Title

Accounting for Omitted Variable Bias in Hierarchical Linear Models with Group-Varying Treatment Assignment Processes

Author

Jacobson, Thomas Abram

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Accounting for Omitted Variable Bias in Hierarchical Linear Models with Group-Varying Treatment Assignment Processes

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science
in Statistics

by

Thomas Abram Jacobson

2023

ABSTRACT OF THE THESIS

Accounting for Omitted Variable Bias
in Hierarchical Linear Models with
Group-Varying Treatment Assignment Processes

by

Thomas Abram Jacobson
Master of Science in Statistics
University of California, Los Angeles, 2023
Professor Chad J. Hazlett, Chair

In multisite observational studies where level-one units are nested within level-two groups and treatment assignment occurs within (as opposed to between) groups, treatment assignment processes may vary between groups. A possible consequence of such group-varying treatment assignment processes is that the conditional exchangeability assumption may hold for some groups in a given sample while other groups are susceptible to omitted variable bias. This paper employs a simulation study to explore the potential for leveraging information about group-specific treatment assignment processes in order to mitigate omitted variable bias in hierarchical linear models with random intercepts and treatment effect slopes. The simulation demonstrates that an "augmented" model that incorporates information about group-varying treatment assignment can substantially reduce bias in treatment main effect estimates and also reduce the mean squared error of treatment random effect variance estimates.

The thesis of Thomas Abram Jacobson is approved.

Mark Stephen Handcock

Michael H. Seltzer

Chad J. Hazlett, Committee Chair

University of California, Los Angeles

2023

Contents

List of Tables

List of Figures

Acknowledgements

My deepest thanks go to my committee chair, Chad Hazlett, for his indispensable guidance in formulating and refining this work. I am also profoundly grateful to my other committee members, Michael Seltzer and Mark Handcock, for their generosity, wisdom, and patience over the years.

Minjeong Jeon and my classmates in her research apprenticeship course provided kind and helpful feedback at various stages. Special thanks to Marilyn Gray and the graduate consultants at the UCLA Graduate Writing Center, and especially to my friend and colleague Alex Kwako, for buoying me throughout the writing process.

Heartfelt, affectionate thanks to Ash Marshall for her steadfast support and encouragement, always.

1 Introduction

An important concern for social scientists, policymakers, and practitioners across various fields is to understand the causal effects of social policy interventions (or "treatments") in settings where individuals are nested within groups. Hierarchical linear models (*e.g.*, Raudenbush and Bryk (2002), Gelman and Hill (2007), and others) are useful tools in these contexts because they enable investigators to estimate an overarching fixed or "main" effect of the treatment while also modeling group-level variation in treatment effects.

Although randomized controlled trials (RCTs) are held up as the "gold standard" for evaluating causal relationships in education (US Department of Education; Institute of Education Sciences, 2003) and other fields, RCTs often are not a realistic option for studying the effects of social policies. This may be due to ethical concerns, practical infeasability, or other obstacles that interfere with randomly assigning participants to different treatment conditions. In such cases non-experimental observational data may be the only available empirical evidence for investigating important causal questions in community health, education, public safety, and other policy arenas.

A pervasive problem with using observational data for causal inference in social research, however, is selection bias. Naive comparisons of non-randomized comparison groups generally do not lead to valid causal effect estimates because individuals' likelihood of being exposed to different treatment conditions can depend on "confounding variables," *i.e*, background characteristics that also predict the outcomes of interest. While some of these confounding variables (or, at least, reasonable proxies for them) may be directly observed, investigators working with observational data must always be mindful of how their causal estimates might be biased by any unobserved confounders that their models necessarily omit.

Methodologists in recent decades have developed an array of so-called "quasi-experimental" strategies that can potentially identify unbiased causal effect estimates from observational data in certain contexts. In every case, however, the robustness of the causal inferences that we might hope to make from a quasi-experimental research design will hinge on satisfying some specific set of relevant assumptions. One such approach, which underpins widely-used strategies such as propensity score matching (Rosenbaum & Rubin, 1983b), is sometimes referred to as the conditional exchangeability assumption. Conditional exchangeability implies that—given some set of observed pre-treatment data—an individual's actual treatment status is independent of the potential outcomes that would be associated with the respective treatment conditions they could have experienced.

When dealing with nested observational data where any individual in any group could potentially experience any of the treatment conditions (*i.e.*, where the treatment is not simply assigned at the group level), the process for assigning treatment status to individuals may vary from one group to another (see Rickles (2011),

for example). Investigators who possess qualitative knowledge about how the treatment assignment process varies between groups (specifically with regard to how the association between observed and unobserved variables to treatment status varies among different groups) might then potentially use that information as leverage in obtaining an unbiased estimate of the main effect of the treatment, as well as in making inferences about the magnitude of the omitted variable bias that might be present in groups where unobserved factors are understood to influence treatment assignment.

To date, the potential for exploiting this kind of information about group-varying treatment assignment processes for estimating causal effects has received little attention in the literature around causal inference and hierarchical linear models. Kim and Seltzer (2007) considered different approaches for estimating propensity scores and treatment effects in multi-site settings where the factors influencing treatment selection vary between sites, but did not address sensitivity analysis to unobserved confounders writ large or the possibility that the magnitude of confounding may vary by site. Seltzer, Kim, and Frank (2006) usefully extended the approach outlined in Frank (2000) to multi-site evaluations where treatment assignment occurs at the group level and unobserved group-level confounders may bias the effect estimates. Less attention, however, has centered on how selection into treatment may vary between groups when treatment assignment occurs within the group level or the implications of heterogeneous selection for estimating the main effect of a treatment, the effect of a group-level moderator variable on group-level treatment effects, or group-level treatment effect variance.

This thesis uses simulated data to explore a possible framework for accounting for omitted variable bias in nested data structures where the treatment assignment process varies between groups such that the presence or absence of omitted variable bias may also vary from one group to another. If we assume a set of nested observational data where the conditional exchangeability assumption is tenable for a subset of groups but selection bias may inhere in other groups, how well can information about group-level variation in the treatment assignment process aid us in obtaining unbiased estimates of key parameters of interest?

The next section reviews the potential outcomes framework for causal inference and outlines necessary relevant assumptions for the particular data structures under consideration. Section 3 describes a specific hypothetical data generating process and specifies some hierarchical linear models that actual investigators might use in this context. Section 4 outlines the simulation study procedure for data generation and analysis, and Section 5 reviews the results of the simulation study. Finally, Section 6 sketches out some related methodological concerns that merit further attention as well as possible implications of this work for applied researchers.

2 Core Assumptions for Causal Inference in this Context

2.1 Potential Outcomes Framework

This discussion takes as a starting point the so-called "potential outcomes" model for causal inference, outlined in Rubin (1974), Holland (1986), Hernán and Robins (2020), and elsewhere. A simple case of the potential outcomes framework supposes a population of units $i = 1, \ldots, N$ that could potentially be exposed to some binary treatment condition (or "cause") $D_i \in \{0, 1\}$, and that exhibit some measurable post-treatment outcome Y_i.

The causal effect τ_i, then, of the treatment D_i on the outcome Y_i for unit i would be the difference between the value that Y_i would take if unit i had been exposed to the treatment condition (which we write as $Y_{1i} \equiv Y_i \mid D_i = 1$) and the counterfactual value Y_i would take had unit i instead been exposed to the control condition ($Y_{0i} \equiv Y_i \mid D_i = 0$), *i.e.*,

$$\tau_i = Y_{1i} - Y_{0i}. \tag{1}$$

The "Fundamental Problem of Causal Inference" (Holland, 1986) is that only one of the potential outcomes in $\{Y_{0i}, Y_{1i}\}$ can possibly be observed. Its counterfactual is always missing and can only, at best, be inferred.

The common strategy for estimating an average treatment effect given that only one potential outcome is observed for each unit i is to randomize treatment assignment so that unit i's potential outcomes are independent of its treatment status, *i.e.*,

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i. \tag{2}$$

Thus, the unobserved potential outcomes are missing completely at random and we can assume that treatment assignment is *exchangeable*.

An estimand of primary interest will be the average treatment effect τ_{ATE}, which we define as

$$\begin{aligned} \tau_{\text{ATE}} \equiv \mathbb{E}[\tau_i] &= \mathbb{E}[Y_{1i} - Y_{0i}] \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(Y_{1i} - Y_{0i} \right). \end{aligned} \tag{3}$$

We cannot compute τ_i or τ_{ATE} directly from the data since only one of $\{Y_{0i}, Y_{1i}\}$ is observed for each unit i. But if the randomization process is valid, the exchangeability assumption in Equation (2) is assured.

If we can moreover assume *consistency*—that the observed Y_i is the outcome that unit i would exhibit given the specific treatment condition $D_i \in \{0, 1\}$ it experienced—then we can write

$$Y_i = Y_{di} \mid D_i = d; \tag{4}$$

equivalently,

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}. \tag{5}$$

And if we can also rely on the *stable unit treatment value assumption (SUTVA)*, implying that the potential outcomes for unit i do not depend on the treatment assignment mechanism or on the treatment status of any other unit i' in the data (Rubin, 1986), then we can obtain an estimate of the average treatment effect $\hat{\tau}_{\text{ATE}}$ from the observed data:

$$\begin{aligned}
\hat{\tau}_{\text{ATE}} = \mathbb{E}[Y_{1i} - Y_{0i}] &= \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0] && \text{(by exchangeability)} \\
&= \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i =\mid D_i = 0] && \text{(by consistency)} \\
&= \frac{1}{N_1} \sum_{i=1}^{N_1} (Y_i) - \frac{1}{N_0} \sum_{i=1}^{N_0} (Y_i)\,,
\end{aligned} \tag{6}$$

where N_0 and N_1 represent a partition of the data into control ($D_i = 0$) and treatment ($D_i = 1$) units, respectively.

2.2 Additional Relevant Assumptions

Some other assumptions that will be important for this discussion include the following:

Positivity. Individual units could potentially be exposed to treatment or control conditions (Hernán & Robins, 2020), *i.e.*,

$$0 < \Pr(D_i = d) < 1. \tag{7}$$

This is important because it enables us to contrast the outcomes associated with different treatment levels. Otherwise it would be impossible to quantify an average causal effect from the data.

