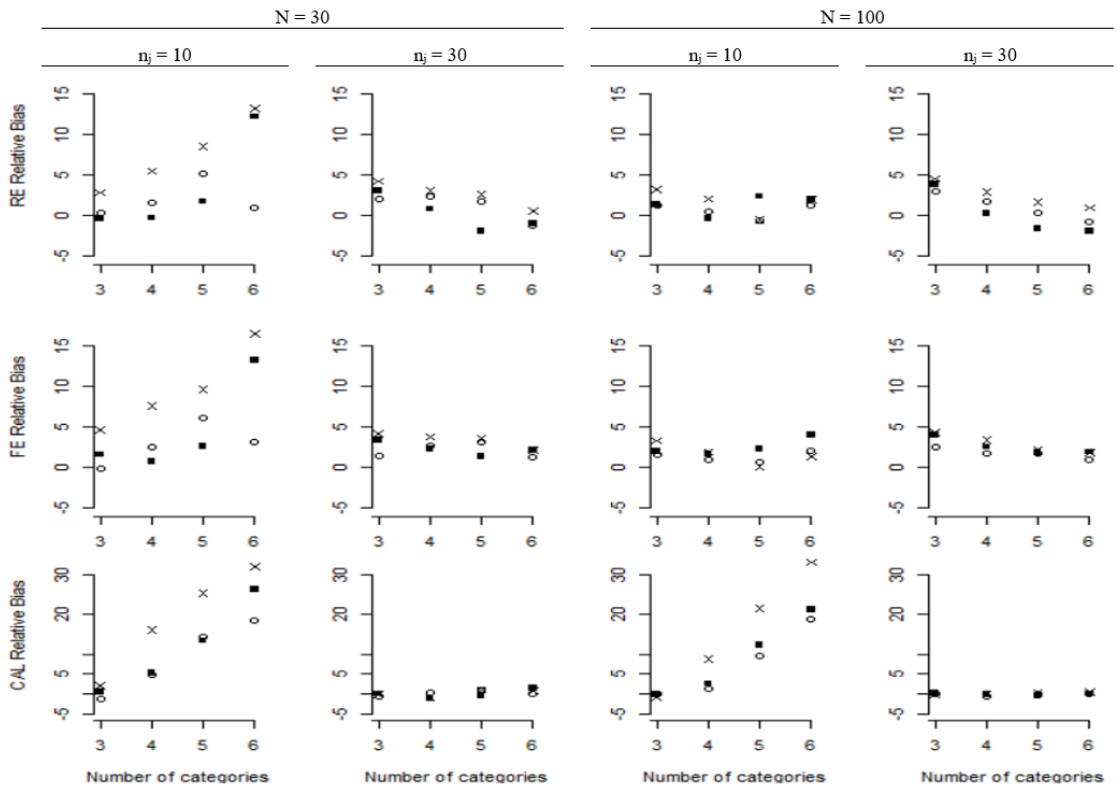
categories) and 16.6% probability (6 categories).[j] The plots in Figure 2.2 present the main results, which are as follows.
With a few small clusters ($N = 30, n_j = 10$), the random-effects and fixed-effects estimators are able to control for mild confounding at both levels with up to 6 categories, though with moderate confounding at level one the limit becomes 5 categories. With moderate confounding at level two, these two estimators fail already at 4 categories. In contrast, under the same cluster count and size scenario, the bias of the calibration estimator rises quickly with the number of categories in all three confounding scenarios. Importantly, this bias does not diminish significantly when the cluster count is larger ($N = 100, n_j = 10$), unlike the bias of the random-effects and fixed-effects estimators which becomes reasonable for up to 6 categories under any of the confounding scenarios.
For all estimators, a cluster size of 30 units was enough to control for confounding with up to 6 categories under all three confounding scenarios.

2.3 Example: Effect of Private and Group Tutoring on Math Outcomes

Kim, Steiner and Lim (2016) used the Singapore PISA 2012 dataset to compare the mathematics scores of students who had received personal tutoring against the scores of those who had not. To demonstrate the use of the propensity score weighting estimators above, in this section we extend the analysis of Kim, Steiner and Lim (2016) to a multicategorical treatment study. Like those authors, we take the students that received no additional tutoring as a baseline group (category 0; $N_0 = 697$), but compare their achievement against that of students that received personal tutoring only (category 1; $N_1 = 663$), students who received additional lessons organized by a commercial company only (category 2; $N_2 = 218$), and students who received both personal tutoring and additional lessons (category 3; $N_3 = 446$). We also emulate the identification strategy of these authors by controlling for the same six student-level covariates: ESCS, whether the student speaks a foreign language at home, whether the father has a university degree or higher, whether the student lives in a two-parent household, whether they have more than 25 books at home, and whether their learning strategy is searching for more information to clarify math problems (rather than repeating examples or thinking about how they relate to everyday problems);

[j] Under each of the three confounding scenarios ((0.05,0.05), (0.05,0.1) and (0.1,0.05)), the intercepts are as follows, respectively. For 4 categories: $\alpha_{00} = -0.26, -0.44, -0.32$. For 5 categories: $\alpha_{00} = -0.378, -0.62, -0.495$. For 6 categories: $\alpha_{00} = -0.32 - 0.54, -0.4$.

Figure 2.2: Simulation Study 3: Relative Bias (%) for Estimation of δ_1 as a Function of the Number of Clusters, the Cluster Size, the Number of Categories, and the Strength of Confounding at Level 1 and Level 2.



Note: FE indicates the fixed-effects estimator, RE is random-effects, and CAL the calibration estimator. N is the number of clusters and nj the cluster size. Circles indicate low confounding at both levels; crosses indicate low L1 and moderate L2 confounding; squares indicate moderate L1 and low L2 confounding.

and the same four school-level covariates: whether the school is public or private, an aggregate of the disciplinary climate in the school, an aggregate of student-teacher relations, and an aggregate of teacher support. However, the covariate set in our analysis differs from theirs in that we cluster-demeaned the student-level covariates and included the cluster means as school-level aggregates in our random-effects model.

The Singapore sample of PISA 2012 contains 5,546 students from 172 schools, with an average of 32.2 sampled students per school (standard deviation of 2.6). Upon listwise deletion of cases with missing data on at least one covariate, the sample reduces to 2,763 students, but the number of schools remains the same, averaging 16 students per school (standard deviation of 3.0). Incomplete covariate data could have been remedied using multiple imputation (e.g., Leyrat et al., 2019). Finally, with the removal of 49 schools that did not fulfill the within-school positivity condition (i.e., did not contain students from every category), the sample was further reduced to 2,024 students clustered in 123 schools, with an average of 16.5 students per school (min=9, max=22).[k] On average, schools in the reduced sample have 5.7 students from the baseline category 0 (min=1, max=11), 5.4 from category 1 (min=1, max=12), 1.8 from category 2 (min=1, max=6) and 3.6 from category 3 (min=1, max=12).

Table 2.7: Point estimates and standard errors for the effect of receiving personal tutoring, attending commercial lessons and both

Estimator	Personal		Commercial		Both	
	ATE	S.E.	ATE	S.E.	ATE	S.E.
Naïve	2.2	5.4	-30.9	7.4	17.6	6.9
Random Effects	14.0	4.5	-12.8	7.2	22.4	4.8
Fixed Effects	14.3	3.9	-9.3	4.8	17.7	4.0
Calibration	16.8	6.7	-11.1	9.5	20.7	8.3

Note. FE indicates the fixed-effects estimator, RE is random-effects, and CAL the calibration estimator

Table 2.7 presents the point estimates and the Balanced-Repeated-Replication standard errors (OECD, 2009) for the effects of interest. Each point estimate is an average of five estimates produced using the five plausible values for the mathematics score (PV1MATH-PV5MATH in the PISA dataset); in turn, those five estimates were computed through the procedures described in the sections above, using the final student weights (W_FSTUWT) as sampling probabilities. With the exception of the random-effects propensity scores, which were estimated

[k] As was mentioned before, rather than directly dropping all clusters that violate within-cluster positivity, a more principaled approach would be to estimate propensities with the "full" subsamples and then carry out influence diagnostics. For simplicity, here we proceed with the "reduced" subsample.

using the mixed-effects logit command of Stata (melogit), all computations were carried out in R. All scripts are available in the repository for this article.
For the effect of personal tutoring, the three estimators closely agree on a statistically significant effect, somewhere between 14 and 16.8 exam points. For the effect of commercial lessons, the three estimators again agree on the magnitude of the effect, but none of the estimates are significantly different from zero. Finally, for the combination of both private tutoring and commercial lessons, results are mixed: the calibration estimator places the effect about 4 points higher than the effect of private lessons on their own; the fixed-effects estimator only 3 points higher; but the random-effects estimator suggests the benefit of combining both extracurricular strategies could be as high as 8 points above the effect of private lessons on their own.

