Observational Studies 4 (2018) 111-149 Submitted 10/17; Published 3/18

Potential for Bias Inflation with Grouped Data: A Comparison of Estimators and a Sensitivity Analysis Strategy

Marc A. Scott **marc.scott@nyu.edu**
New York University, New York, USA
Ronli Diakow **rdiakow@gmail.com**
NYC Department of Education, Brooklyn, NY, USA
Jennifer L. Hill **jennifer.hill@nyu.edu**
New York University, New York, USA
Joel A. Middleton **joel.middleton@gmail.com**
University of California, Berkeley, USA

Abstract

We are concerned with the unbiased estimation of a treatment effect in the context of non-experimental studies with grouped or multilevel data. When analyzing such data with this goal, practitioners typically include as many predictors (controls) as possible, in an attempt to satisfy ignorability of the treatment assignment. In the multilevel setting with two levels, there are two classes of potential confounders that one must consider, and attempts to satisfy ignorability conditional on just one set would lead to a different treatment effect estimator than attempts to satisfy the other (or both). The three estimators considered in this paper are so-called "within," "between" and OLS estimators. We generate bounds on the potential differences in bias for these competing estimators to inform model selection. Our approach relies on a parametric model for grouped data and omitted confounders and establishes a framework for sensitivity analysis in the two-level modeling context. The method relies on information obtained from parameters estimated under a variety of multilevel model specifications. We characterize the strength of the confounding and corresponding bias using easily interpretable parameters and graphical displays. We apply this approach to data from a multinational educational evaluation study. We demonstrate the extent to which different treatment effect estimators may be robust to potential unobserved individual- and group-level confounding.

Keywords: Multilevel models, Sensitivity analysis, Bias amplification

1. Introduction

In non-experimental settings researchers face considerable challenges in the identification of treatment effects. Chief among these is the need to control for all confounders – the covariates that predict both the treatment and outcome. Failure to include all confounders can result in bias (Angrist and Pischke 2009; Greene 2003; Wooldridge 2010). The desire to satisfy this requirement, sometimes referred to as the ignorability of treatment assignment (Rubin 1978), can lead researchers to include as many covariates as possible in the model and sometimes leads to the false intuition that each confounder (or set of confounders) added will reduce the bias of the estimate.

However, recent work on bias amplification (Pearl 2012; Wooldridge 2010; Schisterman et al. 2009) suggests that under certain conditions, including particular types of covariates actually increases the bias due to omitted confounders, making the variable selection process more complex. In related work, Middleton et al. (2016) and Clark and Linzer (2012) show that bias amplification can be a problem in so-called "fixed effects" models (models that yield a "within" estimator), which are often used when the data exhibit group, or multilevel, structure. Thus, the decision to include group fixed effects is not without risk.

While it might seem counterintuitive, ignoring group structure can sometimes lead to estimators with less absolute bias as compared to estimators that account for group structure. In particular this arises because in many studies, ignorability is unlikely to be satisfied, and the bias of any estimator depends on the properties of any remaining, uncontrolled confounders. One approach to understanding the potential impact of unobserved confounders is a sensitivity analysis (see, for example, Rosenbaum and Rubin 1983; Rosenbaum 1987; Greenland 1996; Gastwirth et al. 1998; Lin et al. 1998; Robins et al. 2000; Rosenbaum 2002; Imbens 2003; McCandless et al. 2007; Rosenbaum 2010; VanderWeele and Arah 2011; Harada 2013; Carnegie et al. 2016; Dorie et al. 2016). Beginning with Cornfield et al. (1959) and made more precise in Rosenbaum and Rubin (1983), a sensitivity analysis relates the uncontrolled confounders to treatment and outcome processes, allowing the researcher to evaluate "what if" scenarios, which establish the strength of the two aforementioned relationships needed to reduce or eliminate a treatment effect. Imbens (2003) built on the framework of Rosenbaum and Rubin (1983) to establish a model-based sensitivity analysis for a continuous outcome, binary treatment and binary confounder.

Building on model-based approaches, Middleton et al. (2016) show that the common assumption that including indicators for groups "might help and never hurts" is wrong, contextualizing the roles of bias amplification and unmasking. In this paper, we take this idea further, and show that the components of variance, estimable in multilevel models, provide additional information about how much the improper choice of estimator or variables in the model could hurt. We are able to bound the bias associated with different estimators in a framework that merges ideas from sensitivity analysis with information gleaned from multilevel model parameter estimates (Gelman and Hill 2007; Singer and Willett 2003). Somewhat surprisingly, a well-known variance ratio called the intraclass correlation coefficient (ICC; Shrout and Fleiss 1979) informs the practitioner's choice of estimation model and confidence in the findings.

The recommendation to include known and measurable confounders as controls is based on the desire to satisfy the assumption that all confounders have been measured (also known as the ignorability assumption). When the treatment effect is not very sensitive to the addition or subtraction of individual covariates it is tempting to interpret this as evidence that the threat due to unobserved confounding has been substantially reduced or eliminated. For indeed the opposite holds – if one notices a change in the treatment effect upon removing a covariate then said effect is sensitive to model specification therefore it seems more likely that confounding remains. However when treatment effects are relatively stable it is also possible that the included covariates simply have little effect on the treatment estimate. Our methods offer some protection against drawing incorrect conclusions from either scenario by incorporating variance components in the sensitivity analysis. The precise way in which our methods relate to these two scenarios is described in Section 5.

In this paper, we develop a sensitivity analysis framework for the multilevel setting by extending the work of Imbens (2003), Carnegie et al. (2016) and Middleton et al. (2016) via the inclusion of group-level controls and confounders. Initially, we use information contained in the multilevel structure to compare the asymptotic bias of three different treatment effect estimators: the within, between and OLS estimators (see Greene 2003; Angrist and Pischke 2009). This yields a sensitivity analysis in which competing estimators' performance can be compared conditional on posited levels of confounding. We take this an important step further: by capitalizing on the different bias properties of each estimator, we identify *limits* on the potential impact of remaining unobserved confounders. While the omitted confounders are by definition unobservable, information obtained by exploiting properties of multilevel models reveals a "feasibility set" for the omitted confounders, greatly reducing the researcher's uncertainty about the potential for bias and greatly refining the sensitivity analysis.[1]

The organization of the paper is as follows. In section 2, we propose a data generating process that captures the key elements of confounding in the multilevel context, in which the coefficient on a single predictor represents the targeted treatment effect, usually in an observational study or broken randomized experiment. In section 3, we derive the asymptotic bias for several estimators and examine their difference in absolute bias. From

[1] Another way to understand this is that the covariance structure inherent in grouped data provides information about feasible mean structure. This post hoc evaluation of potential confounding is similar in spirit to post hoc power analysis.

these we can establish direct comparisons in a confounding space derived from our model parameters, and importantly, from some variance components and parameters that may be estimated unbiasedly. Using identifiable parameters, we can infer some of the properties of the confounding space, and specifically whether and when so-called within estimators are more or less biased than OLS estimators, potentially due to amplification (Middleton et al. 2016; Pearl 2012). We then show that by fitting sequential multilevel models (MLMs), we can learn what subset of the confounding space is consistent with our data and underlying model assumptions. This is a key contribution that characterizes and restricts the set of possible unobserved confounders dramatically and can greatly improve one's confidence about estimated treatment effects. In section 4, we apply this approach on real data, presenting a multilevel sensitivity analysis across a range of confounding situations routinely encountered. We conclude and discuss further work in section 5.