Conditional exchangeability. This is a weaker form of the exchangeability assumption in Equation (2)

entailing that unit i's potential outcomes are independent of treatment assignment given the observed unit-level pre-treatment covariates X_i (Dawid, 1979; Rosenbaum & Rubin, 1983a):

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i \mid X_i = x \text{ for any } x \in \mathcal{X}. \tag{8}$$

Note that the conditional exchangeability assumption can be falsified, for example, by obtaining a nonzero estimate of the treatment's effect on an outcome that could not have been affected by the treatment, such as a lagged measure of the outcome that is taken before units receive the treatment. Alternatively, if the untreated group can be partitioned into two or more groups, it may be possible to estimate the effect of a "pseudo-treatment" that distinguishes the different control groups and is known *a priori* not to affect the original outcome (Imbens, 2004; Imbens & Rubin, 2015). But since conditional exchangeability cannot dispositively be proven true, the plausibility of this assumption hinges on investigators' domain knowledge about whether and how strongly any unobserved set of confounders could conceivably influence treatment assignment and the outcome of interest. Focusing on the specific context of nested observational data with group-varying treatment selection processes, we will also assume the following for the purposes of this discussion:

Within-group positivity. There is a positive probability of assignment to each treatment level for any unit i in each group j,

$$0 < \Pr(D_{ij} = d) < 1, \tag{9}$$

such that it is possible to estimate a group-specific average treatment effect $\hat{\tau}_j$ for each group j. (In other words, individual units within groups get assigned to different treatment levels rather than treatment being assigned at the group level.)

Varying degrees of unobserved confounding. Investigators' contextual knowledge includes qualitative information about group-level variation in the treatment assignment process. Specifically, although selection into treatment is not random, some unobserved covariate z_{ij} potentially influences treatment assignment for units that belong to some groups but it *is not* associated with treatment assignment within other known groups.

3 Data Structure and Model Framework

3.1 Hypothetical Data Structure and Data Generating Process

Data structure

Consider a relatively simple data structure that investigators might plausibly encounter in an observational study of some multi-group educational or other social policy where

- there is a consistent treatment-control contrast of interest;
- a goal of the study is to understand or account for between-group variation in treatment effects (including random variation or nonrandom variation explained by some group-level moderator);
- the investigators possess qualitative knowledge about how the treatment assignment process varies between groups, and specifically can distinguish between a set of groups where the conditional exchangeability assumption in Equation (8) is tenable on the basis of observed data and another set of groups where unobserved variables may influence treatment assignment.

Given the preceding assumptions, suppose we have data comprising $i = 1, \ldots, N$ level-one units (*e.g.*, individual students) nested within one of $j = 1, \ldots, J$ level-two groups (*e.g.*, schools) and that we are able to observe the following variables:

y_{ij} is a continuous outcome of interest for unit i in group j;

d_{ij} is a binary treatment indicator for for unit i in group j;

x_{ij} is a level-one pre-treatment covariate for unit i in group j;

w_j is a level-two variable for group j.

We also assume that an important omitted variable exists:

z_{ij} is an unobserved level-one pre-treatment covariate for unit i in group j.[1]

In addition, we assume that the treatment and control conditions are available within each group, meaning that any level-one unit i in any level-two group j could potentially experience treatment or not; and that the relative influence of the observed covariate x_{ij} to the unobserved covariate z_{ij} in determining treatment

[1]For the sake of this discussion, we will assume that z_{ij} is some omitted variable that is associated with both treatment status d_{ij} and outcome y_{ij} but is *uncorrelated* with the observed covariate x_{ij} so that including x_{ij} in a linear model that omits z_{ij} does not "soak up" any of the variance in y_{ij} or d_{ij} that is explained by z_{ij} (which would consequently attenuate the model's omitted variable bias). For this reason it might be more precise to write z_{ij} as $z_{ij}^{\perp x}$, but for ease of notation the remaining discussion will treat the "$\perp x$" superscript for z_{ij} as implied.

status d_{ij} varies between groups. To make the latter point more explicit, we could write a logit model for the treatment assignment process as follows:

$$\text{logit}(\mathbb{E}[d_{ij} \mid x_{ij}, z_{ij}]) = \phi_{0j} + \phi_{1j}x_{ij} + \phi_{2j}z_{ij}, \tag{10a}$$

such that the slope parameters ϕ_{1j} and ϕ_{2j} corresponding to x_{ij} and z_{ij}, respectively, vary by group. Then the expected treatment propensity for unit i in group j would be

$$\mathbb{E}[\Pr(d_{ij} = 1 \mid x_{ij}, z_{ij})] = \frac{1}{1 + \exp[-\phi_{0j} - \phi_{1j}x_{ij} - \phi_{2j}z_{ij}]}. \tag{10b}$$

Given our stipulation above that investigators possess some general qualitative knowledge regarding the heterogeneous treatment assignment process outlined in Equations (10a) and (10b), we assume that an additional dummy variable c_j can be encoded into the observed data where each group j is classified such that

$$c_j = \begin{cases} 0 \text{ if } \phi_{2j} = 0 \\ \\ 1 \text{ otherwise.} \end{cases} \tag{11}$$

In other words, $c_j = 0$ for the groups where the conditional exchangeability assumption is tenable on the basis of observed x_{ij} alone (because z_{ij} has no association with treatment status in those groups), and $c_j = 1$ for the groups where both x_{ij} and z_{ij} may be associated with treatment assignment.

We also suppose that group membership is not completely random, *i.e.*, that units associated with a given group "hang together" to some degree such that there is variation in the level-two means of x_{ij} and z_{ij}, denoted as follows:

$\bar{x}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$ is the level-two mean of x_{ij} for group j with mean $\mu_{\bar{x}}$ and variance $\sigma^2_{\bar{x}}$; and

$\bar{z}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} z_{ij}$ is the level-two mean of z_{ij} for group j with mean $\mu_{\bar{z}}$ and variance $\sigma^2_{\bar{z}}$.

Accordingly, we could compute the intraclass correlation of x_{ij}, *i.e.*, the proportion of variance that is between groups as

$$\text{ICC}_x = \sigma^2_{\bar{x}}/\sigma^2_x, \tag{12}$$

where $\sigma^2_{\bar{x}}$ is the variance of the level-two means $\bar{x}_{.j}$ and σ^2_x is the total overall sample variance in x_{ij}. Likewise, we could, in theory, compute the analogous proportion of between-group to total variance for $\bar{z}_{.j}$ and z_{ij}

were z_{ij} observed.

Data generating process

Given the data structure outlined above, we assume the following data generating process:

Level 1:

$$y_{ij} = \alpha_j + \tau_j d_{ij} + \beta x_{ij} + \gamma z_{ij} + e_{ij}, \quad e_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right) \tag{13a}$$

Level 2:

$$\begin{aligned} \alpha_j &= \alpha_{\cdot} + \eta_1 w_j + r_{\alpha j} \\ \tau_j &= \tau_{\cdot} + \eta_2 w_j + r_{\tau j} \end{aligned}, \quad r_{\alpha j},\ r_{\tau j} \overset{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \Omega\right) \tag{13b}$$

This two-level hierarchical data generating process comprises a random intercept (α_j) and a random treatment effect slope (τ_j), along with six fixed effects:

- β is the fixed effect of the observed level-one covariate x_{ij} on the level-one outcome y_{ij} (holding constant d_{ij} and z_{ij});
- γ is the fixed effect of the unobserved level-one covariate z_{ij} on the level-one outcome y_{ij} (holding constant d_{ij} and x_{ij});
- $\alpha_{\cdot}$ is the fixed intercept;
- $\tau_{\cdot}$ is the treatment fixed effect;
- η_1 is the fixed effect of the level-two moderator variable on the group-specific intercept;
- η_2 is the fixed effect of the level-two moderator variable on the group-specific treatment effect.

We also have three residual terms: e_{ij} at level one and $r_{\alpha j}$ and $r_{\tau j}$ at level two.

Omitted variable bias in this context

Focusing, for the time being, on the level-one component of the data generating process outlined in Equation (13a), a hypothetical investigator who was unable to observe z_{ij} would be limited to estimating the following restricted (biased) level-one model,

$$y_{ij} = \hat{\alpha}_j^{(\text{r})} + \hat{\tau}_j^{(\text{r})} d_{ij} + \hat{\beta}^{(\text{r})} x_{ij} + \hat{e}_{ij}^{(\text{r})}, \tag{14a}$$

whereas the unrestricted, unbiased level-one model the investigator would wish to estimate would include z_{ij}:

$$y_{ij} = \hat{\alpha}_j^{(\mathrm{u})} + \hat{\tau}_j^{(\mathrm{u})} d_{ij} + \hat{\beta}^{(\mathrm{u})} x_{ij} + \hat{\gamma} z_{ij} + \hat{e}_{ij}^{(\mathrm{u})}. \tag{15a}$$

The $\widehat{\mathrm{bias}}_j$ inherent to the sample estimate of the treatment effect for group j, then, is the difference between the estimate an actual investigator would obtain from the restricted level-one model specified in Equation (14a) and the estimate one would obtain from the full unrestricted model specified in Equation (15a), were z_{ij} observed:

$$\widehat{\mathrm{bias}}_j = \hat{\tau}_j^{(\mathrm{r})} - \hat{\tau}_j^{(\mathrm{u})}. \tag{16}$$

This is the traditional definition of omitted variable bias outlined in Angrist and Pischke (2009) and elsewhere. As shown in Cinelli and Hazlett (2020), we can decompose the group-level $\widehat{\mathrm{bias}}_j$ into "impact" and "imbalance" components by employing the Frisch-Waugh-Lovell theorem (Frisch & Waugh, 1933; Lovell, 1963, 2008) to "partial out" the observed covariate X_j. (Here, X_j, as well as D_j and Z_j respectively denote the $n_j \times 1$ column vectors of the values of x_{ij}, d_{ij}, and z_{ij} for group j):

$$\begin{aligned}
\hat{\tau}_j^{(\mathrm{r})} &= \frac{\mathrm{cov}(D_j^{\perp X_j}, Y_j^{\perp X_j})}{\mathrm{var}(D_j^{\perp X_j})} \\
&= \frac{\mathrm{cov}(D_j^{\perp X_j}, \hat{\tau}_j^{(\mathrm{u})} D_j^{\perp X_j} + \hat{\gamma} Z_j^{\perp X_j})}{\mathrm{var}(D_j^{\perp X_j})} \\
&= \hat{\tau}_j^{(\mathrm{u})} \left(\frac{\cancel{\mathrm{cov}(D_j^{\perp X_j}, D_j^{\perp X_j})}}{\mathrm{var}(D_j^{\perp X_j})} \right) + \hat{\gamma} \left(\frac{\mathrm{cov}(D_j^{\perp X_j}, Z_j^{\perp X_j})}{\mathrm{var}(D_j^{\perp X_j})} \right) \\
&= \hat{\tau}_j^{(\mathrm{u})} + \hat{\gamma} \underbrace{\left(\frac{\mathrm{cov}(D_j^{\perp X_j}, Z_j^{\perp X_j})}{\mathrm{var}(D_j^{\perp X_j})} \right)}_{\hat{\delta}_j} \\
&= \hat{\tau}_j^{(\mathrm{u})} + \underbrace{\hat{\gamma} \hat{\delta}_j}_{\widehat{\mathrm{bias}}_j}.
\end{aligned} \tag{17}$$