2.4 Concluding remarks

In this article, we investigated the behavior of three well-known, propensity-score estimators from the multilevel literature in a setting that, to the best of our knowledge, had not been studied before: weighting estimation of treatment effects at level one with multiple, unordered categories. Specifically, we took propensity scores estimated with procedures that take into account the hierarchical structure of the data, and inserted them into the weighting method developed by Imbens (2000) for single-level, multicategorical treatment effect estimation.
Through simulation studies, we found that these multilevel, multicategorical estimators largely follow some of the behaviors that have been observed in their counterparts from the multilevel, binary treatment literature (Arpino & Mealli, 2011; Li et al., 2013; Fuentes et al., 2021). The calibration procedure of Yang (2018) was able to reliably recover treatment effects in many scenarios, including conditions with small cluster sizes and strong confounding. Nevertheless, this estimator required moderate or large cluster sizes to produce unbiased estimates when treatment prevalences are low. The weighting estimator with fixed-effects propensities performed well under moderate to strong confounding and balanced prevalences, but struggled under low prevalences even in moderate cluster size conditions. Still, like the calibration estimator, it has the advantage that analysts need not have all level-two confounders available in their dataset. In contrast, the weighting estimator with random-effects propensities requires that all level-two confounders be observed and, even under this ideal scenario, it did not perform much better than the fixed-effects estimator.
Our simulations also yielded insights that are unique to the multicategorical

multilevel setting. First, as the number of categories grows, the prevalence of each category falls (holding cluster size fixed), making it difficult for these methods to produce unbiased estimates. This is particularly true for the calibration estimator, so that the fixed-effects and random-effects estimators may be the better choice when clusters are small, regardless of the number of clusters. Still, in rich data settings (i.e., many moderate-sized or large clusters), the calibration estimator should be preferred. Second, we note that applying the procedure of Imbens (2000) to clustered data presents a choice that does not emerge in single-level settings, namely, the choice of whether to include in estimation clusters that violate the sample positivity condition (i.e., clusters that do not contain units from every category). Our simulations show that, assuming the treatment assignment models hold in the population, some of those clusters may have fallen into violation of positivity due to extreme level-two covariate and/or extreme level-two residual values. Given this insight, we advice practitioners to inspect such clusters as potential high-leverage and/or outlying level-two units.

In practice, some of the assumptions and simplifications made in this article may not be appropriate. Our simulations assumed throughout that all level-one covariates were observed, whereas covariate selection and omitted variable bias at level one are major hurdles in the application of propensity score weighting for causal inference (Brookhart et al., 2006). In addition, our treatment assignment models only considered the possibility of random intercepts, although testing for and modelling random slopes is also an important concern in practice (Snijders & Bosker, 2012). For simplicity, we only studied the behavior of the most basic forms these estimators, but recent research has shown that these basic strategies can be improved upon through, for example, various trimming procedures (Yang & Ding, 2018; Fuentes et al., 2021). The bias of the random-effects estimator has also been reduced in the binary treatment literature through pooling of low-prevalence clusters (Lee et al., 2019), a strategy that may pay off in the multicategorical setting. Finally, other important extensions and related weighting methods were not studied here, such as the overlap weights of Li et al. (2018) which, for a homogeneous binary treatment, have been shown to perform as well as or better than the three traditional estimators we did investigate (Fuentes et al., 2021).

Appendix

Table 2.8: Simulation Study 1: Relative bias (%) for estimation of δ_2, as a function of the number of clusters, the cluster size and the strength of confounding at level 1 and level 2

	N = 30								N = 100							
	$n_j = 10$				$n_j = 30$				$n_j = 10$				$n_j = 30$			
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Random Effects	0.1	1.863	6.248	7.66	0.1	4.117	6.655	13	0.1	2.184	4.757	7.71	0.1	3.67	7.668	8.7
	0.2	1.513	6.032	10.3	0.2	4.744	8.789	13.2	0.2	2.889	4.774	7.62	0.2	6.503	7.864	9.62
	0.3	2.47	7.429	8.7	0.3	5.017	6.881	14.2	0.3	-1.03	3.179	6.44	0.3	3.046	9.39	10.1
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Fixed Effects	0.1	-0.71	8.589	7.65	0.1	2.919	6.371	11.5	0.1	3.115	5.196	4.15	0.1	3.266	7.553	7.16
	0.2	3.758	10.45	12	0.2	5.653	8.953	12.6	0.2	3.915	4.605	5.66	0.2	6.828	8.425	9.97
	0.3	4.949	12.21	14	0.3	9.441	11.36	18.1	0.3	4.387	6.069	6.49	0.3	8.109	13.06	14.1
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Calibration Estimator	0.1	-1.97	5.242	7.66	0.1	-0.55	-1.22	0.64	0.1	0.092	-1.41	2.8	0.1	-1.14	-0.28	-0.1
	0.2	2.679	5.352	7.28	0.2	0.015	-0.42	-1.5	0.2	0.394	0.015	5.3	0.2	0.268	-0.64	0.05
	0.3	-0.28	6.151	12.7	0.3	0.278	-2.37	2.2	0.3	-0.28	2.327	9.03	0.3	0.612	1.372	0.54

Note. R^2 L1 and R^2 L2 are the explained variances for the treatment equations, at level one and level two, respectively. N is the number of clusters and n_j the cluster size.

Table 2.9: Simulation Study 1: RMSE for estimation of δ_2, as a function of the number of clusters, the cluster size and the strength of confounding at level 1 and level 2

	N = 30								N = 100							
	$n_j = 10$				$n_j = 30$				$n_j = 10$				$n_j = 30$			
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Random Effects	0.1	0.169	0.199	0.22	0.1	0.099	0.117	0.14	0.1	0.096	0.104	0.13	0.1	0.057	0.073	0.08
	0.2	0.193	0.217	0.23	0.2	0.12	0.138	0.16	0.2	0.101	0.119	0.13	0.2	0.071	0.084	0.1
	0.3	0.208	0.245	0.27	0.3	0.159	0.167	0.2	0.3	0.124	0.148	0.16	0.3	0.1	0.115	0.13
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Fixed Effects	0.1	0.182	0.219	0.25	0.1	0.103	0.121	0.15	0.1	0.102	0.117	0.17	0.1	0.055	0.074	0.09
	0.2	0.194	0.232	0.26	0.2	0.115	0.144	0.16	0.2	0.105	0.132	0.16	0.2	0.069	0.085	0.1
	0.3	0.215	0.257	0.29	0.3	0.138	0.154	0.19	0.3	0.123	0.155	0.17	0.3	0.084	0.107	0.13
		R^2 L1				R^2 L1				R^2 L1				R^2 L1		
	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15	R^2 L2	0.05	0.1	0.15
Calibration Estimator	0.1	0.157	0.175	0.19	0.1	0.088	0.088	0.09	0.1	0.086	0.092	0.1	0.1	0.046	0.05	0.05
	0.2	0.176	0.185	0.2	0.2	0.092	0.1	0.1	0.2	0.089	0.097	0.11	0.2	0.049	0.054	0.05
	0.3	0.197	0.202	0.22	0.3	0.108	0.111	0.12	0.3	0.106	0.112	0.12	0.3	0.058	0.06	0.06

Note. R^2 L1 and R^2 L2 are the explained variances for the treatment equations, at level one and level two, respectively. N is the number of clusters and n_j the cluster size.