2. Multilevel framework

2.1 Data generating process

Our notation may be described as the composite form for nested multilevel structure (Scott et al. 2013; Singer and Willett 2003; Singer 1998). We use subscript i to index the individual (sometimes known as varying at level 1) and j for the group index (level 2). Examples include students "nested" in schools, patients in hospitals, or individuals in neighborhoods.

In addition to the usual model formulation for studies with a univariate outcome Y and single treatment Z, we follow the tradition of Lin et al. (1998) and Imbens (2003), among others, by including in our model an unobserved confounder as well as sensitivity parameters, which establish the magnitude of the confounding. The parameters link the confounder to both the treatment and outcome and may be operationalized in many different ways. Lin et al. (1998) posit a confounder U with two coefficients, γ_0 and γ_1 that may differ for control and treatment, respectively. Imbens (2003) posits a binary confounder U with parameters α and δ calibrating the impact on treatment and outcome, respectively. Using some distributional assumptions, both authors integrate out the confounder, leaving an adjusted treatment effect in terms of sensitivity parameters, often in the form of a product of coefficients.

We build on this sensitivity framework by positing an additional layer of confounding that operates exclusively at the group level. We name that confounder V, and specify how that confounder links to treatment and response at the group level. We are able to add this second confounder because we have more information in the multilevel structure, which can be directly modeled, and the structure imposed by the model constrains the feasible confounding at the individual and group level.

The sensitivity analysis framework is developed by specifying unobserved (confounding) covariates to the presumed data generating process (DGP). We include U_{ij}, which varies only within groups as well as V_j, which varies only between groups. In the educational setting, with students nested in schools and a skill assessment as outcome, an example of an unmeasured individual-level confounder might be student motivation. At the school level, information about administrators' adherence to system-wide guidelines could be an unmeasured confounder. This is a natural multilevel extension of Imbens (2003) which posits a single unobserved confounder (the vast majority of sensitivity analysis frameworks posit only one unobserved confounder). It is also a practical way to establish more realistic relationships between treatment, predictors, random effects and errors as suggested in Hill (2013). As is quite standard in multilevel models, we include both group varying errors α_j (at level 2) and subject-specific errors ϵ_{ij} (at level 1). Following Imbens (2003), we assume that U and V are independent of each other, the observed covariates, and the additional error terms α and ϵ. This is reasonable if we conceive of U and V as representing the portion of unobserved confounding that is orthogonal to what we have observed. Henceforth we refer to the model for Y conditional on everything else (Equation 1) as the *response surface* and the model for Z conditional on everything but the outcome (Equation 2) as the *assignment mechanism.*

The equations and assumptions that specify our data generating process (DGP) are summarized here (see also Appendix A):

$$Y_{ij} = \tau Z_{ij} + X_{ij}\beta^y + \zeta^y U_{ij} + W_j\gamma^y + \delta^y V_j + \alpha_j^y + \epsilon_{ij}^y \tag{1}$$

$$Z_{ij} = X_{ij}\beta^z + \zeta^z U_{ij} + W_j\gamma^z + \delta^z V_j + \alpha_j^z + \epsilon_{ij}^z \tag{2}$$

$$\alpha^y \sim N(0, \psi^y), \quad \alpha^z \sim N(0, \psi^z), \quad \text{cov}(\alpha^y, \alpha^z) = 0 \tag{3}$$

$$\epsilon^y \sim N(0, \sigma_y^2), \quad \epsilon^z \sim N(0, \sigma_z^2), \quad \text{cov}(\epsilon^y, \epsilon^z) = 0 \tag{4}$$

$$\text{cov}(\alpha^y, \epsilon^y) = 0, \;\; \text{cov}(\alpha^z, \epsilon^z) = 0, \;\; \text{cov}(\alpha^y, \epsilon^z) = 0, \;\; \text{cov}(\alpha^z, \epsilon^y) = 0$$

$$U_{ij} \sim N(0, \sigma_u^2), \quad V_j \sim N(0, \sigma_v^2), \quad \text{cov}(U, V) = 0 \tag{5}$$

$$\text{cov}(\Upsilon, \epsilon^y) = 0, \;\; \text{cov}(\Upsilon, \epsilon^z) = 0, \;\; \text{cov}(\Upsilon, \alpha^y) = 0, \;\; \text{cov}(\Upsilon, \alpha^z) = 0, \;\; \Upsilon \in \{U, V\}. \tag{6}$$

Here Y_{ij} represents our outcome for subject i in group j, and Z_{ij} is the corresponding subject-specific treatment, which we specify as continuous. Observed covariates divide naturally into two types: X_{ij}, which vary only within groups (e.g., the student's sex), and W_j, which vary only between groups (e.g., average teacher salary in the school). Note that any observed covariate can be split into these two parts via group mean centering (Enders and Tofighi 2007; Neuhaus and Kalbfleisch 1998; Raudenbush 2009; Townsend et al. 2013), and we do this automatically in software.[2]

Note that while group effects in the DGP are assumed to follow a normal distribution, this is not overly restrictive if one allows for large values of the variance parameters (ψ^y, ψ^z). This is consistent with Gelman and Hill (2007), who argue that unrestricted group effects are simply those derived from a distribution with infinite variance.[3] Moreover, our results follow from orthogonality, not normality, assumptions. As we will see, estimation often relies on simpler assumptions that intentionally only partially align with the DGP.

2.2 Estimating a treatment effect τ with grouped data

We wish to estimate τ, the treatment effect, using the model specified for Y, but we will never observe the U or V leading to individual- and group-level confounding. Individual-level confounding violates the strict exogeneity assumption of "fixed effect" models (within-estimator models). Group-level confounding violates the so-called "random effects assumption" of the between estimator model. The OLS estimator is biased when either assumption is violated. See Greene (2003) for additional details.

Although each individual estimator will likely yield biased estimates, we will show how the properties of each reveal important information that characterizes the unobserved confounding. We first define these estimators of τ in the context of grouped data.