Expressed this way, $\widehat{\mathrm{bias}}_j$, *i.e.*, the bias of the sample estimate of the treatment effect for group j, comprises two components:

- $\hat{\gamma}$, the (fixed) "impact" of a unit change in the confounder z_{ij} on the linear expectation of the outcome y_{ij}, holding constant the observed covariate x_{ij} and treatment status d_{ij} as parameterized in Equation

(15a), the unrestricted level-one model:

$$y_{ij} = \hat{\alpha}_j^{(\mathrm{u})} + \hat{\tau}_j^{(\mathrm{u})} d_{ij} + \hat{\beta}^{(\mathrm{u})} x_{ij} + \hat{\gamma} z_{ij} + \hat{e}_{ij}^{(\mathrm{u})}; \text{ and}$$

- $\hat{\delta}_j$, the (group-varying) "imbalance" of the confounder z_{ij} in relation to the treatment d_{ij} after accounting for x_{ij} in group j. More specifically, if we could regress the confounder z_{ij} on d_{ij} and x_{ij} within each group,

$$z_{ij} = \hat{\iota}_j + \hat{\delta}_j d_{ij} + \hat{\psi}_j x_{ij} + \ \hat{e}_{zij}, \tag{18}$$

  then $\hat{\delta}_j$ would represent the expected change in the confounder z_{ij} in group j given a unit change in the treatment d_{ij}, holding constant the observed covariate x_{ij}.

Implications of group-varying treatment selection and conditional exchangeability

If we suppose that the $j = 1 \ldots J$ groups in our sample can be partitioned into subsets $J = \{J', J''\}$ where the unobserved confounder $z_{ij'}$ has no association with treatment status $d_{ij'}$ for one subset of the groups $j' = 1 \ldots J'$, but that $z_{ij''}$ potentially *is* associated with treatment status $d_{ij''}$ among the remaining groups $j'' = 1 \ldots J''$, then we can say that $\hat{\delta}_{j'} = 0$ in Equation (18) for the groups that constitute the former subset, and consequently

$$\hat{\tau}_j^{(\mathrm{r})} = \begin{cases} \hat{\tau}_{j'}^{(\mathrm{u})} & \text{for group } j' \text{ (because } \hat{\delta}_{j'} = 0); \\ \hat{\tau}_{j''}^{(\mathrm{u})} + \hat{\gamma}\hat{\delta}_{j''} & \text{for group } j''. \end{cases} \tag{19}$$

In other words, the group-specific treatment effect estimate $\hat{\tau}_j^{(\mathrm{r})}$ obtained from the restricted model is equivalent in expectation to the estimate from the corresponding unrestricted model (and therefore unbiased) for group j', but potentially biased for group j''.

3.2 A General Two-Level Random Intercept and Treatment Effect Model

Given the grouping structure and data generating process outlined in the previous section, a hypothetical investigator might be interested in estimating $\hat{\tau}_{\cdot}$ to get a sense of the main effect of treatment, $\hat{\omega}_{22}$ to understand random group-level variation in the treatment effect, and $\hat{\eta}_2$ to assess the group-level moderator variable's effect on the group-level treatment effect. Below we consider the differences between the ideal

model the investigator would wish they could run, a naive (and likely biased) restricted model that an investigator could feasibly run with actually-observed data, and an augmented restricted model that incorporates additional qualitative information about group-level variation in the treatment selection process in order to ameliorate omitted variable bias.

"Omniscient" unrestricted outcome model (had z_{ij} been observed)

Ideally, were it possible to observe z_{ij}, an investigator would want to estimate the following unrestricted hierarchical linear model, which would precisely mirror the data generating process in Equations (13a) and (13b), and would provide unbiased estimates of the parameters of interest:

Level 1:

$$y_{ij} = \hat{\alpha}_j^{(\mathrm{u})} + \hat{\tau}_j^{(\mathrm{u})} d_{ij} + \hat{\beta}^{(\mathrm{u})} x_{ij} + \hat{\gamma} z_{ij} + \hat{e}_{ij}^{(\mathrm{u})} \tag{20a}$$

Level 2:

$$\begin{aligned} \hat{\alpha}_j^{(\mathrm{u})} &= \hat{\alpha}_{\cdot}^{(\mathrm{u})} + \hat{\eta}_1^{(\mathrm{u})} w_j + \ \hat{r}_{\alpha}^{(\mathrm{u})}{}_j \\ \hat{\tau}_j^{(\mathrm{u})} &= \hat{\tau}_{\cdot}^{(\mathrm{u})} + \hat{\eta}_2^{(\mathrm{u})} w_j + \ \hat{r}_{\tau}^{(\mathrm{u})}{}_j, \end{aligned} \tag{20b}$$

where we assume the level-one and level-two residuals are distributed as follows:

$$\hat{e}_{ij}^{(\mathrm{u})} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \hat{\sigma}^{(\mathrm{u})^2}\right), \qquad \begin{pmatrix} \hat{r}_{\alpha}^{(\mathrm{u})}{}_j \\ \hat{r}_{\tau}^{(\mathrm{u})}{}_j \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \hat{\omega}_{11}^{(\mathrm{u})} & \hat{\omega}_{12}^{(\mathrm{u})} \\ \hat{\omega}_{21}^{(\mathrm{u})} & \hat{\omega}_{22}^{(\mathrm{u})} \end{bmatrix} \right). \tag{20c}$$

The "(u)" superscript in Equations (20a), (20b), and (20c) above distinguishes the unrestricted model's parameter estimates from the analogous parameter estimates produced by the restricted models below.

"Naive" restricted outcome model (given that z_{ij} is actually unobserved)

In practice, however, an investigator who was blind to the true data generating process, was unable to observe z_{ij}, and who did not have any additional information about group-level heterogeneity in the treatment assignment process might estimate the following group-mean-centered hierarchical linear model, where the "(r)" superscript distinguishes the naive restricted model's parameter estimates from the other models discussed in this section:

Level 1:

$$y_{ij} = \hat{\alpha}_j^{(\mathrm{r})} + \hat{\tau}_j^{(\mathrm{r})} d_{ij}^{\mathrm{mc}} + \hat{\beta}^{(\mathrm{r})} x_{ij}^{\mathrm{mc}} + \hat{e}_{ij}^{(\mathrm{r})}, \tag{21a}$$

Level 2:

$$\begin{aligned}\hat{\alpha}_j^{(\mathrm{r})} &= \hat{\alpha}_{\cdot}^{(\mathrm{r})} + \hat{\zeta}_{11}^{(\mathrm{r})}\bar{d}_{\cdot j} + \hat{\zeta}_{12}^{(\mathrm{r})}\bar{x}_{\cdot j} + \hat{\eta}_1^{(\mathrm{r})}w_j + \ \hat{r}_{\alpha j}^{(\mathrm{r})} \\ \hat{\tau}_j^{(\mathrm{r})} &= \hat{\tau}_{\cdot}^{(\mathrm{r})} + \hat{\zeta}_{21}^{(\mathrm{r})}\bar{d}_{\cdot j} + \hat{\zeta}_{22}^{(\mathrm{r})}\bar{x}_{\cdot j} + \hat{\eta}_2^{(\mathrm{r})}w_j + \ \hat{r}_{\tau j}^{(\mathrm{r})},\end{aligned} \tag{21b}$$

with the level-one and level-two residuals assumed to be distributed as follows:

$$\hat{e}_{ij}^{(\mathrm{r})} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \hat{\sigma}^{(\mathrm{r})^2}\right), \qquad \begin{pmatrix} \hat{r}_{\alpha j}^{(\mathrm{r})} \\ \hat{r}_{\tau j}^{(\mathrm{r})} \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \hat{\omega}_{11}^{(\mathrm{r})} & \hat{\omega}_{12}^{(\mathrm{r})} \\ \hat{\omega}_{21}^{(\mathrm{r})} & \hat{\omega}_{22}^{(\mathrm{r})} \end{bmatrix}\right). \tag{21c}$$

The naive restricted model above differs from the omniscient unrestricted model in two key respects. First, it omits the unobserved confounder z_{ij} but also (given the hypothetical investigator's lack of knowledge about the true data generating process) conservatively accounts for group-level means of the treatment indicator d_{ij} and observed covariate x_{ij}: At level two, $\bar{d}_{\cdot j}$ and $\bar{x}_{\cdot j}$ are the group-level means of d_{ij} and x_{ij}, respectively, and at level one, $d_{ij}^{\text{mc}} = d_{ij} - \bar{d}_{\cdot j}$ and $x_{ij}^{\text{mc}} = x_{ij} - \bar{x}_{\cdot j}$ are the group-mean-centered values of d_{ij} and x_{ij}. (This centering approach follows recommendations outlined in Raudenbush and Bryk (2002), Schunck (2013), Hazlett and Wainstein (2022), and elsewhere to mitigate potential bias induced by correlated random effects.) This naive model is nevertheless problematic, however, because the fixed and group-level treatment effect estimates ($\hat{\tau}_{\cdot}^{(\mathrm{r})}$ and $\hat{\tau}_j^{(\mathrm{r})}$) that it produces would be susceptible to bias due to the omission of z_{ij}. We might also be concerned about the reliability of other parameter estimates of interest, such as $\hat{\eta}_2^{(\mathrm{r})}$, the estimated effect of the group-level moderator variable w_j on the group-level treatment effect, or $\hat{\omega}_{22}^{(\mathrm{r})}$, the estimated variance of the group-level treatment effect.