Chapter 3

Partial Pooling in Propensity Score Weighting with Clustered Data

Motivation

Propensity score (PS) strategies are a well-established tool for estimation of causal effects with observational data (Morgan & Winship, 2014; Schafer & Kang, 2008). An adequately estimated PS carries information on covariates that confound the relationship between the treatment and outcome of interest (Rosenbaum & Rubin, 1983). Once obtained, this summary can be used to match comparable units through PS matching algorithms (Stuart, 2010), to stratify the sample into comparable groups of units (Lunceford & Davidian, 2004), or to construct pseudopopulations where confounding is not present through weighting with the PS (Li et al., 2018; Leite et al., 2015). Methods that employ the latter strategy, PS weighting, are robust, simple to implement, computationally light relative to matching procedures, and easy to combine with the survey structure that is present in many observational datasets (Leite et al., 2015). In a context of clustered data, where unit-level treatment effects may be confounded by unit-level and cluster-level covariates, PS weighting has been shown to perform well with both traditional estimators of the propensity score (i.e., fixed effects and random effects; Arpino & Mealli, 2011); Fuentes et al., 2021) and with more recent algorithms that calibrate to confounder moments (Yang, 2018; Kim et al., 2017). Still, unbiased estimation is difficult when clusters are too small or too few, when covariate effects vary by cluster (i.e., when there are random slopes), and, particularly in the case of the calibration estimators, when units from some category are rare (Fuentes et al., 2021; Fuentes & Lüdtke, 2022). Thus, there is

room for improvement for all of these estimators under difficult conditions. Looking to improve upon the random-effects estimator, Lee et al. (2019) proposed "partial pooling", a method whereby clusters with similar prevalence are grouped together prior to estimation. For a binary treatment, they showed analytically and in simulations that a grouping based on cluster prevalences (i.e., the number of treated in each cluster) can reduce the bias of the random-effects estimator by reducing the confounding present at the cluster level. A related strategy is that of Kim et al. (2017), who suggested a grouping based on either known or latent similarities between clusters. Their latent-class approach to grouping clusters differs from the partitioning around medoids (PAM) applied by Lee et al. (2019), and only this latter strategy is the subject of our study.

The present article asks whether the partial pooling approach of Lee et al. (2019) can improve the performance of the fixed-effects PS weighting estimator and of the calibration estimator of Yang (2018). We build upon the simulation study of Fuentes & Lüdtke (2022), who applied fixed-effects PS weighting and calibration to the estimation of multicategorical treatment effects. In a multicategorical setting, treatment prevalences are low by construction at the cluster level, since multiple categories have to share the same cluster. This makes it difficult for the estimators to control for cluster-level confounding, particularly in settings of small and scarce clusters. Our simulations take the most difficult conditions from that study, and investigate whether pooling prior to estimation can remove any of the bias the estimators could not control on their own. Additionally, we investigate whether pooling can help the estimators control for confounding when random slopes are present.

The next section presents the theory of propensity score weighting, the methods employed for estimation of the propensity score, the partial pooling approach of Lee et al. (2019), and some additional notes on the application of these methods in a context of multiple treatments. Section 2 presents our simulation studies. Section 3 concludes.

3.1 Propensity score weighting theory

Table 3.1 illustrates the basic structure of a clustered dataset with a three-category treatment. The $N = 9$ units of this sample are grouped into $J = 3$ clusters, with i the index for units and j the index for clusters. Units have either been exposed to treatment 1 ($A_{ij} = 1$) or treatment 2 ($A_{ij} = 2$) or, say, no treatment at all ($A_{ij} = 0$). Interest lies in estimating the average effect of receiving treatments 1 and 2, rather than no treatment, on the outcome Y_{ij}:

$$ATE_1 = E\left[Y_{ij}(1) - Y_{ij}(0)\right] = \mu_1 - \mu_0 \tag{3.1}$$

$$ATE_2 = E\left[Y_{ij}(2) - Y_{ij}(0)\right] = \mu_2 - \mu_0 \tag{3.2}$$

Here, $Y_{ij}(a)$ are the so-called potential outcomes: the values of the outcome Y_{ij} that are or would have been observed if unit ij received or would have received treatment a (Imbens & Rubin, 2015). The expectations of these potential outcomes (i.e., μ_0, μ_1 and μ_2) can be estimated to compute estimates of the two average treatment effects, $\hat{ATE}_1$ and $\hat{ATE}_2$. Notice that only one potential outcome is observed for each unit, and the observed outcome Y_{ij} is:

$$Y_{ij} = Y_{ij}(0)A_{0,ij} + Y_{ij}(1)A_{1,ij} + Y_{ij}(2)A_{2,ij} \tag{3.3}$$

with $A_{0,ij}$, $A_{1,ij}$ and $A_{2,ij}$ the binary indicators constructed from the categorical treatment A_{ij}. Computation of the average potential outcomes is then hampered by the fact that we only observe the potential outcomes partially, as depicted by the shading in columns $Y(0)$, $Y(1)$ and $Y(2)$ of Table 3.1. When the shaded values are "missing at random" (Rosenbaum & Rubin, 1983), that is, when the potential outcomes and their missingness are independent given a set $\boldsymbol{X}_{ij}$ of characteristics of the units and a set $\boldsymbol{V}_j$ of characteristics of the clusters:

$$Y_{ij}(a) \perp A_{a,ij} | \boldsymbol{X}_{ij}, \boldsymbol{V}_j \quad for \quad a \in (0,1,2) \tag{3.4}$$

average potential outcomes can be recovered from the observed outcome Y_{ij}, and we say that the average treatment effects are weakly unconfounded (Imbens, 2000; Yang et al., 2016). Using weights constructed from propensity scores, the observed outcome values Y_{ij} can be weighted to make their distribution proportional to the distribution of each of the full potential outcome columns, such that averages of the weighted sample can be used to estimate population expectations (Li et al., 2018). The procedure, developed by Imbens (2000), is the following.

For our three-category example, a propensity score $p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)$ is the probability that unit i in cluster j received treatment $a \in (0,1,2)$, conditional on some characteristics $\boldsymbol{X}_{ij}$ of the units and some characteristics $\boldsymbol{V}_j$ of the clusters:

$$p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j) = p(A_{a,ij} = 1 | \boldsymbol{X}_{ij}, \boldsymbol{V}_j) \quad for \quad a \in (0,1,2) \tag{3.5}$$

The scores p_0, p_1 and p_2 are summaries of the confounding information of $\boldsymbol{X}_{ij}$ and $\boldsymbol{V}_j$ such that, under weak unconfoundedness, we may condition on $p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)$ rather than $\boldsymbol{X}_{ij}$ and $\boldsymbol{V}_j$ themselves to obtain expectations (i.e.,

Table 3.1: An artificial sample with three clusters of three units

j	i	A	A0	A1	A2	Y	Y(0)	Y(1)	Y(2)	X	V
1	1	2	0	0	1	515			515	2.5	1.5
1	2	0	1	0	0	541	541			3	1.5
1	3	1	0	1	0	510		510		2.1	1.5
2	1	0	1	0	0	537	537			3.5	0.7
2	2	0	1	0	0	545	545			2.7	0.7
2	3	1	0	1	0	512		512		2.3	0.7
3	1	2	0	0	1	530			530	2.8	1.2
3	2	0	1	0	0	520	520			2.8	1.2
3	3	0	1	0	0	526	526			3.2	1.2

unlike in Equation (3.4)):

$$\mu_a = E[E[Y_{ij}|A_{a,ij} = 1, p_{a,ij}]] \quad for \quad a \in (0,1,2) \tag{3.6}$$