- $\hat{\tau}_W$ (within estimator): This can be implemented in many ways, one of which is to group-mean center all variables (including Z) and then regress the outcome on these transformed predictors, reporting the coefficient on centered Z. The resulting estimate $\hat{\tau}_W$ is equivalent to the so-called fixed effects estimate, $\hat{\tau}_{FE}$, obtained by OLS estimation on a model with indicators for each group (see Townsend et al. 2013, for further details.). Translating the exogeneity assumption to the notation of our DGP, this estimator assumes $E(Z_{ij}(\zeta^y U_{ij} + \epsilon^y_{ij})) = 0$, which will only hold when either $\zeta^y = 0$ or $\zeta^z = 0$.

- $\hat{\tau}_B$ (between estimator): This can also be implemented by group-mean centering all predictors (including Z), adding the group means to the regression, and then reporting the coefficient on the group mean of Z (in practice, estimation of between and within effects can be made simultaneously). The random effects assumption translates to

[2] Available at: https://github.com/priism-center/pre-CRAN/tree/master/compBiasMLM. Relately we can also conceive of U and V as distinct parts of the same unobserved covariate – the part that works at the individual level and the part that works at the group level.

[3] We implicitly use this unrestricted version with indicator variables for group effects in so-called fixed effects models.

$E(Z_{ij}(\delta^y V_j + \alpha_j^y)) = 0$, which will only hold when either $\delta^y = 0$ or $\delta^z = 0$. For several different approaches and discussion, see Allison (2006); Neuhaus and Kalbfleisch (1998); Griliches and Hausman (1986).

- $\hat{\tau}_{OLS}$ (OLS estimator): Here, we estimate the (pooled; see Gelman and Hill 2007) treatment effect without group-mean centering Z. The ordinary least squares estimator ignores the group structure; we demonstrate the utility of this estimator in what follows.

We do not include the generalized least squares (GLS) estimator, also known as the random effects estimator, in our discussion. It is a weighted-average of the within- and between-estimators, but the weights introduce an additional parameter that we are unable to unbiasedly estimate, and as such, we cannot easily "learn" from this model. See Townsend et al. (2013) and Appendix B for some discussion. Causal researchers may note that these estimators are sometimes use to target *different* estimands. However, in this case we assume a DGP with constant treatment effects, therefore each of these estimators is estimating the same estimand, τ.

2.3 Bias

We know that U and V are unobserved and introduce omitted variables bias when we attempt to estimate the treatment effect τ. However, asymptotically, the within-estimator ($\hat{\tau}_W$) is robust to bias introduced by V and the between estimator ($\hat{\tau}_B$) is robust to bias introduced by U. The OLS estimator ($\hat{\tau}_{OLS}$) is robust to neither, but provides a potential estimation strategy which offers a different hedge against either form of bias through an implicit down-weighting process. In Appendix B, we provide more derivation details for the asymptotic omitted variables bias under the various estimators $\hat{\tau}_W$, $\hat{\tau}_B$ and $\hat{\tau}_{OLS}$. We now explore how these bias formulas may be exploited to bound the bias in our estimators.

Fixing the number of groups J and letting the sample size N go to infinity, the bias of the fixed effects (within) estimator is

$$\text{Bias}[\hat{\tau}_W] = \frac{\zeta^y \zeta^z \sigma_u^2}{\sigma_z^2 + (\zeta^z)^2 \sigma_u^2} \tag{7}$$

The asymptotic bias for the between estimator, holding group size N_J constant (and letting the number of groups $J = N/N_J \to \infty$) is:

$$\text{Bias}[\hat{\tau}_B] = \frac{\zeta^y \zeta^z \sigma_u^2 / N_J + \delta^y \delta^z \sigma_v^2}{(\sigma_z^2 + (\zeta^z)^2 \sigma_u^2)/N_J + \psi^z + (\delta^z)^2 \sigma_v^2} \tag{8}$$

Asymptotically, the bias is:

$$\text{Bias}[\hat{\tau}_B] = \frac{\delta^y \delta^z \sigma_v^2}{\psi^z + (\delta^z)^2 \sigma_v^2} \tag{9}$$

We see that, asymptotically, the within estimator is unbiased with respect to group confounding (δ^z and δ^y do not appear in the expression for the bias). On the other hand for large group size the between estimator is unbiased with respect to individual confounding (ζ^z and ζ^y do not appear in the expression for the bias). In practice however, group size may be relatively small, in which case equation (8) is more accurate and the bias for the between estimator depends on both individual and group-level confounding.

For the OLS estimator, asymptotically, the bias is:

$$\text{Bias}[\hat{\tau}_{OLS}] = \frac{\zeta^y \zeta^z \sigma_u^2 + \delta^y \delta^z \sigma_v^2}{\sigma_z^2 + (\zeta^z)^2 \sigma_u^2 + \psi^z + (\delta^z)^2 \sigma_v^2} \tag{10}$$

Clearly, it is a function of both individual and group-level confounding. However, the σ_z^2 and ψ^z terms in the denominator offer an additional down-weighting of the bias induced by the terms in the numerator, as compared to the asymptotic bias formulas for the between and within estimators. These reveal the latter's vulnerability: should the variance terms in the denominator be relatively small, bias-inflation (of the numerator) is likely to occur. See Middleton et al. (2016); Pearl (2012); Clark and Linzer (2012) for further discussion.

2.3.1 Comparison of Biases: "danger" zones

These bias equations demonstrate that no single estimator is uniformly preferable; bias is situation dependent. We can "map" the zones in which within, between or OLS estimators would be preferred, conditional on assumptions about sensitivity parameter product terms $\zeta^{yz} = \zeta^y \zeta^z$ and $\delta^{yz} = \delta^y \delta^z$. We define the confounding space to be the coordinate plane with x-axis ζ^{yz} and y-axis δ^{yz}. As we move away from the origin, in which there is no confounding, the degree of between and within group confounding grows. Each of our estimators will be more or less biased under these differing conditions. This form of sensitivity analysis, in which properties of an unobserved confounder are mapped to bias, were developed in Rosenbaum (2002) and Rosenbaum and Rubin (1983), but our approach is based more closely on Imbens (2003) and the generalization thereof by Carnegie et al. (2016) and Middleton et al. (2016). The extension to multilevel forms of confounding is underexplored in the literature.

For this discussion, we can assume $\sigma_u^2 = \sigma_v^2 = 1$ without loss of generality. We will primarily use version (9) of the between estimator in which group size is large for model development, but may have the option of reverting to the finite group size results with real data, where the group size might not be considered large. To proceed we define two terms (see also Appendix A):

$$\begin{aligned} c_W &= \sigma_z^2 + (\zeta^z)^2 \\ c_B &= \psi^z + (\delta^z)^2 \end{aligned} \tag{11}$$

Crucially, these sums are estimable unbiasedly from a model based on the DGP for Z; they correspond to the within and between group level variance components, respectively (the two components in each sum are not separately identifiable). Moreover we can benchmark the magnitude of these terms because a common measure of between versus within variation, $ICC_Z = c_B/(c_B + c_W)$, is a function of these terms and roughly characterizes different scenarios that we might expect in real data.