"Augmented" restricted outcome model

Supposing, on the other hand, that it were possible to incorporate the kind of qualitative knowledge about group-level variation in the treatment assignment process described in Section 3.1, our hypothetical investigator could augment the naive restricted model at level two with the group-level dummy variable c_j defined in Equation (11). This model would theoretically account for the omitted variable bias owing to the unobserved covariate z_{ij}'s association with treatment status d_{ij} among the subset of groups where treatment assignment was confounded:

Level 1:

$$y_{ij} = \hat{\alpha}_j + \hat{\tau}_j d_{ij}^{\text{mc}} + \hat{\beta} x_{ij}^{\text{mc}} + \hat{e}_{ij}, \tag{22a}$$

Level 2:

$$\begin{aligned}\hat{\alpha}_j &= \hat{\alpha}_{\cdot} + \hat{\zeta}_{11}\bar{d}_{\cdot j} + \hat{\zeta}_{12}\bar{x}_{\cdot j} + \hat{\eta}_1 w_j + \hat{\theta}_1 c_j + \hat{r}_{\alpha j} \\ \hat{\tau}_j &= \hat{\tau}_{\cdot} + \hat{\zeta}_{21}\bar{d}_{\cdot j} + \hat{\zeta}_{22}\bar{x}_{\cdot j} + \hat{\eta}_2 w_j + \hat{\theta}_2 c_j + \hat{r}_{\tau j},\end{aligned} \tag{22b}$$

with the level-one and level-two residuals assumed to be distributed as follows:

$$\hat{e}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \hat{\sigma}^2\right), \qquad \begin{pmatrix} \hat{r}_{\alpha j} \\ \hat{r}_{\tau j} \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \hat{\omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{21} & \hat{\omega}_{22} \end{bmatrix}\right). \tag{22c}$$

If the inclusion of the group-level dummy variable c_j accurately flags the groups affected by omitted variable bias via confounded treatment assignment, then $\hat{\tau}_{\cdot}$ in the level-two model in Equation (22b) should be an unbiased estimate of the treatment fixed effect $\tau_{\cdot}$. The coefficient $\hat{\theta}_2$ also becomes an additional parameter of potential interest, corresponding to the estimated fixed effect of group-level confounded treatment assignment on the group-level treatment effect.

The next section outlines a simulation study that examines the utility of this sort of augmented model for mitigating omitted variable bias in different grouped data structures where investigators are interested in estimating a main effect of the treatment while also accounting for random and nonrandom variation in the group-specific treatment effects.

4 Methods

4.1 Data Generation

Nested data structures

The simulation study considers three different nested data structures:

(a) $J = 10$ groups with $n_j = 28$ units per group;

(b) $J = 20$ groups with $n_j = 20$ units per group;

(c) $J = 50$ groups with $n_j = 8$ units per group.

Recognizing that the overall number and specific choices of different nested data structures to include in the simulation study is arbitrary, these particular selections take the following criteria into consideration: First, they should correspond to sample sizes that that applied researchers in education or other social policy fields might employ, given a reasonable set of cost constraints for recruiting groups of participants. In other words, the sample size should be neither unrealistically large nor so small as to make the design drastically underpowered.

Second, the range of data structures should be appropriate for estimating a variety of different combinations of focal parameter magnitudes. In multisite study design, expanding the number of groups in the sample (while holding n_j constant) increases statistical power much faster than increasing the number of units per group while holding J constant (Raudenbush & Liu, 2000). In practice, however, it will often be very costly to add more sites to a sample and relatively inexpensive to recruit additional participants within each site.

Also, depending on the context, an investigator will probably be most concerned with one or two parameter estimates from the hierarchical linear model and consequently less interested in the others. For example, a researcher who suspects that the variability between group-specific treatment effects is relatively small would be inclined to focus on obtaining an accurate estimate of the treatment main effect and much less worried about estimating group-level moderator effects or random effect variance. On the other hand if there is reason to believe *a priori* that the main effect of the treatment is relatively small but substantial between-group treatment effect variation exists, it may be much more interesting to understand the main effect random variance or to isolate the effect of a group-level moderator variable on the group-level treatment effect.

With those considerations in mind, this simulation study examines three two-level nested data structures from Raudenbush and Liu (2000), which outlined a range of multisite study designs to optimize statistical power for estimating key parameters of interest as a function of treatment main effect size, main effect variance, group moderator effect, and the cost ratio of sampling additional sites relative to sampling additional participants per site. Table 1 shows a set of presupposed true parameter values for the treatment main effect $\tau_{.}$, intercept and treatment main effect variances ω_{11} and ω_{22}, as well as the group moderator effect on group-level treatment effect η_2, along with the corresponding statistical power for estimating each parameter, given its magnitude, via a two-level hierarchical linear model with each of the three nested data structures used in the simulation:

# of units per group n_j	# of groups J	main effect $\tau_{.}$	main effect variance ω_{11}, ω_{22}	group moderator effect η_2	power for $\tau_{.}$	power for ω_{11}, ω_{22}	power for η_2
28	10	0.4	0.10	0.4	0.629	0.344	0.205
20	20	0.3	0.05	0.4	0.721	0.185	0.395
8	50	0.2	0.15	0.6	0.405	0.350	0.732

Table 1: Nested data structures, effect sizes, and corresponding power for simulated data sets (power calculations for hypothetical study designs from Raudenbush and Liu (2001)).

The scenario in the first row of Table 1, with 10 groups and 28 units per group, supposes a medium-sized treatment main effect ($\tau_{.} = 0.4$), medium-sized random effect variance ($\omega_{11} = \omega_{22} = 0.10$), and a medium-

sized group-level moderator effect ($\eta_2 = 0.4$). The scenario in the second row, with 20 groups and 20 units per group, supposes a small-to-medium-sized treatment main effect ($\tau_{\cdot} = 0.3$), a small random effect variance ($\omega_{11} = \omega_{22} = 0.05$), and a medium-sized group-level moderator effect ($\eta_2 = 0.4$). Finally, the scenario in the third row, with 50 groups and 8 units per group, supposes a small treatment main effect ($\tau_{\cdot} = 0.2$), a relatively large random effect variance ($\omega_{11} = \omega_{22} = 0.15$), and a large group-level moderator effect ($\eta_2 = 0.6$). In practice, the first two sample structures might be appropriate (if slightly under-powered) for investigators primarily interested in estimating the main effect of treatment, while the third sample structure would be more useful for an investigator focused on understanding between-group variance in treatment effects.

Group-varying treatment selection and variation in confounded treatment assignment

For each of the three nested data structure scenarios described above, the simulation study varies the proportion of groups with unconfounded treatment assignment processes (*i.e.*, groups where the unobserved covariate z_{ij} has no association with treatment status d_{ij}) from 10% to 30% to 50%. Given that the omitted variable bias in the naive restricted model approaches zero as the proportion of unconfounded groups in the sample approaches 100%, our interest centers around scenarios where the naive restricted model would perform the worst and where the augmented model has the most potential to reduce bias.

In addition to varying the proportion of unconfounded groups in each sample design, the simulation study also varies the degree of group-level confounding by modulating the coefficient ϕ_{2j} on unobserved z_{ij} from Equations (10a) and (10b) among the groups with confounded treatment assignment from 1 to 2 to 3 to 6, while holding the coefficient ϕ_{1j} on observed x_{ij} constant at 1. This enables comparison between the naive and augmented restricted models' performance over a range of magnitudes of omitted variable bias.

To summarize the various permutations, the simulation study looks at three different nested sample designs and varies the proportion of unconfounded groups three ways over four different magnitudes of omitted variable bias. For each of these 36 permutations, the simulation randomly generated 1000 data sets in order to compare the relative performance of the three models outlined in Section 3.2.

Covariates, parameters, and outcome

The simulated data sets were generated using R Statistical Software (R Core Team, 2023) with each iteration proceeding as follows:

1. The intraclass correlations of x_{ij} and z_{ij}, *i.e.*, the proportion of the total variance of each variable that is between groups, is set to 0.1:[2]

[2] An intraclass correlation of 0.1 is within the range of what might conventionally be regarded as a "medium-sized" intraclass correlation in the context of an education research study involving students nested within intact schools (Murnane & Willett, 2010).

$\text{ICC}_x = \text{ICC}_z = 0.1$.

2. The fixed effects are set as follows:

   $\tau_{\cdot} \in \{0.2, 0.3, 0.4\}$, depending on the nested data structure (see Table 1);

   $\beta = 0.75$;

   $\gamma = 0.75$;

   $\eta_2 \in \{0.4, 0.6\}$, depending on the nested data structure (see Table 1).

   These values correspond to what might be considered small (0.2) to medium-sized (0.4) treatment fixed effects ($\tau_{\cdot}$) and medium-sized 0.4 to large 0.6 level-two moderator fixed effects (η_2). The level-one covariates x_{ij} and z_{ij} are both strongly predictive of the level-one outcome y_{ij} with their fixed effects (β and γ, respectively) set at 0.75. The fixed intercept ($\alpha_{\cdot}$) and the level-two moderator effect on the group-specific intercept (η_1), both of which might be regarded as "nuisance parameters," are assigned some arbitrary value from a normally distributed random draw with mean 0 and standard deviation 0.2.

3. The level-one covariates x_{ij} and z_{ij} are generated randomly from a standard normal draw and their covariance is set to zero:

$$\begin{pmatrix} x_{ij} \\ z_{ij} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right).$$

4. The level-two moderator variable w_j is generated randomly from a standard normal draw:

   $w_j \sim \mathcal{N}(0, 1)$.

5. Depending on the iteration, 10%, 30%, or 50% of the J groups in the data set are randomly selected to be "unconfounded," with the remainder designated to exhibit unobserved confounding in their group-level treatment assignment process.

6. Treatment propensity for each unit is then computed according to Equation (10b):

$$p_{ij} = \Pr(d_{ij} = 1 \mid x_{ij}, z_{ij}) = \frac{1}{1 + \exp[-\phi_{0j} - \phi_{1j} x_{ij} - \phi_{2j} z_{ij}]},$$

   with the coefficients $\phi_{0j}, \phi_{1j}, \phi_{2j}$ set to

   $\phi_{0j} = 0$

$$\phi_{1j} = 1$$

$$\phi_{2j} = \begin{cases} 0 \text{ for unconfounded groups where } z_{ij} \text{ plays no role in treatment assignment} \\ \phi_{2j} \in \{1, 2, 3, 6\} \text{ otherwise (to vary the resulting magnitude of omitted variable bias).} \end{cases}$$

Then, treatment status d_{ij} is determined randomly from a binomial draw with parameters $n = 1$ and $p = p_{ij}$:

$$d_{ij} \sim \mathcal{B}(1, p_{ij})$$

7. The level-one residual (e_{ij}) values are generated randomly from a standard normal draw:

$$e_{ij} \sim \mathcal{N}(0, 1),$$

and the level-two residuals ($r_{\alpha j}$) and ($r_{\tau j}$) are generated randomly from a bivariate normal distribution with zero covariance and differing variances depending on the number of level-two groups and level-one units per group, as outlined in Table 1:

$$\begin{pmatrix} r_{\alpha j} \\ r_{\tau j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \omega_{11} & 0 \\ 0 & \omega_{22} \end{bmatrix} \right);$$

$\omega_{11}, \omega_{22} \in \{0.05, 0.10, 0.15\}$, depending on the nested data structure (see Table 1).