The conditioning methods we study in this article involve weighting the observed outcomes with the inverse of the propensity score (Guo & Fraser, 2014; Rosenbaum & Rubin, 1983):

$$\begin{aligned} E\left[\frac{A_{a,ij}Y_{ij}}{p_{a,ij}}\right] &= E\left[\frac{A_{a,ij}Y_{ij}(a)}{p_{a,ij}}\right] = E\left[E\left[\frac{A_{a,ij}Y_{ij}(a)}{p_{a,ij}}|\boldsymbol{X}_{ij}\boldsymbol{V}_j\right]\right] \\ &= E\left[\frac{E[A_{a,ij}|\boldsymbol{X}_{ij}\boldsymbol{V}_j]E[Y_{ij}(a)|\boldsymbol{X}_{ij}\boldsymbol{V}_j]}{p_{a,ij}}\right] \\ &= E\left[\frac{p_{a,ij}E[Y_{ij}(a)|\boldsymbol{X}_{ij}\boldsymbol{V}_j]}{p_{a,ij}}\right] \\ &= E\left[E[Y_{ij}(a)|\boldsymbol{X}_{ij}\boldsymbol{V}_j]\right] = E[Y_{ij}(a)] = \mu_a \end{aligned} \tag{3.7}$$

The weighted sample averages we will use to estimate these expectations are of the form:

$$\hat{\mu}_a = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{a,ij}A_{a,ij}Y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{a,ij}A_{a,ij}} \tag{3.8}$$

where $\hat{\omega}_{a,ij} = 1/p_{a,ij}$ are the inverse-probability weights constructed from estimated propensity scores and the denominator of Equation (3.8) normalizes these weights to sum to one. The differences $\mu_1 - \mu_0$ and $\mu_2 - \mu_0$ are then estimates for the ATEs of Equations (3.2) and (3.3).

3.1.1 Estimation of the propensity score

To achieve unbiased estimation, the estimation methods we study here require that all relevant unit characteristics X_{ij} have been observed, but they allow for at least some of the cluster characteristics in V_j to be unobserved (Arpino & Mealli, 2011; Yang, 2018). In previous studies, their ability to capture this unobserved cluster-level information has been largely responsible for the differences in their performance (Fuentes et al., 2021).
Note that $p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)$ in Equation (3.5) is the regression curve of the category-a indicator $A_{a,ij}$ given the unit-level confounders $\boldsymbol{X}_{ij}$ and the cluster-level confounders $\boldsymbol{V}_j$. The logistic fixed-effects estimator (Hong & Raudenbush, 2006) assumes a generalized linear specification:

$$logit(p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)) = \gamma_{0,FE} + \boldsymbol{X}_{ij}\gamma_{X,FE} + \sum_{j=1}^{J-1} I_j\gamma_j \tag{3.9}$$

where an indicator variable I_j is included for all but one of the clusters (i.e., $J-1$), with $I_j = 1$ if unit i belongs to cluster j, and $I_j = 0$ otherwise. This complete set of indicators will be collinear with any cluster-level characteristics, and therefore no elements of V_j can be included in the regression. The analyst then relies completely on the indicator variables to represent all relevant cluster-level information, even if cluster-level characteristics are available in the dataset. In the literature, the indicator variables of the fixed-effects estimator have been shown to capture cluster intercepts effectively (Arpino & Mealli, 2013), although capturing cluster-specific slopes typically requires large clusters (Thoemmes & West, 2011; Fuentes et al., 2021).[a] Weights constructed from a fixed-effects propensity have preformed well in many scenarios, but can have unreasonably large values when units in the sample are assigned to some category with a very small probability; methods to overcome this instability have been shown to be effective in the single-level and clustered data contexts (Li et al., 2018; Fuentes et al., 2021). In a typical application, once a propensity score has been estimated with the fixed-effects method, the analyst would then check that the unit-level characteristics $\boldsymbol{X}_{ij}$ are indeed balanced within levels of the estimated propensity score. That is, within strata of the propensity score, the analyst will expect the elements of $\boldsymbol{X}_{ij}$ to have similar averages among treated and control units (i.e., similar according to some distance measure, such as Cohen's d). If that is not the case, estimation is attempted with a different specification of the propensity score (e.g., including interactions, squares, etc.)

[a] We discuss how this estimator can be modified to account for cluster-specific slopes in our presentation of Simulation 2 below.

and balance is checked again, repeating this process until the average differences of $\boldsymbol{X}_{ij}$ among treated and controls are not statistically significant.
This process of estimation, balance checks and respecification has been criticized in the literature (e.g., Imai et al., 2008) because of the difficulty of achieving balance for all covariates and the unreliability of balance statistics. Improving upon this strategy, calibration estimators have been developed that target exact balance (Hainmueller, 2012; Imai & Ratkovic, 2014; Kim et al., 2017; Yang, 2018). That is, once a specification has been chosen, the estimator ensures all $\boldsymbol{X}_{ij}$ averages are identical across treatment groups after conditioning. Among these estimators, the calibration procedure of Yang (2018) has been shown to be robust in many clustered-data scenarios (Fuentes et al., 2021). Given an initial vector of weights ω^*_{ij}, which may be fixed-effects propensity score weights or simply a constant vector, the procedure of Yang (2018) finds the vector $\hat{\omega}_{a,ij}$ that is closest in Kullback-Leibler distance:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{a,ij} ln \frac{\hat{\omega}_{a,ij}}{\omega^*_{ij}} \tag{3.10}$$

while fulfilling the two balance constraints:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{a,ij}A_{a,ij}\boldsymbol{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{\omega}_{a,ij}(1-A_{a,ij})\boldsymbol{X}_{ij} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\boldsymbol{X}_{ij} \tag{3.11}$$

$$\sum_{i=1}^{n_j}\hat{\omega}_{a,ij}A_{a,ij} = \sum_{i=1}^{n_j}\hat{\omega}_{a,ij}(1-A_{a,ij}) = n_j \quad for \quad j = 1, ..., J \tag{3.12}$$

The constraint of Equation (3.11) forces the averages of every unit-level covariate in $\boldsymbol{X}_{ij}$ to be the same across treatment groups and equal to the overall average, while the constraint of Equation (3.12) forces the weighted count of treated to be the same as the weighted count of controls within each cluster. In other words, once we weight the sample with the resulting $\hat{\omega}_{a,ij}$, the treated and control units are comparable in $\boldsymbol{X}_{ij}$, and every cluster has the same number of treated and controls. The weights that achieve this balance are:

$$\hat{\omega}_{a,ij} = n_j \frac{\omega^*_{ij} exp\left\{\boldsymbol{X}_{ij}\hat{\boldsymbol{\lambda}}\right\}}{\sum_{h=1}^{n_j} A_{a,hj}\omega^*_{hj} exp\left\{\boldsymbol{X}_{hj}\hat{\boldsymbol{\lambda}}\right\}} \tag{3.13}$$

where $\hat{\boldsymbol{\lambda}}$ are the Lagrange Multipliers from the constrained minimization of the Kullback-Leibler distance in Equation (3.10).
In addition to the advantage of obtaining perfect balance, results from previous simulation studies suggest this calibration strategy is more robust than the fixed-effects estimator and a similar calibration method proposed by Kim et al.,

2017 in many scenarios (Yang, 2018; Fuentes et al., 2021). An important exception are scenarios with rare treatments, that is, where cluster prevalences are low for some category (Fuentes et al., 2021; Fuentes & Lüdtke, 2022). Our simulation studies below investigate this weakness of the calibration estimator, as well as its performance in the presence of random slopes.

3.1.2 Partial pooling

As mentioned above, both the logistic fixed-effects propensity score weighting estimator and the calibration estimator of Yang (2018) can control for cluster-level confounders without observing them. In previous simulation studies, however, the degree to which these and other estimators are able to control for unobserved cluster-level confounding varies across simulation conditions. In particular, settings with small cluster sizes and unbalanced prevalences have proven difficult (Fuentes et al., 2021).