We first consider the case in which $\text{Bias}[\hat{\tau}_B] \leq \text{Bias}[\hat{\tau}_W]$. This inequality translates into an observable condition:

$$\left|\frac{\delta^y\delta^z}{c_B}\right| \leq \left|\frac{\zeta^y\zeta^z}{c_W}\right| \iff \left|\frac{\delta^y\delta^z}{\zeta^y\zeta^z}\right| \leq \frac{c_B}{c_W} \tag{12}$$

When the magnitude of between-to-within group confounding (as captured in our parameters δ and ζ) is smaller than the estimable ratio of between-to-within group variance in the treatment model (net of predictors), the between estimator will exhibit less absolute bias than the within estimator.[4] While we do not know the values of $\zeta^{yz} = \zeta^y\zeta^z$ or $\delta^{yz} = \delta^y\delta^z$, we can identify portions of the confounding space (defined by these parameters) for which one estimator is less biased (asymptotically) than another. The inequality establishes a set of lines in the coordinate plane, intersecting at the origin, forming a partition into four regions, where each side of the partition represents superior performance of one estimator over another. The magnitude of the ratio drives the magnitude of the difference.

While there are several pairwise comparisons of methods that one could make, this paper focuses on the within versus OLS estimator comparison. This mirrors the decision to control for group effects using the so-called fixed effects estimator or not, as discussed in Middleton et al. (2016) and Clark and Linzer (2012). In order for the bias in OLS to be less than or equal to the bias in the within estimator, we require:

$$\left|\frac{\zeta^y\zeta^z + \delta^y\delta^z}{c_W + c_B}\right| \leq \left|\frac{\zeta^y\zeta^z}{c_W}\right| \iff \left|\frac{\zeta^y\zeta^z + \delta^y\delta^z}{\zeta^y\zeta^z}\right| \leq \frac{c_W + c_B}{c_W} \iff \left|1 + \frac{\delta^y\delta^z}{\zeta^y\zeta^z}\right| \leq 1 + \frac{c_B}{c_W} \tag{13}$$

Note that $1 + c_B/c_W = 1/(1 - ICC_Z)$.

Figure 1 displays the difference in bias (OLS minus within) that would be incurred in a set-up corresponding to our DGP under two different assumptions about the ICC for the assignment mechanism, denoted ICC_Z. Thus the blue regions correspond to products of

[4] Going forward, we use the term bias and absolute bias to mean the latter, for convenience of exposition.

the sensitivity parameters when the OLS model leads to worse results than the fixed effects model. The pink regions correspond to combinations of the sensitivity parameters when the OLS model leads to better results than the fixed effects model. In order to make a fair comparison, the quantity $c_W + c_B$ was held constant for the two ICC_Z scenarios. As in the prior discussion, we assume the variance of our confounders to be one, as these set the scale of ζ^{yz} and δ^{yz}.

When ICC_Z is moderate at 0.50 (left panel, figure 1), OLS provides a reasonable hedge against fairly large bias. Referring to (13), when $c_W = c_B$, we see that the line $\zeta^{yz} = \delta^{yz}$ (indicated on plot) defines scenarios in which we are indifferent as to whether or not to control for group effects. This suggests that while one will always be indifferent between the within-estimator and the OLS estimator at the origin (when there is no confounding), the line that indicates which is preferred will only fall on the 45° line when ICC_Z=0.50. On either side of the line, in the upper right and lower left quadrants (I and III), half of the confounding space (blue) suggests that including fixed effects is the best solution and half the space (pink) suggests that excluding fixed effects is the best solution. However, when $sign(\zeta^{yz}) \neq sign(\delta^{yz})$, which is true for the upper left and lower right quadrants (II and IV), very few scenarios would recommend including group (fixed) effects. This is due to the strictly positive signs of c_W and c_B.

In the right panel of Figure 1, when ICC_Z is smaller at 0.25, the within estimator has lower absolute bias than OLS in a larger portion of the plane, particularly in quadrants I and III. On average, however, the bias difference is smaller than the left panel (indicated by lighter shades of both colors).

In general, a practitioner can estimate c_B and c_W unbiasedly through variance component estimates from multilevel models for Z. From these, one can construct the absolute bias difference plot and assess whether inclusion of group effects is prudent or whether caution is advised. In real data analyses, we do not know in which portion of the confounding space we reside, but our substantive knowledge about a research question may help us assign probabilities to different regions of the plot. Moreover, as we shall demonstrate, we can constrain the possibilities further.

This "confounding space" framework suggests that one can and should evaluate the potential implications of unobserved confounders *for the particular problem.* For example, in order for the within-estimator to be unbiased, one must assume that unobserved confounding is on the x-axis of our confounding space; alternatively, one would have to assume that the ICC for Z is very small, forcing the bias difference also to be small.[5] We examined several

[5] This only suggests that neither method is superior; they may both be highly biased, of course, should confounding be strong or susceptible to amplification.

papers that document the ICC for different variables and domains[6] and found that an ICC of about 0.25 was quite common in fields such as education, social and behavioral studies, while it tends to be a bit lower in biological and health studies (Altonji and Mansfield 2011; Donner and Koval 1980; Hedges and Hedberg 2007; Hedges et al. 2007; Thompson et al. 2012). The ICC tends to reduce to about half of that size with full controls in a model. That the ICC is rarely zero suggests that the estimation method (and thus model specification) matters; if one estimates the ICC and it is near 0.50, then our framework suggests that OLS has lower absolute bias over a wider range of the confounding space (for any square centered at the origin), whereas for ICC below 0.25, the within estimator has less absolute bias over a greater range of scenarios. If one wants to make a safer bet on the treatment effect, the ICC and the resulting bias maps allow a more informed choice.