8. Finally, the outcome y_{ij} is determined based on data generation process outlined above in Equations (13a) and (13b) (shown below in combined form),

$$y_{ij} = \underbrace{(\alpha_{\cdot} + \eta_1 w_j + r_{\alpha j})}_{\alpha_j} + \underbrace{(\tau_{\cdot} + \eta_2 w_j + r_{\tau j})}_{\tau_j} d_{ij} + \beta x_{ij} + \gamma z_{ij} + e_{ij}.$$

4.2 Data Analysis

For each of the 36,000 simulated data sets generated according to the procedure above, the unbiased "omniscient" unrestricted as well as the "naive" and "augmented" restricted models were then estimated using restricted maximum likelihood via the `lmer()` function from the R Statistical Software package `lme4` (Bates, Mächler, Bolker, & Walker, 2015).

The following section outlines the naive and augmented models' comparative performance relative to the unrestricted model across the different permutations of sample structures, proportions of unconfounded groups, and magnitudes of unobserved confounding in the treatment assignment process, focusing on the model parameters likely to be of substantive interest to investigators working with these sorts of study designs, including

(a) $\hat{\tau}_{\cdot}^{(\mathrm{r})}$ and $\hat{\tau}_{\cdot}$, the treatment main effect;

(b) $\hat{\eta}_2^{(\mathrm{r})}$ and $\hat{\eta}_2$, the effect of the group-level moderator w_j on estimated group-level treatment effect $\hat{\tau}_j^{(\mathrm{r})}$ or $\hat{\tau}_j$, respectively (holding constant the group-level means $\bar{d}_{\cdot j}$ and $\bar{x}_{\cdot j}$);

(c) $\hat{\omega}_{22}^{(\mathrm{r})}$ and $\hat{\omega}_{22}$, the random variance of the treatment main effect;

(d) $\hat{\beta}^{(\mathrm{r})}$ and $\hat{\beta}$, the fixed level-one effect of covariate x_{ij} on outcome y_{ij}, holding treatment status d_{ij} constant; and

(e) $\hat{\theta}_2$, the effect of the group-level confounded treatment assignment indicator c_j on estimated group-level treatment effect $\hat{\tau}_j$ in the augmented model.

5 Results

5.1 Comparison of Naive and Augmented Restricted Model Performance

Table 2 shows selected mean parameter estimates from the naive and augmented models for the simulations where omitted variable bias was maximized (*i.e.*, $\phi_{2j} = 6$ in the treatment assignment process detailed in Equation (10a) for groups with confounded treatment assignment),[3] and contrasts the mean parameter estimate over each permutation of 1000 simulated data sets to the corresponding true parameter value specified in the data generation procedure.

For treatment fixed effect $\tau_{\cdot}$, the naive model estimates had a mean bias of between 0.49 to 1.00, depending on the study design and the proportion of groups with unconfounded treatment assignment. The mean bias of the corresponding estimates from the augmented model ranged from 0.02 to 0.07. In other words, inclusion of the "confounded treatment assignment" dummy variable in the level-two model absorbed between 88% and 98% of the omitted variable bias present in the naive model.

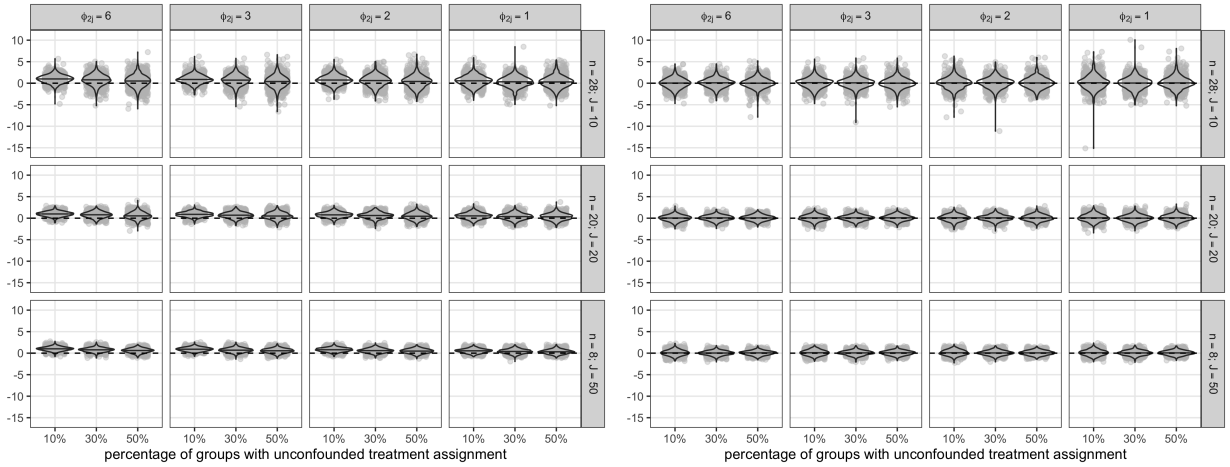
Figure 1 plots the error of the naive, augmented, and omniscient model estimates of the treatment fixed effect relative to the true value of $\tau_{\cdot}$ for each simulated data set. Although the augmented model outperformed the naive model in terms of bias, the augmented model's error variance was substantially larger than the unrestricted model that accounts for the unobserved confounder z_{ij} at level one. The large error variance of the restricted models is especially problematic when the number of groups is relatively small, as in the top row of the subfigures where $J = 10$.

We can estimate the bias of the naive and augmented models' estimates of the treatment fixed effect for each data set by examining their respective differences in relation to the unrestricted model's estimate (*i.e.*,

[3] Analogous tables for the other values of ϕ_{2j} appear in the Appendix.

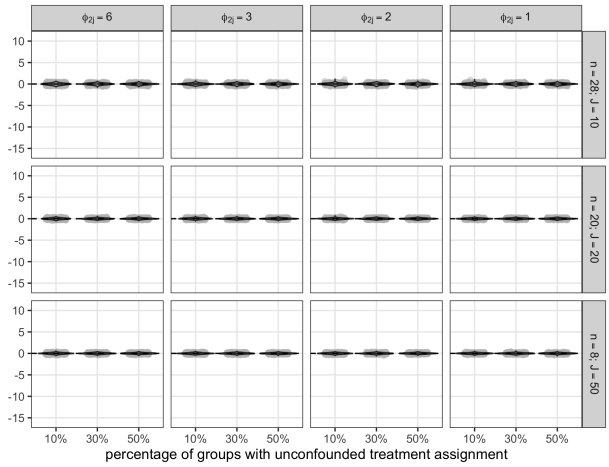
Table 2: Selected mean parameter estimates from simulations where $\phi_{2j} = 6$

parameter of interest	study design	% of groups unconfounded	true value	mean "naive" estimate	mean "augmented" estimate	"naive" bias	"augmented" bias	bias reduction
$\tau_.$	n=28; J=10	10%	0.40	1.381	0.419	0.981	0.019	0.962
		30%	0.40	1.162	0.473	0.762	0.073	0.689
		50%	0.40	0.894	0.418	0.494	0.018	0.476
	n=20; J=20	10%	0.30	1.274	0.318	0.974	0.018	0.956
		30%	0.30	1.091	0.353	0.791	0.053	0.738
		50%	0.30	0.872	0.367	0.572	0.067	0.505
	n=8; J=50	10%	0.20	1.203	0.273	1.003	0.073	0.930
		30%	0.20	0.967	0.228	0.767	0.028	0.739
		50%	0.20	0.794	0.244	0.594	0.044	0.550
η_2	n=28; J=10	10%	0.40	0.419	0.414	0.019	0.014	0.005
		30%	0.40	0.424	0.416	0.024	0.016	0.008
		50%	0.40	0.393	0.392	-0.007	-0.008	-0.001
	n=20; J=20	10%	0.40	0.398	0.400	-0.002	0.000	0.002
		30%	0.40	0.397	0.397	-0.003	-0.003	0.000
		50%	0.40	0.404	0.398	0.004	-0.002	0.002
	n=8; J=50	10%	0.60	0.591	0.592	-0.009	-0.008	0.001
		30%	0.60	0.606	0.602	0.006	0.002	0.004
		50%	0.60	0.605	0.606	0.005	0.006	-0.001
ω_{22}	n=28; J=10	10%	0.10	0.105	0.044	0.005	-0.056	-0.051
		30%	0.10	0.215	0.048	0.115	-0.052	0.063
		50%	0.10	0.272	0.056	0.172	-0.044	0.128
	n=20; J=20	10%	0.05	0.051	0.017	0.001	-0.033	-0.032
		30%	0.05	0.127	0.016	0.077	-0.034	0.043
		50%	0.05	0.154	0.017	0.104	-0.033	0.071
	n=8; J=50	10%	0.15	0.105	0.056	-0.045	-0.094	-0.049
		30%	0.15	0.221	0.069	0.071	-0.081	-0.010
		50%	0.15	0.245	0.058	0.095	-0.092	0.003
β	n=28; J=10	10%	0.75	0.686	0.690	-0.064	-0.060	0.004
		30%	0.75	0.692	0.699	-0.058	-0.051	0.007
		50%	0.75	0.704	0.713	-0.046	-0.037	0.009
	n=20; J=20	10%	0.75	0.674	0.682	-0.076	-0.068	0.008
		30%	0.75	0.680	0.693	-0.070	-0.057	0.013
		50%	0.75	0.692	0.709	-0.058	-0.041	0.017
	n=8; J=50	10%	0.75	0.675	0.684	-0.075	-0.066	0.009
		30%	0.75	0.675	0.695	-0.075	-0.055	0.020
		50%	0.75	0.683	0.708	-0.067	-0.042	0.025



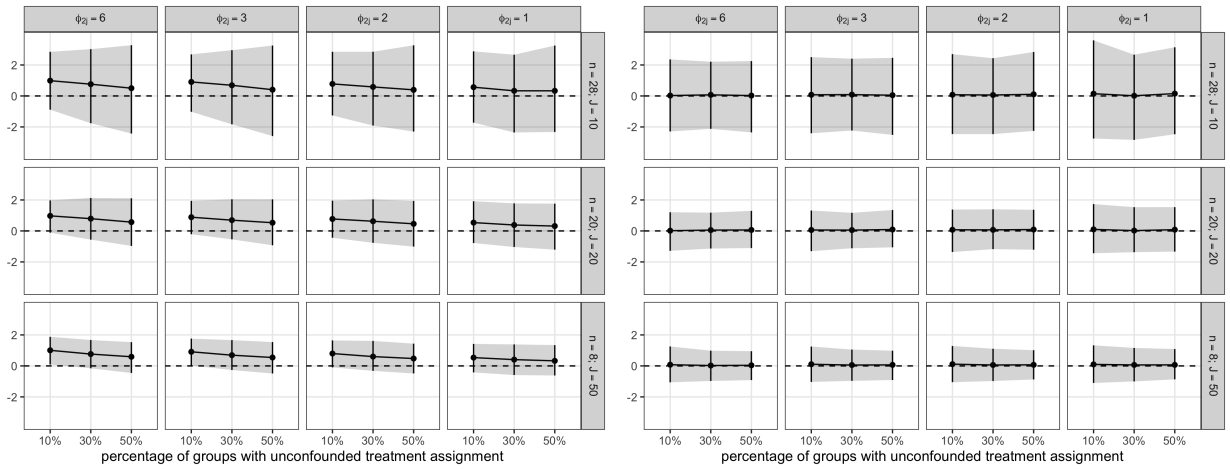
(a) Naive restricted model error in estimates of treatment main effect: $\text{error}^{(\text{r})} = \hat{\tau}_{\cdot}^{(\text{r})} - \tau_{\cdot}$.