To help control for unobserved cluster-level confounding, Lee et al. (2019) suggested a strategy of grouping together similar clusters, estimating within those subsamples, then averaging the estimates. In their method, cluster similarity is defined by the treatment prevalence within clusters and, to group clusters according to these prevalences, they employed Partitioning Around Medoids (PAM; Rousseeuw & Kaufman, 2005), an algorithm that partitions the sample of J clusters into a predetermined number of groups G that minimize some within-group dissimilarity measure d. With P_j the prevalence of cluster j and P_g the average prevalence of the group to which cluster j is assigned, these authors showed that the grouping that minimizes $d = P_j - P_g$ (i.e., the within-group difference in prevalences) can reduce or eliminate the bias of cluster-level confounding, and that grouping by the values of observed confounders was less effective. They demonstrated this effect analytically for a binary treatment, and in simulations of the random-effects propensity score weighting estimator (Hong & Raudenbush, 2006) also with a binary treatment.

In the next section, we present simulations that investigate whether partial pooling can also improve the performance of the fixed-effects estimator and of the calibration estimator. Importantly, unlike the binary treatment study of Lee et al. (2019), we evaluate this strategy in a multicategorical context, where more than one prevalence is available to group by. Given three categories, our simulations attempt partitioning based on one prevalence and partitioning based on two prevalences. Our reasoning for investigating a partitioning based on more than one prevalence is the following. Recall that both estimators of μ_a (i.e., the fixed-effects and calibration estimators) are blind to observed cluster-level

confounders, and that they both use the within-cluster prevalence of A_a as their only source of information regarding the values of the observed cluster-level confounders, the unobserved cluster-level confounders, and the cluster-level residual. Because of this, we suspect these estimators may benefit from the information an additional vector of prevalences may carry regarding the values of those cluster-level elements.

3.1.3 Full and reduced samples

Before turning to the presentation of our simulation studies, we briefly address the practical issue of full and reduced samples. Consider again the mock sample of Table 3.1, where cluster 2 has no category-2 units and cluster 3 has no category-1 units. To estimate the category-1 propensity p_1 with, for example, the fixed-effects method, we need to regress the vector A_1 on X and on a set of cluster indicators. Before setting up this regression, we would be well advised to drop cluster 3 from the sample, since the absence of treated units in cluster 3 of the A_1 vector will surely result in an extreme value for that estimated cluster intercept and extreme weights for the units of the cluster. What is not immediately clear, however, is whether we should keep or drop cluster 2 in this regression, since it has both treated and control units in the A_1 vector, but is actually missing units from category-2. Similar issues are present for the estimation of p_2, where cluster 2 is clearly useless (since it has only zeroes in the A_2 vector) but it is unclear whether cluster 3 should be kept.
Fuentes & Lüdtke (2022) compared estimates computed with "reduced" samples (i.e., the subsample of only clusters that contain units from every category) and full samples. They found that estimating with reduced samples resulted in lower bias and variance across all their simulation conditions, and attributed this result to the quality of the clusters used in estimation, since the clusters where some category is absent tend to be outlying (i.e., they have very large or very small cluster intercepts) or high leverage (i.e., they have very large or very small values in their cluster-level covariates). The simulations of the next section also distinguish between estimates computed with full samples and reduced samples to test whether pooling together clusters with similar prevalences can help reduce the bias of estimation with full samples.

3.2 Simulation studies

We present two simulation studies. In the first, we simulate a random-intercepts scenario where confounding is strong, and where clusters are few and small, as

this combination proved difficult for multicategorical estimation in a previous study (Fuentes & Lüdtke, 2022). The aim here is to check whether pooling by prevalences can help the fixed-effects estimator and the calibration estimator overcome these adverse conditions by reducing the cluster-level confounding they have to deal with. The second simulation introduces a random slope, which has also been shown to make estimation difficult, even with a relatively large sample (Fuentes et al., 2021; Li et al., 2013). Again, the aim is to check whether pooling can improve the performance of the estimators in that case. Additionally, we take a closer look at the behavior of the calibration estimator of Yang (2018), specifically at how it handles clusters that have a prevalence of 1 for some category, and at whether there is a relationship between these "problematic" clusters and the bias we observe.

3.2.1 Simulation 1: Few and small clusters under strong confounding

We generate the three-category indicator A_{ij}, where each category $a \in (0,1,2)$ has a propensity score:

$$p_a(\boldsymbol{X}_{ij}, Z_j) = \frac{exp\left\{\alpha_{0a} + X_{1,ij}\alpha_{a,X1} + X_{2,ij}\alpha_{a,X2} + Z_j\alpha_{a,Z} + U_{A,j}\right\}}{\sum_{k=0}^{2} exp\left\{\alpha_{0k} + X_{1,ij}\alpha_{k,X1} + X_{2,ij}\alpha_{k,X2} + Z_j\alpha_{k,Z} + U_{A,j}\right\}} \quad (3.14)$$

Here, the unit-level confounders $X_{1,ij}$ and $X_{2,ij}$ are standard Normal; the cluster-level confounder Z_j is standard Normal; and the cluster-level residual $U_{A,j}$ is Normal, centered at zero and with a variance that implies a residual ICC of 0.2. The slopes of these treatment equations are set like in the most difficult condition from Simulation 1 of Fuentes & Lüdtke (2022), which achieve at the unit level ("level one", or L1) an R^2_{L1} of approximately 0.15, and at the cluster level ("level two", or L2) an R^2_{L2} of approximately 0.2. With $\alpha_X = 0.8701$ and $\alpha_Z = 1.036$, the slopes of $X_{1,ij}$ and $X_{2,ij}$ for the baseline category were both set to $-\alpha_X/2$, while for category 1 they were set to $\alpha_X/4$ and $3\alpha_X/4$, respectively, and for category 2 to $3\alpha_X/4$ and $\alpha_X/4$, respectively. The slope of Z_j was set to $-\alpha_Z$ for the baseline category, and to α_Z for categories 1 and 2. The intercepts α_{01} and α_{02} were set to zero, while the intercept of the baseline category, α_{00}, was set to -0.52 so that the average probability of the three categories is balanced at around 33%.

Given these three propensity scores, the three binary indicators $A_{0,ij}$, $A_{1,ij}$ and $A_{2,ij}$ are straightforward to generate with, for instance, the *sample* function of

the base package in R (code in the supplement). The observed outcome is then:

$$Y_{ij} = Y_{0,ij}A_{0,ij} + Y_{1,ij}A_{1,ij} + Y_{2,ij}A_{2,ij} \tag{3.15}$$

where the potential outcomes are:

$$Y_{0,ij} = X_{1,ij}\beta_{X1} + X_{2,ij}\beta_{X2} + Z_j\beta_Z + U_{Y,j} + \epsilon_{ij} \tag{3.16}$$

$$Y_{1,ij} = A_{1,ij}\delta_1 + X_{1,ij}\beta_{X1} + X_{2,ij}\beta_{X2} + Z_j\beta_Z + U_{Y,j} + \epsilon_{ij} \tag{3.17}$$

$$Y_{2,ij} = A_{2,ij}\delta_2 + X_{1,ij}\beta_{X1} + X_{2,ij}\beta_{X2} + Z_j\beta_Z + U_{Y,j} + \epsilon_{ij} \tag{3.18}$$

Here, the treatment effects to be estimated are $\delta_1 = 0.5$ and $\delta_2 = 0.3$; the slopes are $\beta_{X1} = \beta_{X2} = \beta_Z = 0.25$; ϵ_{ij} is a Normal residual at level one, with mean zero and variance 0.8; and $U_{Y,j}$ is Normal residual at level two, with mean zero and variance 0.2. Given this parametrization, the explained variations in the outcome equation are approximately $R^2_{L1} = 0.1$ and $R^2_{L2} = 0.05$; and the total explained variations of $A_{1,ij}$ and $A_{2,ij}$ in Equation (3.15) are approximately 0.045 and 0.016, respectively.