To reiterate, there is information in the data, revealed via multilevel modeling, that can inform the analyst as to which scenarios lead to greater absolute bias and what is the potential size of that bias difference. To assume that confounding is very close to the origin and that the ICC for Z is small is imprudent and unnecessary; the former may be overly optimistic and the latter may be directly evaluated *a posteriori.*

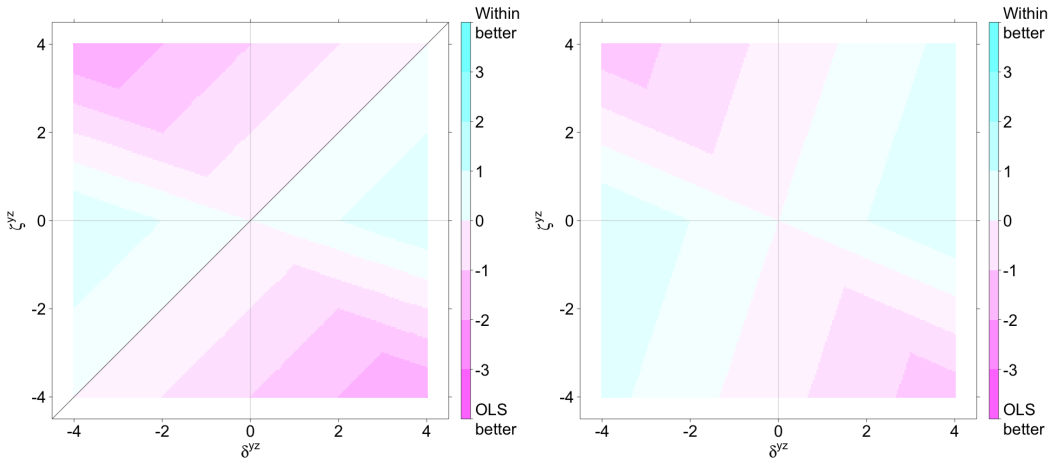

Figure 1: Difference in absolute bias when $ICC_Z = 0.50$ and 0.25, respectively: $|\text{bias}[\hat{\tau}_{OLS}]| - |\text{bias}[\hat{\tau}_W]|$ estimators for a range of within and between group confounding. The blue regions correspond to combinations of the sensitivity parameter products when the OLS model leads to worse results than the fixed effects model; the pink regions correspond to the opposite situation.

[6] The goal of these papers was primarily to document ICC for use in power studies; thus, these estimates are an attempt to consolidate substantive knowledge about what to commonly expect in real studies.

3. Bounds on parameters from models of Y and Z

The above discussion emphasizes estimable components of variance from the assignment mechanism and makes no attempt to constrain the sensitivity parameters. This is consistent with the view that such parameters are given, or ranges of them are explored in a sensitivity analysis, which is true. We now shift the focus to implicit constraints on these parameters imposed by the data and DGP. As such, we subsequently refer to $\{\zeta^y, \zeta^z, \delta^y, \delta^z\}$ as *confounding parameters*. Given our DGP, we have clear expressions for the bias for each of our estimators as a function of components of the response surface. Thus, if two models estimate the same effect but with different biases, then the difference in these biased estimates provides an estimate of a bias difference relationship. We now develop models and expressions for bias differences to exploit the fact that the functional form of this bias difference is known asymptotically, conditional on our confounding parameters.

It is common in the multilevel setting to estimate models with fewer predictors first and then include additional predictors, documenting changes in the proportion of variance components explained. A set of nested sequential models explain different components of variance in the same way that R^2 may increase as one adds predictors to a regression model. Under our causal inferential framework, estimates of the treatment effect are biased whenever there are omitted confounders, and it would seem that this MLM sequential model approach induces additional bias through the intentional omission of predictors. However, under the DGP given by Equations (1)-(6), we precisely know the form of that bias. We even have consistent estimators of the bias difference using multiple model specifications. A key insight is that these bias differences are consistent when confounders are defined as per our DGP. We essentially begin our analysis in a state of ignorance, in which confounding, as parameterized by ζ^{yz} and δ^{yz} can take on any value in $\mathbb{R}^2$, but through multiple multilevel model estimates, we will learn that only a limited subspace of $\mathbb{R}^2$ is consistent with the data and our understanding of its multilevel structure. This approach allows us to partially identify the characteristics of the confounders (see Gustafson 2015, for discussion of partial identification).

3.1 Multistage models for Y

We build Models M_0^Y and M_1^Y for Y. The first model contains no predictors. The second model contains group-mean centered individual-level predictors, denoted $X_{ij} - \bar{X}_{\cdot j}$, the corresponding group means, $\bar{X}_{\cdot j}$, as well as any other group level predictors W_j. This process will yield a set of variance components that will change from model M_0^Y to model M_1^Y, and we will use these to identify plausible values for the confounding parameters. Both models contain the terms $(Z_{ij} - \bar{Z}_{\cdot j})$ and $\bar{Z}_{\cdot j}$, allowing us to estimate the within and between treatment effect estimators simultaneously, under two different model specifications (and thus bias). The orthogonality of terms in this hybrid model specification (Neuhaus and Kalbfleisch 1998) ensures that predictor effects are isolated to specific components of variance as well.

We name the parameters in M_0^Y and M_1^Y slightly differently to distinguish them from prior MLM equations and to emphasize the smaller number of identifiable parameters in these models:

$$Y_{ij} = \beta_{00}^y + \tau_{W0}(Z_{ij} - \bar{Z}_{\cdot j}) + \tau_{B0}\bar{Z}_{\cdot j} + \alpha_{j0}^y + \epsilon_{ij0}^y \ \ [M_0^Y] \tag{14}$$

$$Y_{ij} = \beta_{01}^y + (X_{ij} - \bar{X}_{\cdot j})\beta_{W1}^y + \bar{X}_{\cdot j}\beta_{B1}^y + \tau_{W1}(Z_{ij} - \bar{Z}_{\cdot j}) + \tau_{B1}\bar{Z}_{\cdot j} + W_j\gamma_1^y + \alpha_{j1}^y + \epsilon_{ij1}^y \ [M_1^Y] \tag{15}$$

with $\alpha_{j0}^y \sim N(0, \sigma_{\alpha_0^y}^2)$, $\alpha_{j1}^y \sim N(0, \sigma_{\alpha_1^y}^2)$, $\epsilon_{ij0}^y \sim N(0, \sigma_{\epsilon_0^y}^2)$, and $\epsilon_{ij1}^y \sim N(0, \sigma_{\epsilon_1^y}^2)$, all mutually independent, where τ_{W0}, τ_{W1} are the corresponding within-group treatment effects and τ_{B0}, τ_{B1} are the corresponding between-group treatment effects. These are models, not DGPs, so under our true DGP, $\tau_{W0} \neq \tau_{W1} \neq \tau_{B0} \neq \tau_{B1} \neq \tau$, due to (intentional) model misspecification and omitted confounders U and V. We will exploit the fact that estimates of $\{\tau_{W0}, \tau_{W1}, \tau_{B0}, \tau_{B1}\}$ will vary depending on the magnitude, direction and type of confounding. Note as well that in our estimated models, we separate within- and between-group predictors and use group-mean centering to enforce this. Restricting the predictors to vary strictly within or between groups allows us to attribute the change in variance components from M_0^Y to M_1^Y as strictly derived from one or the other type of predictor as well as expect the variances to be non-increasing as such predictors are added. Thus, going forward, all within-group predictors are given by $X^* = X_{ij} - \bar{X}_{\cdot j}$ and all between-group predictors by $W^* = \{W_j, \bar{X}_{\cdot j}\}$ to emphasize the separation, which is central to the analysis.