(b) Augmented restricted model error in estimates of treatment main effect: error $= \hat{\tau}_{\cdot} - \tau_{\cdot}$.



(c) Omniscient unrestricted model error in estimates of treatment main effect: $\text{error}^{(\text{u})} = \hat{\tau}_{\cdot}^{(\text{u})} - \tau_{\cdot}$.

Figure 1: Error of (a) "naive" restricted, (b) "augmented" restricted, and (c) "omniscient" unrestricted model estimates of treatment fixed effect relative to true parameter ($\tau_{\cdot}$). Each scatterplot point corresponds to a simulated data set. Violin plots represent density of estimates over y-axis for each permutation with solid horizontal line segment denoting the median of each distribution. Dashed lines set at zero.



(a) Naive restricted model estimates of $\widehat{\text{bias}}_{\tau_\cdot}^{(\text{r})} = \hat{\tau}_\cdot^{(\text{r})} - \hat{\tau}_\cdot^{(\text{u})}$.

(b) Augmented restricted model estimates of $\widehat{\text{bias}}_{\tau_\cdot} = \hat{\tau}_\cdot - \hat{\tau}_\cdot^{(\text{u})}$.

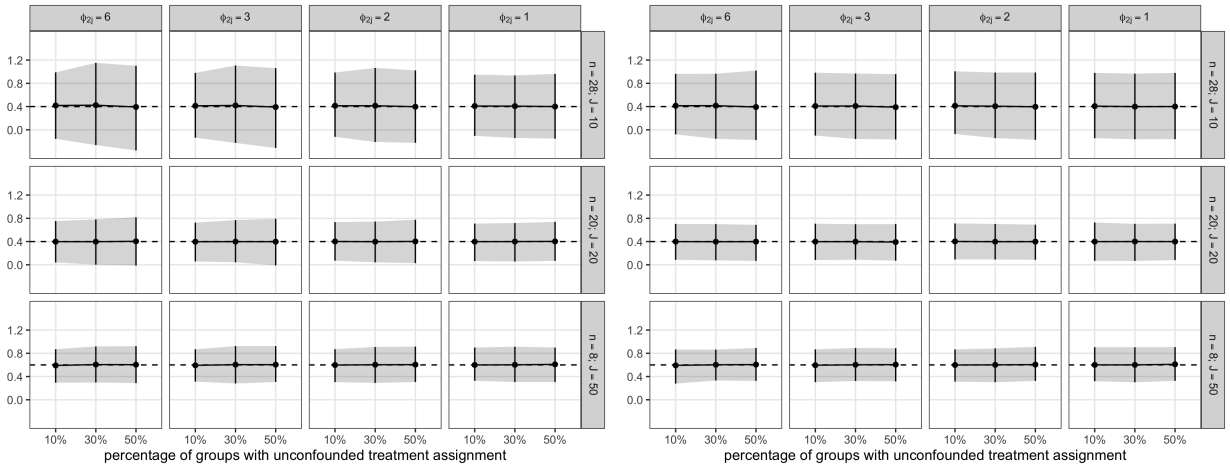
Figure 2: Means and 95% confidence intervals of (a) "naive" and (b) "augmented" restricted model estimates of treatment fixed effect omitted variable bias relative to omniscient model. Solid points correspond to mean estimates and vertical lines extend to upper and lower bounds of 95% confidence interval over 1000 simulated data sets for each permutation. Dashed lines set at zero.

$\widehat{\text{bias}}_{\tau_\cdot}^{(\text{r})} = \hat{\tau}_\cdot^{(\text{r})} - \hat{\tau}_\cdot^{(\text{u})}$ for the naive model, and $\widehat{\text{bias}}_{\tau_\cdot} = \hat{\tau}_\cdot - \hat{\tau}_\cdot^{(\text{u})}$ for the augmented model). Figure 2 plots the means and 95% confidence intervals of the treatment fixed effect bias estimates for each permutation of 1000 data sets.

As shown in Table 2, the naive model produced relatively unbiased mean estimates of the group-level moderator effect (η_2), within the range of -0.01 and 0.02 of the true parameter values. The augmented model's estimates were no worse on average, falling within approximately the same range. Figure 3 plots the means and 95% confidence intervals of the two models' estimates of η_2 in each scenario in relation to the corresponding true parameter values specified in the data generation procedure. The two models are broadly similar in terms of their confidence interval ranges and relative lack of bias in each instance.

As shown in Table 2 above and Figure 4 below, the augmented model did not perform straightforwardly better than the naive model in terms of bias in estimating the random variance of the group-level treatment effect (ω_{22}). In each of the simulated scenarios, the augmented model underestimated ω_{22} on average. On the other hand, the naive model exhibited a larger bias in the opposite direction in some scenarios, especially as the proportion of groups with unconfounded treatment assignment increased.

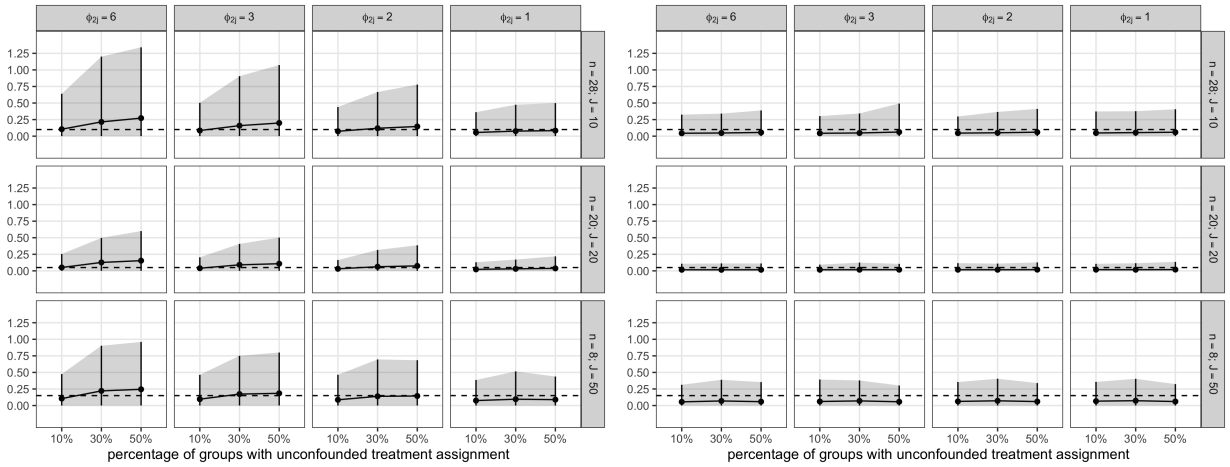
Although the augmented model did not exhibit a substantial improvement over the naive model in terms of bias reduction in estimating the variance of the treatment random effect, the mean squared error of the augmented model's estimates of ω_{22} was smaller than the MSE of the naive model's estimates, as is visually apparent in Figure 4.



(a) Naive restricted model estimates of $\hat{\eta}_2^{(\mathrm{r})}$.

(b) Augmented restricted model estimates of $\hat{\eta}_2$.

Figure 3: Means and 95% confidence intervals of (a) "naive" and (b) "augmented" restricted model estimates of fixed effect of group-level moderator variable on group-level treatment effect. Solid points correspond to mean estimates and vertical lines extend to upper and lower bounds of 95% confidence interval over 1000 simulated data sets for each permutation. Dashed lines correspond to the true parameter value for each study design structure.



(a) Naive restricted model estimates of $\hat{\omega}_{22}^{(\mathrm{r})}$.

(b) Augmented restricted model estimates of $\hat{\omega}_{22}$.

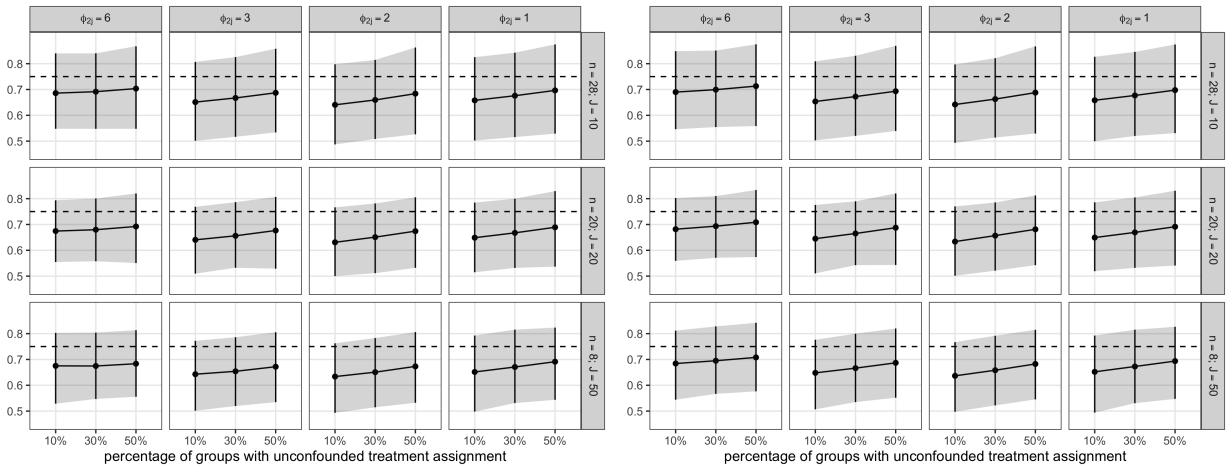
Figure 4: Means and 95% confidence intervals of (a) "naive" and (b) "augmented" restricted model estimates of group-level treatment random effect variance. Solid points correspond to mean estimates and vertical lines extend to upper and lower bounds of 95% confidence interval over 1000 simulated data sets for each permutation. Dashed lines correspond to the true parameter value for each study design structure.