We simulated 1,000 samples of 30 clusters with 10 units per cluster, and estimated the two treatment effects, δ_1 and δ_2, with each sample.[b] Table 3.2 presents the bias and RMSE results for six different versions of the fixed-effects (FE) and calibration (CAL) estimators of δ_1. The versions differ by whether the "full" or "reduced" sample was used in estimation, by whether or not the clusters were pooled into two groups using PAM prior to estimation, and by whether PAM was applied using one or two prevalences. D=1 stands for "one dimension", and indicates PAM was carried out using only the prevalence of the category in question (e.g., the prevalence of category zero when estimating μ_0). D=2 indicates the prevalences of both category 0 and category 1 were used to pool clusters, regardless of which average potential outcome was being estimated. Partitioning around medoids was implemented with the *pam* function of the cluster package in R. The code for the fixed-effects regressions, which were carried out using the *glm* function in R, and the calibration estimation is available in the supplement. The main results are as follows.

First, as was observed in previous simulation studies, estimates computed with the full sample are severely biased. For the fixed-effects estimator, the results show this bias cannot be corrected through pooling. And in the case of the calibration estimator, pooling with either one or two dimensions substantially

[b] The results for δ_1 follow. See Table 3.6 in the Appendix for the results of δ_2

Table 3.2: Simulation study 1: Relative bias, standard deviation and RMSE of the estimators of δ_1 that use the full or reduced samples, with and without grouping clusters by one or two prevalences

Sample	PAM D	Estimator	Bias	SD	RMSE
Full	-	CAL	24	0.188	0.202
Full	-	FE	29	0.267	0.281
Full	1	CAL	38	0.19	0.222
Full	1	FE	28	0.258	0.272
Full	2	CAL	40	0.191	0.225
Full	2	FE	28	0.264	0.277
Reduced	-	CAL	9	0.218	0.22
Reduced	-	FE	9	0.292	0.293
Reduced	1	CAL	23	0.212	0.223
Reduced	1	FE	10	0.284	0.285
Reduced	2	CAL	25	0.206	0.219
Reduced	2	FE	10	0.292	0.294

Note. Reduced sample indicates that only clusters with units from every category were used in estimation. PAM D indicates the number of prevalences (i.e., dimensions) used when Partitioning Around Medoids. FE = fixed-effects estimator; CAL = calibration estimator.

worsens the bias. We explore this worsening in the subsection below with random, rather than fixed, cluster sizes.

Second, as would also be expected for this very adverse scenario of few and small clusters with strong confounding at both levels, even the reduced-sample fixed-effect and reduced-sample calibration estimators struggle to control for confounding. Interestingly, even using the reduced sample, pooling more than doubles the bias of the calibration estimator, while the fixed-effects estimator performs about the same with and without pooling.

Third, whether clusters were pooled using one or two dimensions made little difference. This result suggests that, for the estimation of an average potential outcome μ_a, the prevalences of the categories other than the category in question (i.e., category a) carry no useful information beyond what is present in the binary indicator A_a.

Random cluster sizes

We now take a closer look at the behavior of the calibration estimator, in an attempt to understand why pooling not only does not improve its performance, but instead induces a strong bias. We suspected this could be related to how the calibration estimator assigns weights when clusters have only one unit from some category. Recall from Equation (3.12) that the within-cluster sum of calibration weights must equal the cluster size, and suppose that among our sample there are $h > 1$ clusters of size 10. Suppose further that those h clusters each contain only one unit from the baseline category, that is, category 0. To fulfill the condition of Equation (3.12) in the estimation of μ_0, the calibration estimator must assign an identical weight of 10 to all of these singleton units, regardless of

how different they may be in terms of their unit-level covariate values. In other words, the weights of the calibration estimator are unable to represent individual differences in this situation. With this in mind, we suspected that pooling by prevalences aggravates this issue because, following the example, one of the groups created by pooling will contain all the problematic clusters (i.e., all h clusters with a prevalence of 1 for the baseline category will be pooled into the same group). We now ask whether these few problematic units (i.e., units that are assigned identical weights but are not identical in covariate values) are in fact behind the bias observed.

We simulated an additional 1,000 samples of 30 clusters each, only this time the size of the clusters was drawn from the discrete uniform distribution $U(8, 12)$, such that we have an average cluster size of 10, rather than a fixed cluster size of 10. In the simulation above with fixed cluster sizes, of the 1,000 samples generated, 97% had multiple clusters with a single treated and/or multiple clusters with a single control for the estimation of μ_0, 81% for the estimation of μ_1 and 80% for μ_2. For the 1,000 samples we now generated with random cluster sizes, these percentages are 77%, 50% and 50%, respectively, since choosing the cluster size at random reduces the probability of generating multiple clusters with both the same size and prevalence. Thus, any bias reduction we observe in the results of this simulation vs the simulation above can be attributed to an attenuation of the extreme prevalence problem of the calibration estimator (if there is, in fact, such a problem).

Table 3.3 presents the δ_1 results for the calibration estimator with no pooling, with one-dimensional pooling and with two-dimensional pooling.[c] In all cases, only a small bias reduction was observed (i.e., against the fixed cluster size results of Table 3.2), despite the large reduction in the number of problematic clusters. We interpret this to mean that, although the problematic clusters do affect the performance of the calibration estimator, they are not the main source of bias in this scenario of small clusters and strong confounding.

Table 3.3: Simulation study 1: Relative bias, standard deviation and RMSE of the calibration estimator of δ_1 for samples generated with random cluster sizes

Sample	PAM D	Estimator	Bias	SD	RMSE
Reduced	-	CAL	6	0.215	0.216
Reduced	1	CAL	22	0.202	0.212
Reduced	2	CAL	21	0.214	0.223

Note. Reduced sample indicates that only clusters with units from every category were used in estimation. PAM D indicates the number of prevalences (i.e., dimensions) used when Partitioning Around Medoids. CAL = calibration estimator.

[c] See Table 3.7 in the Appendix for the results of δ_2

3.2.2 Simulation 2: Random slope

When clusters are numerous and reasonably large, unbiased estimates may still be difficult to obtain if the effect of covariates differs from cluster to cluster, that is, if the data generating processes of the treatment and the outcome have random slopes (Thoemmes & West, 2011). This has been shown for both the fixed-effects and calibration estimators in a binary treatment setting (Fuentes et al., 2021). To determine whether pooling can help capture this slope variation and reduce bias, we added a random slope to the data generating processes of Simulation 1 as follows.
Let:

$$\begin{aligned} p_0^*(\boldsymbol{X}_{ij}, Z_j) = exp\{&\alpha_{00} + X_{1,ij}(\alpha_{0,X1} + \alpha_{X_1U_A}U_{A,j}) \\ &+ X_{2,ij}\alpha_{0,X2} + Z_j\alpha_{0,Z} + U_{A,j}\} \end{aligned} \tag{3.19}$$

$$p_1^*(\boldsymbol{X}_{ij}, Z_j) = exp\{\alpha_{01} + X_{1,ij}\alpha_{1,X1} + X_{2,ij}\alpha_{1,X2} + Z_j\alpha_{1,Z} + U_{A,j}\} \tag{3.20}$$

$$p_2^*(\boldsymbol{X}_{ij}, Z_j) = exp\{\alpha_{02} + X_{1,ij}\alpha_{2,X1} + X_{2,ij}\alpha_{2,X2} + Z_j\alpha_{2,Z} + U_{A,j}\} \tag{3.21}$$

Here, the parametrization is identical to Equation (3.14), except for the random slope $\alpha_{X_1U_A}$ which was set to $-\alpha_X/4$ (see the specification of Simulation 1). The propensity score of the binary indicators for each category $a \in (0, 1, 2)$ was set to:

$$p_a(\boldsymbol{X}_{ij}, Z_j) = \frac{p_a^*}{\sum_{k=0}^{2} p_k^*} \tag{3.22}$$

The observed outcome is as in Equation (3.15), and the potential outcomes for category 1 and 2 are as in Equations (3.17) and (18). But the potential outcome for the baseline category 0 is now:

$$Y_{0,ij} = X_{1,ij}(\beta_{X1} + \beta_{X1U_Y}U_{Y,j}) + X_{2,ij}\beta_{X2} + Z_j\beta_Z + U_{Y,j} + \epsilon_{ij} \tag{3.23}$$

where the parametrization is identical to Equation (3.16), except for the random slope β_{X1U_Y} which was set to $\beta_{X1}/2$.
To estimate under the presence of a random slope, the fixed-effects specification of Equation (3.9) must be augmented with interactions of $X_{1,ij}$ with the cluster indicators:

$$logit(p_a(\boldsymbol{X}_{ij}, \boldsymbol{V}_j)) = \gamma_{0,FE} + X_{1,ij}(\gamma_{X1,FE} + \sum_{c=1}^{J-1} I_c\gamma_{cX}) + X_{2,ij}\gamma_{X2,FE} + \sum_{j=1}^{J-1} I_j\gamma_j \tag{3.24}$$