We first fit model M_0^Y, which has intentionally omitted predictors X^* and W^*. The parameter estimate $\hat{\tau}_{W0}$ is equivalent to a within-group estimator (Allison 2006; Neuhaus and Kalbfleisch 1998). The exact form of the bias term (due to *all* omitted predictors) is given in Appendix B expression (B.11), but we repeat it here for exposition purposes under the newly introduced notation, and note that in the appendix, X represents individual-level predictors and W represents group-level predictors, so we are just changing the notation slightly to be consistent with that derivation.

In what follows, we use the asymptotic expectations for the bias expressions *without indicating the limit notationally*, for convenience; without loss of generality, continue to assume $\sigma_u^2 = \sigma_v^2 = 1$. Note as well that any parameters listed are *true* values from the DGP equations (1)-(6), not estimated in our models – the latter parameters have subscripts to differentiate them.

$$\text{Bias}[\hat{\tau}_{W0}] = \frac{\zeta^y\zeta^z + \beta^{y\prime}V(X^*)\beta^z}{\sigma_z^2 + (\zeta^z)^2 + \beta^{z\prime}V(X^*)\beta^z} \tag{16}$$

Fit model M_1^Y. Let $\hat{\tau}_{W1}$ be the estimate of τ from a within-group estimator for this model. The exact form of the bias term is given in Appendix B expression (B.5), but we repeat it here, as a function of the confounding parameters:

$$\text{Bias}[\hat{\tau}_{W1}] \quad = \quad \frac{\zeta^y \zeta^z}{\sigma_z^2 + (\zeta^z)^2} \tag{17}$$

We now write the estimators in terms of the true values of the treatment effect and the bias: $\hat{\tau}_{W0} = \tau + \text{Bias}[\hat{\tau}_{W0}]$, where τ is the true treatment effect, and $\text{Bias}[\hat{\tau}_{W0}]$ is the bias associated with model M_0^Y, while $\hat{\tau}_{W1} = \tau + \text{Bias}[\hat{\tau}_{W1}]$. Then $\hat{\tau}_{W0} - \hat{\tau}_{W1} = \text{Bias}[\hat{\tau}_{W0}] - \text{Bias}[\hat{\tau}_{W1}]$ (the true τ's cancel). We apply this to the derived expressions for asymptotic bias, and find (as a function of finite common group size N_j):

$$\hat{\tau}_{W0} - \hat{\tau}_{W1} = \text{Bias}[\hat{\tau}_{W0}] - \text{Bias}[\hat{\tau}_{W1}] = \frac{\zeta^y \zeta^z + \beta^{y\prime} V(X^*)\beta^z}{\sigma_z^2 + (\zeta^z)^2 + \beta^{z\prime} V(X^*)\beta^z} - \frac{\zeta^y \zeta^z}{\sigma_z^2 + (\zeta^z)^2} \tag{18}$$

Applying the same procedure, deriving the bias difference for the between estimators of τ for models M_0^Y and M_1^Y, yields

$$\begin{aligned} \hat{\tau}_{B0} - \hat{\tau}_{B1} \quad &= \quad \text{Bias}[\hat{\tau}_{B0}] - \text{Bias}[\hat{\tau}_{B1}] \\ &= \quad \frac{(\zeta^y \zeta^z + \beta^{y\prime} V(X^*)\beta^z)/N_j + \delta^y \delta^z + \gamma^{y\prime} V(W^*)\gamma^z}{(\sigma_z^2 + (\zeta^z)^2 + \beta^{z\prime} V(X^*)\beta^z)/N_j + \psi^z + (\delta^z)^2 + \gamma^{z\prime} V(W^*)\gamma^z} \\ &\quad - \frac{\zeta^y \zeta^z / N_j + \delta^y \delta^z}{(\sigma_z^2 + (\zeta^z)^2)/N_j + \psi^z + (\delta^z)^2}, \end{aligned} \tag{19}$$

which has a simpler form when group size is very large.

3.2 Multistage models for Z

Now fit the corresponding models (without and with controls, respectively) for Z:

$$Z_{ij} = \beta_{00}^z + \alpha_{j0}^z + \epsilon_{ij0}^z \;\; [M_0^Z] \tag{20}$$

$$Z_{ij} = \beta_{01}^z + (X_{ij} - \bar{X}_{\cdot j})\beta_{W1}^z + \bar{X}_{\cdot j}\beta_{B1}^z + W_j \gamma_1^z + \alpha_{j1}^z + \epsilon_{ij1}^z \;\; [M_1^Z] \tag{21}$$

with $\alpha_{j0}^z \sim N(0, \sigma^2_{\alpha_0^z})$, $\alpha_{j1}^z \sim N(0, \sigma^2_{\alpha_1^z})$, $\epsilon_{ij0}^z \sim N(0, \sigma^2_{\epsilon_0^z})$, and $\epsilon_{ij1}^z \sim N(0, \sigma^2_{\epsilon_1^z})$, all mutually independent. We can unbiasedly estimate all of the variance components for the Z process, with $\hat{\sigma}^2_{\epsilon_0^z}$, $\hat{\sigma}^2_{\epsilon_1^z}$, estimates of the denominator sums in (18): $\sigma_z^2 + (\zeta^z)^2 + \beta^{z\prime} V(X^*)\beta^z$ and $\sigma_z^2 + (\zeta^z)^2$, respectively, which we can shorthand as c_{W0} and c_{W1}, referring to a partition of the Z process variance, as per (11). Similarly, $\hat{\sigma}^2_{\alpha_0^z}$, $\hat{\sigma}^2_{\alpha_1^z}$, respectively, are estimates of certain terms in the denominator sums in (19): $\psi^z + (\delta^z)^2 + \gamma^{z\prime} V(W^*)\gamma^z$ and $\psi^z + (\delta^z)^2$, which we can shorthand as c_{B0} and c_{B1} as per (11). We replace products of confounding parameters by using these definitions and notation: $\zeta^{yz} = \zeta^y \zeta^z$, $\delta^{yz} = \delta^y \delta^z$, $\beta^{yz} = \beta^{y\prime} V(X^*)\beta^z$ and $\gamma^{yz} = \gamma^{y\prime} V(W^*)\gamma^z$. Then under our assumptions,

$$\Delta_W = \text{Bias}[\hat{\tau}_{W0}] - \text{Bias}[\hat{\tau}_{W1}] \quad = \quad \frac{\zeta^{yz} + \beta^{yz}}{c_{W0}} - \frac{\zeta^{yz}}{c_{W1}} \tag{22}$$

$$\Delta_B = \text{Bias}[\hat{\tau}_{B0}] - \text{Bias}[\hat{\tau}_{B1}] \quad = \quad \frac{(\zeta^{yz} + \beta^{yz})/N_j + \delta^{yz} + \gamma^{yz}}{c_{W0}/N_j + c_{B0}} - \frac{\zeta^{yz}/N_j + \delta^{yz}}{c_{W1}/N_j + c_{B1}} \tag{23}$$

See Appendix A for summary of notation. We maintain the "finite number of groups" version of the equations in this case, as it is easy to remove the terms later. We note that expressions for β^{yz} and γ^{yz} have only one term each that is not unbiasedly estimable – the parameters $\{\beta^y, \gamma^y\}$ from the equations in models M_0^Y and M_1^Y, respectively. We also note that the above expressions are for fixed group size in the asymptotics.