Table 3: Mean squared error of unrestricted, naive, and augmented model estimates of treatment random effect variance (ω_{22})

selection bias magnitude	study design	% of groups unconfounded	true ω_{22}	unrestricted MSE	"naive" MSE	"augmented" MSE	augmented/naive MSE ratio
$\phi_{2j} = 6$	n=28; J=10	10%	0.10	0.008	0.033	0.012	0.351
		30%	0.10	0.008	0.127	0.017	0.135
		50%	0.10	0.008	0.176	0.015	0.086
	n=20; J=20	10%	0.05	0.002	0.005	0.002	0.427
		30%	0.05	0.002	0.027	0.002	0.078
		50%	0.05	0.002	0.040	0.002	0.057
	n=8; J=50	10%	0.15	0.015	0.020	0.017	0.859
		30%	0.15	0.016	0.075	0.022	0.290
		50%	0.15	0.015	0.077	0.018	0.231
$\phi_{2j} = 3$	n=28; J=10	10%	0.10	0.008	0.024	0.012	0.494
		30%	0.10	0.008	0.074	0.017	0.224
		50%	0.10	0.008	0.100	0.023	0.231
	n=20; J=20	10%	0.05	0.002	0.003	0.002	0.637
		30%	0.05	0.002	0.015	0.002	0.151
		50%	0.05	0.002	0.021	0.003	0.119
	n=8; J=50	10%	0.15	0.016	0.020	0.018	0.878
		30%	0.15	0.015	0.042	0.018	0.436
		50%	0.15	0.015	0.049	0.017	0.355
$\phi_{2j} = 2$	n=28; J=10	10%	0.10	0.008	0.025	0.015	0.587
		30%	0.10	0.008	0.053	0.017	0.332
		50%	0.10	0.008	0.054	0.020	0.361
	n=20; J=20	10%	0.05	0.002	0.003	0.002	0.839
		30%	0.05	0.002	0.008	0.002	0.283
		50%	0.05	0.002	0.011	0.003	0.236
	n=8; J=50	10%	0.15	0.016	0.019	0.018	0.958
		30%	0.15	0.015	0.034	0.020	0.593
		50%	0.15	0.015	0.033	0.017	0.523
$\phi_{2j} = 1$	n=28; J=10	10%	0.10	0.007	0.014	0.013	0.905
		30%	0.10	0.008	0.027	0.018	0.660
		50%	0.10	0.007	0.022	0.017	0.772
	n=20; J=20	10%	0.05	0.002	0.002	0.002	0.904
		30%	0.05	0.002	0.003	0.002	0.819
		50%	0.05	0.002	0.004	0.002	0.638
	n=8; J=50	10%	0.15	0.015	0.019	0.018	0.972
		30%	0.15	0.015	0.025	0.021	0.858
		50%	0.15	0.016	0.020	0.018	0.878

Table 3 details the mean squared error of the omniscient, naive, and augmented models' estimates of the treatment random effect variance parameter ω_{22} in each permutation. In every case, the MSE of the augmented model was smaller than the naive model's MSE, and in many cases the augmented model's MSE was roughly equivalent to (or only slightly larger than) the MSE of the omniscient model that perfectly mirrors the data generating process.

Both the naive and augmented models produced negatively biased mean estimates of β, the fixed effect of the observed covariate x_{ij} on the outcome y_{ij} (holding constant treatment status d_{ij}), with bias ranging from between -0.08 and -0.05 for the observed model and -0.07 and -0.04 for the augmented model. The augmented model performed modestly better, absorbing between 6.3% and 37.3% of the naive model's bias depending on the study design and the proportion of unconfounded groups. For each study design permutation, the augmented model's performance improved relative to the naive model as the proportion of unconfounded



(a) Naive restricted model estimates of $\hat{\beta}^{(r)}$.

(b) Augmented restricted model estimates of $\hat{\beta}$.

Figure 5: Means and 95% confidence intervals of (a) "naive" and (b) "augmented" restricted model estimates of fixed effect of unit-level covariate on unit-level outcome, holding constant unit-level treatment status. Solid points correspond to mean estimates and vertical lines extend to upper and lower bounds of 95% confidence interval over 1000 simulated data sets for each permutation. Dashed lines correspond to the true parameter value ($\beta = 0.75$ for each study design structure).

groups increased from 10% to 30% to 50%. Although β in this context may not be likely to be a parameter of primary interest to investigators, this minor improvement may be an added benefit in the augmented model's favor.

5.2 Estimation of the Effect of Group-Level Confounded Treatment Assignment on Group-Level Treatment Effect

Investigators who are interested in exploiting the characteristics of these kinds of data structures to quantify the magnitude of omitted variable bias among the confounded groups as it relates to the group-level treatment effect estimates may be interested in the augmented model's estimate of θ_2, the coefficient on the group-level "confounded treatment assignment" indicator in the level-two model where the group-level treatment effect is the outcome. Figure 6 plots the means and 95% confidence intervals of $\hat{\theta}_2$ from each permutation. As would be expected, there is substantial uncertainty around the estimates of θ_2 when the proportion of unconfounded groups in the sample is very small but the confidence intervals narrow as the proportion of unconfounded groups approaches 30% or more.

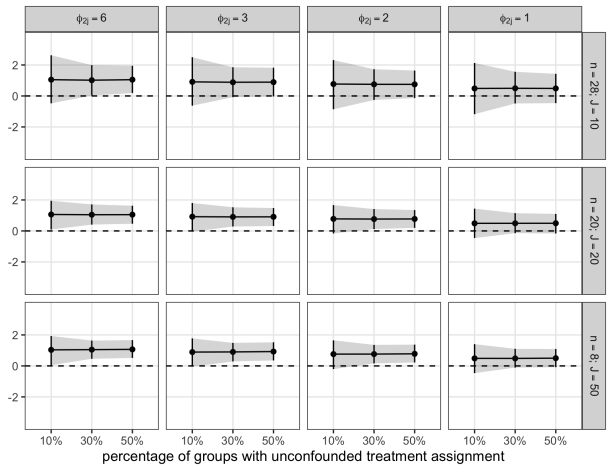


Figure 6: Means and 95% confidence intervals of "augmented" restricted model estimates of fixed effect of group-level confounded treatment assignment indicator on group-level treatment effect ($\hat{\theta}_2$). Solid points correspond to mean estimates and vertical lines extend to upper and lower bounds of 95% confidence interval over 1000 simulated data sets for each permutation. Dashed lines set at zero.

---

6 Discussion

6.1 Implications and Next Steps

As this simulation study shows, incorporating accurate qualitative information about sample-wide variation in group-level treatment selection processes can substantially reduce omitted variable bias in estimating the main effect of treatment and reduce the mean squared error of treatment random effect variance estimates in multisite studies where treatment is assigned to various units within each group.

That said, the hypothetical examples examined here are limited to "best-case" scenarios in which the conditional exchangeability assumption is unequivocally valid among a subset of groups in a sample, and where each of those groups is correctly flagged by the investigators carrying out the data analysis. In practice, such confident assumptions may not be warranted. Cautious researchers in these situations will worry about how reliably they can ascertain whether (or which of) the ostensibly unconfounded groups in their sample actually satisfy the criteria for conditional exchangeability. Further work along these lines could consider how the utility of the approach outlined here would be attenuated under a weaker set of assumptions, for example, by looking at the effect of mis-classifying actually-confounded groups as "unconfounded" based on flawed information about group-specific treatment assignment processes.

This simulation study also only considers situations where groups can be neatly classified as confounded or perfectly unconfounded, which may be an unrealistic oversimplification. Another useful line of inquiry would be to study how the robustness of hierarchical linear model inferences are affected by continuously

varying degrees of group-level confounding within a sample instead of the straightforward binary distinction considered here.

Some other related topics worth probing in future work include the effects of group-varying treatment selection processes with regard to more complex hierarchical linear models with additional random slopes besides the intercept and treatment effect, how interactions between unobserved confounders at the unit level and the group level affect inference, the implications of heterogeneous treatment selection on causal inference in nested data structures with three or more levels, in unbalanced nested data structures with uneven numbers of units within each group, as well as in cross-classified and multiple-membership data structures where units and groups are not tidily nested between levels.

In the absence of a more general approach to sensitivity analysis for multilevel models, the most persuasive way that applied researchers may be able to assert the robustness of causal inferences drawn from hierarchical linear models is by reporting results of *ad hoc* sensitivity analyses appropriate to the context at hand. Ideally, these tests should account for plausible sources and magnitudes of unobserved confounding, given relevant domain knowledge, the specific data structure in question, and the particular specifications of the hierarchical linear models employed in the analyses. Given the parameter estimates obtained from a potentially confounded model, researchers could consider implications of omitted variable bias along the lines of what Cinelli and Hazlett (2020) proposed for OLS regression. A starting point could be to simulate a plausibly strong source of confounding at level one, level two, or both levels, and then report on how unobserved confounding of that magnitude would shift the confidence intervals for key parameter estimates of interest. Hong and Raudenbush (2006) demonstrated a useful example of this approach as it applies to estimating the effects of kindergarten retention on reading and math achievement scores. To provide yet more transparency, researchers could extended their sensitivity analysis reporting to include the magnitude of confounding at either level that would be necessary to render key parameter estimates equal to zero, as well as the minimum degree of confounding that would render a statistically significant estimate with the opposite sign.