The slope of $X_{1,ij}$ for cluster j is then $\gamma_{X1,FE} + \gamma_{jX}$. Similarly, to estimate with the calibration estimator of Yang (2018), the balance condition of Equation (3.11) for $X_{1,ij}$ is replaced with a stricter one that requires balance within clusters:

$$\sum_{i=1}^{n_j} \hat{\omega}_{a,ij} A_{a,ij} X_{1,ij} = \sum_{i=1}^{n_j} \hat{\omega}_{a,ij}(1 - A_{a,ij}) X_{1,ij} = \sum_{i=1}^{n_j} X_{1,ij} \quad for \quad j = 1, ..., J \tag{3.25}$$

We simulated 1,000 samples of 50 clusters with 30 units per cluster, and estimated the two treatment effects, δ_1 and δ_2, with each sample.[d] Though it is customary to simulate and study more than one condition for the number of clusters and their size, this setup is sufficient to determine whether pooling can improve the performance of these estimators when a random slope is present. This is because, if the clusters were larger or more numerous, the estimators of interest would achieve unbiased estimation on their own (i.e., without pooling) under our parametrization; and, with smaller or fewer clusters, the estimates would be severely biased under our parametrization and pooling would not have any chance of improving them sufficiently. Table 3.4 presents the relative bias, standard deviation and RMSE results for the six different versions of the fixed-effects (FE) and calibration (CAL) estimators of δ_1. The main results are as follows.

First, as expected, the estimators that use the full sample are severely biased. Pooling appears to reduce this bias, but not enough to render the estimators useful.

Second, among the reduced-sample estimates, the fixed-effects estimator without pooling achieved the lowest bias, though it is outperformed by the calibration estimator without pooling in terms of RMSE.

Third, when using reduced samples, pooling had no effect on the calibration estimator, and had a small biasing effect on the fixed-effects estimator.

Lastly, as observed in Simulation 1, whether PAM is applied using one or two prevalences seems to make a negligible difference for all of these estimators.

Random cluster sizes

[d] The results for δ_1 follow. See Table 3.8 in the Appendix for the results of δ_2.

Table 3.4: Simulation study 2: Relative bias, standard deviation and RMSE of the estimators of δ_1 that use the full or reduced samples, with and without grouping clusters by one or two prevalences

Sample	PAM D	Estimator	Bias	SD	RMSE
Full	-	CAL	38	0.132	0.175
Full	-	FE	27	0.117	0.141
Full	1	CAL	31	0.11	0.144
Full	1	FE	23	0.123	0.141
Full	2	CAL	31	0.11	0.143
Full	2	FE	22	0.123	0.14
Reduced	-	CAL	18	0.102	0.116
Reduced	-	FE	9	0.139	0.142
Reduced	1	CAL	19	0.105	0.12
Reduced	1	FE	14	0.123	0.131
Reduced	2	CAL	19	0.106	0.12
Reduced	2	FE	14	0.124	0.13

Note. Reduced sample indicates that only clusters with units from every category were used in estimation. PAM D indicates the number of prevalences (i.e., dimensions) used when Partitioning Around Medoids. FE = fixed-effects estimator; CAL = calibration estimator.

With parametrization identical to the random slope simulation above, we again simulated an additional 1,000 samples where the cluster size is randomly drawn from $U(8, 12)$. Whereas, with fixed cluster sizes, 90% of the simulated samples had multiple clusters with a single treated and/or multiple clusters with a single control for the estimation of μ_0, that proportion is only 64% when simulating with random cluster sizes. For the estimation of μ_1 and μ_2, the difference is 29% with fixed cluster sizes vs 1% with random cluster sizes.
Table 3.5 presents the results of estimating δ_1 with three variants of the reduced-sample calibration estimator: with no pooling, with single-prevalence pooling and with two-prevalence pooling.[e] The fact that the results are nearly identical to those obtained with fixed cluster sizes (see table Table 3.4) indicates again that problematic clusters are not behind any substantial part of the bias in this setting.

Table 3.5: Simulation study 2: Relative bias, standard deviation and RMSE of the calibration estimators of δ_1 for samples generated with random cluster sizes

Sample	PAM D	Estimator	Bias	SD	RMSE
Reduced	-	CAL	18	0.105	0.118
Reduced	1	CAL	19	0.109	0.123
Reduced	2	CAL	19	0.109	0.123

Note. Reduced sample indicates that only clusters with units from every category were used in estimation. PAM D indicates the number of prevalences (i.e., dimensions) used when Partitioning Around Medoids. CAL = calibration estimator.

[e]See Table 3.9 in the Appendix for the results of δ_2

3.3 Concluding remarks

In this article, we investigated the possibility of improving two estimators from the multilevel propensity score weighting literature —the fixed-effects inverse probability weighting estimator (Hong & Raudenbush, 2006) and the calibration estimator of Yang (2018)— through a procedure that had previously been shown to improve the random-effects estimator. Although the two estimators had performed relatively well in previous studies, we wondered whether bias under difficult simulated conditions could be decreased further through this procedure, namely, partial pooling.

In our simulation with a random intercept, multiple treatments, and small, scarce clusters, the fixed-effects approach did not benefit from pooling clusters of similar prevalence. In the best of cases, pooling prior to fixed-effects estimation had no effect, while in others it even introduced a slight bias. Partial pooling also did not reduce the bias of the calibration estimator of Yang (2018) in that setting, but instead greatly increased the bias. This was perhaps to be expected, since the calibration estimator is known to struggle under unbalanced prevalences, and the partial pooling procedure creates precisely those conditions: it creates groups where prevalences are necessarily less balanced than in the full sample. We also tested whether some of the bias of the calibration estimator is due to how it handles clusters with a prevalence of one for some category, and found that only a very small portion of the bias could be due to this effect. The question of why the calibration estimator struggles so much under unbalanced prevalences remains open.

Our simulations also found that partial pooling cannot help these estimators control for a random slope. When the effect of a covariate varies strongly from cluster to cluster, it makes it difficult to pinpoint a treatment effect, and large clusters are typically required. We had hoped that, by reducing this variance prior to estimation, pooling would make it easier for the estimators to do their job. Unfortunately, that was not the case.

Finally, our simulations also tested whether pooling could reduce the bias of using a "full" sample in the estimation of a multicategorical treatment effect, that is, using clusters that are missing units from some category. The answer again was no. The bias induced by the outlying and high leverage clusters cannot be removed by grouping those clusters with similar ones prior to estimation.

Note that, although our investigations were carried out in a context of multiple treatments, the findings are directly applicable to the binary treatment case, since the estimation algorithm is identical. The only important difference is that, with a binary treatment, the within-cluster prevalences are typically more

balanced, save in cases of a rare treatment or rare outcome. This balance tends to allow for less biased estimation. Note also that, throughout, our simulations assumed that all relevant unit-level covariates were observed, such that unobserved confounders were only present at the cluster level. In practice, however, covariate choice at both levels is an important part of the specification process (Brookhart et al., 2006). Furthermore, analysts may choose to specify a model for the outcome, in what has been called the "doubly-robust" approach (Hernán & Robins, 2020), but our investigations focus on treatment models.