3.3 A multistage OLS model

There is one more piece of information that we have yet to use. While we cannot fit a GLS (or random-effects) model without introducing an additional parameter λ (see Appendix B, equation (B.8)), we can fit OLS models without introducing any new parameters.[7] We fit new models M_2^Y and M_3^Y, in which we do *not* group-mean center Z, as

$$Y_{ij} = \beta_{02}^y + \tau_{O2} Z_{ij} + \epsilon_{ij2}^y \ \ [M_2^Y]$$

$$Y_{ij} = \beta_{03}^y + (X_{ij} - \bar{X}_{\cdot j})\beta_{OW3}^y + \tau_{O3} Z_{ij} + \bar{X}_{\cdot j}\beta_{OB3}^y + W_j \gamma_{O3}^y + \epsilon_{ij3}^y \ \ [M_3^Y]$$

using OLS (thus the subscript 'O'). These assume simpler error structure with $\epsilon_{ij2}^y \sim N(0, \sigma^2_{\epsilon_2^y})$ and $\epsilon_{ij3}^y \sim N(0, \sigma^2_{\epsilon_3^y})$ independent, and no group effects.

Then the bias difference equations for OLS (maintaining the definitions of X^* and W^*) are:

$$\begin{aligned}\text{Bias}[\hat{\tau}_{O2}] - \text{Bias}[\hat{\tau}_{O3}] \ &= \ \frac{\zeta^y \zeta^z + \beta^{y\prime} V(X^*)\beta^z + \delta^y \delta^z + \gamma^{y\prime} V(W^*)\gamma^z}{\sigma_z^2 + (\zeta^z)^2 + \beta^{z\prime} V(X^*)\beta^z + \psi^z + (\delta^z)^2 + \gamma^{z\prime} V(W^*)\gamma^z} \\ &\quad - \ \frac{\zeta^y \zeta^z + \delta^y \delta^z}{\sigma_z^2 + (\zeta^z)^2 + \psi^z + (\delta^z)^2}\end{aligned}$$

and the asymptotic change in bias between M_2^Y and M_3^Y can be written:

$$\Delta_O = \text{Bias}[\hat{\tau}_{O2}] - \text{Bias}[\hat{\tau}_{O3}] \ = \ \frac{\zeta^{yz} + \delta^{yz} + \beta^{yz} + \gamma^{yz}}{c_{W0} + c_{B0}} - \frac{\zeta^{yz} + \delta^{yz}}{c_{W1} + c_{B1}}. \qquad (24)$$

3.4 Constraints to confounding parameters

We now have three equations (22)-(24), and four unknowns, $\zeta^{yz}, \delta^{yz}, \beta^{yz}, \gamma^{yz}$. Our sequential models M_0^Y-M_3^Y provide estimates of Δ_W, Δ_B and Δ_O. We can add a fourth equation, effectively assigning a value to one of the four unknowns, and then all others will be a function of it. This set of equations constrains the space of confounding parameters from the entire $\delta^{yz} - \zeta^{yz}$ plane to a single line. We can express all relationships with this linear

[7] Adding a new parameter with a new equation is problematic in this context; in particular, we cannot estimate λ unbiasedly.

system of equations (letting group size go to infinity for the between estimators):

$$\begin{pmatrix} \frac{1}{c_{W0}} - \frac{1}{c_{W1}} & 0 & \frac{1}{c_{W0}} & 0 \\ 0 & \frac{1}{c_{B0}} - \frac{1}{c_{B1}} & 0 & \frac{1}{c_{B0}} \\ \frac{1}{c_{W0}+c_{B0}} - \frac{1}{c_{W1}+c_{B1}} & \frac{1}{c_{W0}+c_{B0}} - \frac{1}{c_{W1}+c_{B1}} & \frac{1}{c_{W0}+c_{B0}} & \frac{1}{c_{W0}+c_{B0}} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \zeta^{yz} \\ \delta^{yz} \\ \beta^{yz} \\ \gamma^{yz} \end{pmatrix} = \begin{pmatrix} \Delta_W \\ \Delta_B \\ \Delta_O \\ \eta \end{pmatrix} \tag{25}$$

The fourth row in the set of equations specifies the value of one open parameter, η; we evaluate the conditions for the matrix inversion when the open parameter is in the third or fourth column. The requirement reduces to $\frac{c_{W0}}{c_{W1}} \neq \frac{c_{B0}}{c_{B1}}$, which will be easily satisfied if both types of predictors reduce the unexplained group and individual level variation in different proportions. This can also be understood as the within and between estimators providing different information about the unknown parameters. In addition, for this parametrization, $c_{B1} \neq c_{B0}$ and $N_j > 1$ is required in the finite group-size equations (not shown). In MLMs, this translates to some additional variance of the type associated with the open parameter η needs to be explained by M_1^Z over M_0^Z.

From (25), we can solve for ζ^{yz}, δ^{yz} and β^{yz} given $\gamma^{yz} = \eta$, e.g. As alluded to above, our choice to allow either β^{yz} or γ^{yz} to vary was not arbitrary. First, we are most interested in the relationship between ζ^{yz} and δ^{yz}; we characterize this as the confounder space. Expressing these parameters in terms of a less central parameter enables this.