7 Appendix

7.1 Supplementary Tables

Table 4: Selected mean parameter estimates from simulations where $\phi_{2j} = 3$

parameter of interest	study design	% of groups unconfounded	true value	mean "naive" estimate	mean "augmented" estimate	"naive" bias	"augmented" bias	bias reduction
$\tau_.$	n=28; J=10	10%	0.40	1.302	0.473	0.902	0.073	0.829
		30%	0.40	1.092	0.489	0.692	0.089	0.603
		50%	0.40	0.796	0.441	0.396	0.041	0.355
	n=20; J=20	10%	0.30	1.191	0.366	0.891	0.066	0.825
		30%	0.30	0.997	0.356	0.697	0.056	0.641
		50%	0.30	0.834	0.395	0.534	0.095	0.439
	n=8; J=50	10%	0.20	1.102	0.299	0.902	0.099	0.803
		30%	0.20	0.893	0.256	0.693	0.056	0.637
		50%	0.20	0.746	0.270	0.546	0.070	0.476
η_2	n=28; J=10	10%	0.40	0.413	0.409	0.013	0.009	0.004
		30%	0.40	0.420	0.410	0.020	0.010	0.010
		50%	0.40	0.391	0.390	-0.009	-0.010	-0.001
	n=20; J=20	10%	0.40	0.397	0.398	-0.003	-0.002	0.001
		30%	0.40	0.399	0.398	-0.001	-0.002	-0.001
		50%	0.40	0.398	0.391	-0.002	-0.009	-0.007
	n=8; J=50	10%	0.60	0.593	0.595	-0.007	-0.005	0.002
		30%	0.60	0.604	0.602	0.004	0.002	0.002
		50%	0.60	0.604	0.605	0.004	0.005	-0.001
ω_{22}	n=28; J=10	10%	0.10	0.085	0.043	-0.015	-0.057	-0.042
		30%	0.10	0.159	0.048	0.059	-0.052	0.007
		50%	0.10	0.199	0.062	0.099	-0.038	0.061
	n=20; J=20	10%	0.05	0.040	0.017	-0.010	-0.033	-0.023
		30%	0.05	0.091	0.018	0.041	-0.032	0.009
		50%	0.05	0.108	0.018	0.058	-0.032	0.026
	n=8; J=50	10%	0.15	0.096	0.061	-0.054	-0.089	-0.035
		30%	0.15	0.173	0.070	0.023	-0.080	-0.057
		50%	0.15	0.185	0.057	0.035	-0.093	-0.058
β	n=28; J=10	10%	0.75	0.651	0.654	-0.099	-0.096	0.003
		30%	0.75	0.667	0.673	-0.083	-0.077	0.006
		50%	0.75	0.687	0.693	-0.063	-0.057	0.006
	n=20; J=20	10%	0.75	0.641	0.645	-0.109	-0.105	0.004
		30%	0.75	0.656	0.665	-0.094	-0.085	0.009
		50%	0.75	0.677	0.687	-0.073	-0.063	0.010
	n=8; J=50	10%	0.75	0.643	0.648	-0.107	-0.102	0.005
		30%	0.75	0.654	0.666	-0.096	-0.084	0.012
		50%	0.75	0.672	0.687	-0.078	-0.063	0.015

Table 5: Selected mean parameter estimates from simulations where $\phi_{2j} = 2$

parameter of interest	study design	% of groups unconfounded	true value	mean "naive" estimate	mean "augmented" estimate	"naive" bias	"augmented" bias	bias reduction
$\tau_\cdot$	n=28; J=10	10%	0.40	1.174	0.474	0.774	0.074	0.700
		30%	0.40	0.989	0.468	0.589	0.068	0.521
		50%	0.40	0.788	0.499	0.388	0.099	0.289
	n=20; J=20	10%	0.30	1.071	0.374	0.771	0.074	0.697
		30%	0.30	0.926	0.370	0.626	0.070	0.556
		50%	0.30	0.764	0.394	0.464	0.094	0.370
	n=8; J=50	10%	0.20	0.991	0.306	0.791	0.106	0.685
		30%	0.20	0.800	0.257	0.600	0.057	0.543
		50%	0.20	0.678	0.272	0.478	0.072	0.406
η_2	n=28; J=10	10%	0.40	0.414	0.414	0.014	0.014	0.000
		30%	0.40	0.415	0.407	0.015	0.007	0.008
		50%	0.40	0.397	0.396	-0.003	-0.004	-0.001
	n=20; J=20	10%	0.40	0.401	0.402	0.001	0.002	-0.001
		30%	0.40	0.398	0.398	-0.002	-0.002	0.000
		50%	0.40	0.402	0.399	0.002	-0.001	0.001
	n=8; J=50	10%	0.60	0.598	0.599	-0.002	-0.001	0.001
		30%	0.60	0.603	0.601	0.003	0.001	0.002
		50%	0.60	0.605	0.606	0.005	0.006	-0.001
ω_{22}	n=28; J=10	10%	0.10	0.074	0.046	-0.026	-0.054	-0.028
		30%	0.10	0.119	0.051	0.019	-0.049	-0.030
		50%	0.10	0.146	0.060	0.046	-0.040	0.006
	n=20; J=20	10%	0.05	0.031	0.017	-0.019	-0.033	-0.014
		30%	0.05	0.062	0.017	0.012	-0.033	-0.021
		50%	0.05	0.075	0.018	0.025	-0.032	-0.007
	n=8; J=50	10%	0.15	0.087	0.063	-0.063	-0.087	-0.024
		30%	0.15	0.139	0.072	-0.011	-0.078	-0.067
		50%	0.15	0.144	0.060	-0.006	-0.090	-0.084
β	n=28; J=10	10%	0.75	0.641	0.642	-0.109	-0.108	0.001
		30%	0.75	0.660	0.663	-0.090	-0.087	0.003
		50%	0.75	0.684	0.688	-0.066	-0.062	0.004
	n=20; J=20	10%	0.75	0.631	0.634	-0.119	-0.116	0.003
		30%	0.75	0.651	0.657	-0.099	-0.093	0.006
		50%	0.75	0.674	0.681	-0.076	-0.069	0.007
	n=8; J=50	10%	0.75	0.634	0.637	-0.116	-0.113	0.003
		30%	0.75	0.651	0.658	-0.099	-0.092	0.007
		50%	0.75	0.673	0.683	-0.077	-0.067	0.010

Table 6: Selected mean parameter estimates from simulations where $\phi_{2j} = 1$

parameter of interest	study design	% of groups unconfounded	true value	mean "naive" estimate	mean "augmented" estimate	"naive" bias	"augmented" bias	bias reduction
$\tau_\cdot$	n=28; J=10	10%	0.40	0.957	0.535	0.557	0.135	0.422
		30%	0.40	0.735	0.416	0.335	0.016	0.319
		50%	0.40	0.729	0.549	0.329	0.149	0.180
	n=20; J=20	10%	0.30	0.833	0.393	0.533	0.093	0.440
		30%	0.30	0.686	0.325	0.386	0.025	0.361
		50%	0.30	0.621	0.388	0.321	0.088	0.233
	n=8; J=50	10%	0.20	0.732	0.295	0.532	0.095	0.437
		30%	0.20	0.603	0.261	0.403	0.061	0.342
		50%	0.20	0.528	0.268	0.328	0.068	0.260
η_2	n=28; J=10	10%	0.40	0.409	0.408	0.009	0.008	0.001
		30%	0.40	0.408	0.399	0.008	-0.001	0.007
		50%	0.40	0.400	0.400	0.000	0.000	0.000
	n=20; J=20	10%	0.40	0.398	0.399	-0.002	-0.001	0.001
		30%	0.40	0.400	0.400	0.000	0.000	0.000
		50%	0.40	0.403	0.399	0.003	-0.001	0.002
	n=8; J=50	10%	0.60	0.600	0.601	0.000	0.001	-0.001
		30%	0.60	0.602	0.601	0.002	0.001	0.001
		50%	0.60	0.610	0.610	0.010	0.010	0.000
ω_{22}	n=28; J=10	10%	0.10	0.055	0.048	-0.045	-0.052	-0.007
		30%	0.10	0.076	0.054	-0.024	-0.046	-0.022
		50%	0.10	0.083	0.059	-0.017	-0.041	-0.024
	n=20; J=20	10%	0.05	0.022	0.018	-0.028	-0.032	-0.004
		30%	0.05	0.031	0.018	-0.019	-0.032	-0.013
		50%	0.05	0.038	0.019	-0.012	-0.031	-0.019
	n=8; J=50	10%	0.15	0.075	0.067	-0.075	-0.083	-0.008
		30%	0.15	0.096	0.073	-0.054	-0.077	-0.023
		50%	0.15	0.089	0.061	-0.061	-0.089	-0.028
β	n=28; J=10	10%	0.75	0.658	0.658	-0.092	-0.092	0.000
		30%	0.75	0.676	0.677	-0.074	-0.073	0.001
		50%	0.75	0.697	0.698	-0.053	-0.052	0.001
	n=20; J=20	10%	0.75	0.649	0.649	-0.101	-0.101	0.000
		30%	0.75	0.668	0.669	-0.082	-0.081	0.001
		50%	0.75	0.689	0.691	-0.061	-0.059	0.002
	n=8; J=50	10%	0.75	0.652	0.652	-0.098	-0.098	0.000
		30%	0.75	0.671	0.673	-0.079	-0.077	0.002
		50%	0.75	0.691	0.694	-0.059	-0.056	0.003

References

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*(1). doi:10.18637/jss.v067.i01

Cinelli, C. & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 39–67. doi:10.1111/rssb.12348

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(1), 1–31. Retrieved from https://www.jstor.org/stable/2984718

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, *29*(2), 147–194. doi:10.1177/0049124100029002001

Frisch, R. & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, *1*(4), 387. doi:10.2307/1907330

Gelman, A. & Hill, J. (2007). *Data analysis using regression and hierarchical linear models*. Analytical Methods for Social Research. Cambridge, UK: Cambridge University Press.

Hazlett, C. & Wainstein, L. (2022). Understanding, choosing, and unifying multilevel and fixed effect approaches. *Political Analysis*, *30*(1), 46–65. doi:10.1017/pan.2020.41

Hernán, M. A. & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC. Retrieved from https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

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

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*(1), 4–29. Retrieved from https://www.jstor.org/stable/3211657

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

Kim, J. & Seltzer, M. (2007). Causal inference in multilevel settings in which selection processes vary across schools: (644002011-001). American Psychological Association. doi:10.1037/e644002011-001

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, *58*(304), 993–1010. doi:10.1080/01621459.1963.10480682

Lovell, M. C. (2008). A simple proof of the FWL theorem. *The Journal of Economic Education*, *39*(1), 88–91. doi:10.3200/JECE.39.1.88-91

Murnane, R. & Willett, J. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE Publications, Inc.

Raudenbush, S. W. & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387–401. doi:10.1037/1082-989X.6.4.387

Raudenbush, S. W. & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. doi:DOI:10.1037/1082-989X.5.2.199

Rickles, J. H. (2011). Using interviews to understand the assignment mechanism in a nonexperimental study: The case of eighth grade algebra. *Evaluation Review*, *35*(5), 490–522. doi:10.1177/0193841X11428644

Rosenbaum, P. R. & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society*. B, *45*(2), 212–218. Retrieved from https://www.jstor.org/stable/2345524

Rosenbaum, P. R. & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. Retrieved from https://doi.org/10.2307/2335942

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi:10.1037/h0037350

Rubin, D. B. (1986). Statistics and causal inference: comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*(396), 961–962. doi:10.2307/2289065

Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal: Promoting communications on statistics and Stata*, *13*(1), 65–76. doi:10.1177/1536867X1301300105

Seltzer, M., Kim, J., & Frank, K. A. (2006). *Studying the sensitivity of inferences to possible unmeasured confounding variables in multisite evaluations (CSE Technical Report 701)*. Retrieved from https://files.eric.ed.gov/fulltext/ED495851.pdf

US Department of Education; Institute of Education Sciences. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Council for Excellence in Government. Washington, DC. doi:10.1037/e370412004-001