Appendix

Table 3.6: Simulation study 1: Relative bias, standard deviation and RMSE of the estimators of δ_2 that use the full or reduced samples, with and without grouping clusters by one or two prevalences

Sample	PAM D	Estimator	Bias	SD	RMSE
Full	-	CAL	42	0.19	0.228
Full	-	FE	48	0.273	0.309
Full	1	CAL	65	0.194	0.274
Full	1	FE	48	0.263	0.3
Full	2	CAL	65	0.191	0.272
Full	2	FE	48	0.264	0.301
Reduced	-	CAL	16	0.215	0.22
Reduced	-	FE	15	0.286	0.29
Reduced	1	CAL	39	0.207	0.238
Reduced	1	FE	19	0.27	0.276
Reduced	2	CAL	38	0.206	0.235
Reduced	2	FE	17	0.278	0.283

Note. Reduced sample indicates that only clusters with units from every category were used in estimation. PAM D indicates the number of prevalences (i.e., dimensions) used when Partitioning Around Medoids.FE = fixed-effects estimator; CAL = calibration estimator.

Table 3.7: Simulation study 1: Relative bias, standard deviation and RMSE of the calibration estimator of δ_2 for samples generated with random cluster sizes

Sample	PAM D	Estimator	Bias	SD	RMSE
Reduced	-	CAL	10	0.21	0.212
Reduced	1	CAL	36	0.212	0.239
Reduced	2	CAL	33	0.213	0.235

Note. Reduced sample indicates that only clusters with units from every category were used in estimation. PAM D indicates the number of prevalences (i.e., dimensions) used when Partitioning Around Medoids. CAL = calibration estimator.

Table 3.8: Simulation study 2: Relative bias, standard deviation and RMSE of the estimators of δ_2 that use the full or reduced samples, with and without grouping clusters by one or two prevalences

Sample	PAM D	Estimator	Bias	SD	RMSE
Full	-	CAL	121	0.129	0.385
Full	-	FE	45	0.114	0.177
Full	1	CAL	51	0.11	0.189
Full	1	FE	40	0.116	0.167
Full	2	CAL	51	0.11	0.188
Full	2	FE	39	0.117	0.166
Reduced	-	CAL	31	0.107	0.142
Reduced	-	FE	15	0.139	0.146
Reduced	1	CAL	31	0.108	0.143
Reduced	1	FE	26	0.118	0.141
Reduced	2	CAL	31	0.108	0.142
Reduced	2	FE	25	0.118	0.14

Note. Reduced sample indicates that only clusters with units from every category were used in estimation. PAM D indicates the number of prevalences (i.e., dimensions) used when Partitioning Around Medoids. FE = fixed-effects estimator; CAL = calibration estimator.

Table 3.9: Simulation study 2: Relative bias, standard deviation and RMSE of the calibration estimators of δ_2 for samples generated with random cluster sizes

Sample	PAM D	Estimator	Bias	SD	RMSE
Reduced	-	CAL	31	0.111	0.144
Reduced	1	CAL	31	0.114	0.146
Reduced	2	CAL	30	0.114	0.146

Note. Reduced sample indicates that only clusters with units from every category were used in estimation. PAM D indicates the number of prevalences (i.e., dimensions) used when Partitioning Around Medoids. CAL = calibration estimator.

Bibliography

Allison, P. D. (2009). *Quantitative applications in the social sciences: Fixed effects regression models*. SAGE Publications.
https://doi.org/10.4135/9781412993869
Anguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16(2)*, 270-301. DOI 10.1177/1094428112470848
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 the 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
Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163(12)*, 1149-1156. DOI: 10.1093/aje/kwj149
Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics, 155*, 138-154. DOI: 10.1016/j.jeconom.2009.09.023
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
Dingena Spreeuwenberg, M., Bartak, A., Croon, M. A., Hagenaars, A., Busschbach, J. J. V., Andrea, H., Twisk, J., & Stijnen, T. (2010), The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. *Medical Care, 48(2)*, 166-174.
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
Fuentes, A. (2022), Partial Pooling in Propensity Score Weighting with Clustered Data. [Unpublished manuscript]
Fuentes, A., & Lüdtke, O. (2022), Multiple treatment effect estimation with propensity score weighting for two-level data. [Unpublished manuscript]
Fuentes, A., Lüdtke, O., & Robitzsch, A. (2021). Causal inference with multilevel data: a comparison of different propensity score weighting approaches. *Multivariate Behavioral Research*. DOI: 10.1080/00273171.2021.1925521
Gibbs, B. R., Lytle, R., & Wakefield, W. (2019). Outcome effects of recividism among drug court participants. *Criminal Justice and Behavior, 46(1)*, 115-135. DOI: 10.1177/0093854818800528

Gupta, N. D., & Simonsen, M. (2010). Non-cognitive child outcomes and universal high quality child care. *Journal of Public Economics, 94(1-2)*, 30-43. DOI: 10.1016/j.jpubeco.2009.10.001

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

Hu, L., Ji, J., Ennis, R. D., & Hogan, J. W. (2022). A flexible approach for causal inference with multiple treatments and clustered survival outcomes. *arXiv preprint arXiv*:2202.08318

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. (2000). *The role of propensity score in estimating dose-response functions*. Technical Working Paper 237, National Bureau of Economic Research. Cambridge, MA. Retrieved from http://www.nber.org/papers/T0237

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

Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society, 161(2)*, 121-160.

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., & Li, F. (2018). Propensity score weighting for causal inference with multi-valued treatments. *arXiv preprint arXiv*:1808.05339v3

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

Liou, PY., & Hung, YC. (2014). Statistical techniques utilized in analyzing PISA and TIMSS data in science education from 1996 to 2013: a methodological review. *International Journal of Science and Mathematics Education, 13*, 1449–1468. DOI 10.1007/s10763-014-9558-5

Longford, N. T. (2001). Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society, 162(2)*, 259-273. DOI: 10.1111/1467-985X.00201

Lopez, M. J., & Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science, 32*, 432–454. DOI 10.1214/17-STS612

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., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2012). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine, 32*, 3388–3414. DOI: 10.1002/sim.5753
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
Rose, S., & Normand, S. (2019). Double robust estimation for multiple unordered treatments and clustered observations: evaluating drug-eluting coronary artery stents. *Biometrics, 75*, 289-296. DOI 10.1111/biom.12927
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
Rousseeuw, P. J., & Kaufman, L. (2005). *Finding groups in data: an introduction to cluster analysis*. United States. Wiley.
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 preprint arXiv*: 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
Yang, S., & Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika, 105(2)*, 487-493. DOI: 10.1093/biomet/asy008
Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., & Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics, 72*, 1055–1065. DOI 10.1111/biom.12505
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

Coauthor contributions

Chapter 1. Causal Inference with Multilevel Data: A Comparison of Different Propensity Score Weighting Approaches

Alexander Robitzsch

- Located a gap in the literature and conceived the article
- Co-designed and coded the simulation studies
- Derived mathematical proofs
- Wrote the section on survey weights

Oliver Lüdtke

- Located a gap in the literature and conceived the article
- Co-designed the simulation studies
- Answered reviewer questions and implemented reviewer advice

Alvaro Fuentes

- Assisted in the development of ideas
- Wrote and coded the empirical example
- Co-designed the simulation studies
- Drafted the manuscript

Chapter 2. Multiple Treatment Effect Estimation with Propensity Score Weighting for Two-Level Data

Alvaro Fuentes

- Located a gap in the literature and conceived the article
- Designed and coded the simulation studies
- Drafted the manuscript

Oliver Lüdtke

- Assisted in the development of ideas
- Advised in the design of the simulation studies

Oliver Lüdtke	Date
Alexander Robitzsch	Date

Erklärung zum selbständigen Verfassen der Arbeit

Ich erkläre hiermit, dass ich meine Doktorarbeit „Propensity Score Weighting Procedures for Causal Inference with Clustered Data“ selbstständig und ohne fremde Hilfe angefertigt habe und dass ich als Koautor maßgeblich zu den weiteren Fachartikeln beigetragen habe. Alle von anderen Autoren wörtlich übernommenen Stellen, wie auch die sich an die Gedanken anderer Autoren eng anlehnenden Ausführungen der aufgeführten Beiträge wurden besonders gekennzeichnet und die Quellen nach den mir angegebenen Richtlinien zitiert.

Alvaro Fuentes	Datum