The solution to the system of equations is a line in confounder space. While this is a dramatic improvement, in terms of restricting the plausible confounding scenarios, the line is still of infinite length, covering a wide range. However, practically speaking, confounders exhibiting "effect sizes" beyond one or two standard deviations are implausible, so restricting the line to a segment could be governed by substantive knowledge. Fortunately, we are also able to further restrict the set of possible scenarios by learning from the data and models. We first form a bound for $\gamma^{yz} = \gamma^{y\prime}V(W^*)\gamma^z$ by noting that it is a particular covariance, $\mathrm{Cov}(W^*\gamma^y, W^*\gamma^z)$. From the Cauchy-Schwartz inequality, $|\mathrm{Cov}(W^*\gamma^y, W^*\gamma^z)| \leq \sqrt{\mathrm{Var}(W^*\gamma^y)\mathrm{Var}(W^*\gamma^z)}$, and this translates to $|\gamma^{y\prime}V(W^*)\gamma^z| \leq \sqrt{\gamma^{y\prime}V(W^*)\gamma^y\gamma^{z\prime}V(W^*)\gamma^z}$. The second term in the product, $\gamma^{z\prime}V(W^*)\gamma^z$, may be estimated unbiasedly from (21). The first term, $\gamma^{y\prime}V(W^*)\gamma^y$, requires a little more work; it may be bounded by noting that it captures a portion of between group variation in Y. Unfortunately, even the between group variation estimable from an unconditional means model for Y, call this c_B^y, may underestimate $\gamma^{y\prime}V(W^*)\gamma^y$ when group effects in Z negate the additional group effects in Y, under our DGP. So a *potential* bound for $\gamma^{y\prime}V(W^*)\gamma^y$ is c_B^y, but it may not be sufficient. Bound c_B^y fails when $\tau W\gamma^z = -W\gamma^y$; this is when the contribution of $\gamma^{y\prime}V(W^*)\gamma^y$ to the total variance in Y is masked. This worst case "cancellation" should provide us with an alternative upper bound for $\gamma^{y\prime}V(W^*)\gamma^y$; this is when

$\gamma^y = -\tau\gamma^z$. That condition implies $\gamma^{y\prime}V(W^*)\gamma^y = \tau^2\gamma^{z\prime}V(W^*)\gamma^z$, so that $|\tau|\gamma^{z\prime}V(W^*)\gamma^z$, is an alternative upper bound for $|\gamma^{yz}|$. We set a bound by choosing the maximum of $|\tau|\gamma^{z\prime}V(W^*)\gamma^z$, $\sqrt{c_B^y\gamma^{z\prime}V(W^*)\gamma^z}$, and the (biased, so possibly under-) estimated value of $\sqrt{\gamma^{y\prime}V(W^*)\gamma^y\gamma^{z\prime}V(W^*)\gamma^z}$. Underestimation could occur because c_B^y may be biased. By choosing the largest of three estimates, the overall approach is conservative; the bound under the cancellation scenario ($\gamma^y = -\tau\gamma^z$) will overcompensate when cancellation does not occur. One may be surprised to require a value for τ, but it may be chosen to be conservative. In practice, one often has some sense of the magnitude of the treatment effect size (the sign is not needed for this bound).

3.5 Characteristics of the line in confounder space

The prior results indicate that multistage multilevel models can be used to constrain the confounding space to a line segment. We highlight a few characteristics of that line that are implied by the form of the equations. Additional details are given in Appendix C. The first slightly unexpected implication of the system of equations is that their solution for non-degenerate cases is always a *positively sloped* line in the $\delta^{yz} - \zeta^{yz}$ plane. The slope is the ratio of two variance components, and since these must be non-negative, the sign must be non-negative too. While expressions can be derived for the points at which the line crosses either axis, they do not reduce to a simple form. One insight gained is that only a very limited set of values for the confounding parameters would result in the line crossing quadrant II. This is the scenario in which the signs of unobserved within confounding is positive, while it is negative for the corresponding between group confounding. To date, we have not found a dataset and model yielding a line in quadrant II, but they clearly can be created in simulated data.

3.6 Further implications of the framework

Assuming the DGP approximates the observed process reasonably well, the system of equations (25) have strong implications for what must be true in the absence of confounding or situations in which confounding is minimal. For example, if $\zeta^{yz} = 0$, then the first row of (25) reduces to $\beta^{yz} = c_{W0}\Delta_W$ and we have identified β^{yz} unbiasedly. Forming the contrapositive, we can learn about the plausibility of the unbiasedness assumption, $\zeta^{yz} = 0$. The extent to which $\hat{\beta}^{yz} \neq \hat{c}_{W0}\hat{\Delta}_W$ is evidence *against* the supposition that $\zeta^{yz} = 0$. Similarly, if we make the assumption that $\zeta^{yz} = 0$ and $\delta^{yz} = 0$, then using the third row of (25), we have the expression $\beta^{yz} + \gamma^{yz} = (c_{W0} + c_{B0})\Delta_O$. Under these two assumptions, all terms in this expression can be estimated unbiasedly, providing us with additional necessary conditions for unbiasness to hold. In the example using real data, we demonstrate how versions of these implications may be utilized.

Of course, we could find less directly testable implications under weaker assumptions (for example we might know the sign of some parameters once we know the sign of the bias difference). The complex interplay established by the system of equations suggests that we learn a great deal by simply determining in which portion of the confounding space we are most likely to reside, given our modeling assumptions (DGP) and data.

3.7 Simulation Study

To confirm the intuition discussed above and to begin to understand the inferential properties of the model-based estimates in finite samples, we conduct a simulation study. Rather than being a full model-based approach, our sensitivity analysis relies on the asymptotic bias of a set of related estimators. This framework limits our ability to make inferential statements about point estimates and coverage, but this is common in the realm of sensitivity analysis, in which we explore hypotheticals.

The primary purpose of this simulation study is to verify the somewhat bold implications of our modeling framework, namely, that a set of multilevel model fits to a given dataset imply a highly constrained set of possible confounders *a posteriori*. These, in fact, are limited to a line segment in the $\delta^{yz} - \zeta^{yz}$ plane; the "truth" must be contained in this segment. Does this work in practice? Before delving into the details of the simulation we note that several factors are likely to introduce error in the specification of the line segment, and our goal is to quantify the potential impact that they may have. Examples of error-inducing factors include: the idealized nature of asymptotic expectations, the potential numerical instability of linear equations[8] and the DGP linearity and additivity assumptions themselves. We only control for the latter.

For the simulation study, we focus on a case in which the sample size is moderately sized: 100 groups of size 40. This is close to the "middle of the road," in which asymptotics are unlikely to fully hold, and sampling variability could be moderate. An important choice is the range of ICC_Z that we explore. We report results for the range $0.4 - 0.6$, as it is where problems are more likely to emerge (we computed similar assessments with ICC in the $0.15 - 0.35$ range as well, with comparable or better results in terms of coverage).

To explore the space of confounding, we vary the parameters $\{\zeta^y, \zeta^z, \delta^y, \delta^z\}$ so that they approximately assume the values $\{-1.0, -0.7, -0.5, +0.5, +0.7, +1.0\}$[9], while the values of σ^2_α and σ^2_ϵ are fixed at 1 for both Y and Z models. The treatment effect (τ) is set to one, but the implicit effect size, based on the variance of the Z process, varies from 1.5 to 1.8. The predictors X and W are simulated as standard normals, and enter the DGP for Y

[8] Several terms defining the relationships between confounding parameters are reciprocals of differences in variance components.

[9] Before entering the DGP, ζ^y and δ^z are perturbed by plus or minus 10%, respectively, to prevent singularity of the matrix in the system of equations.

and Z with the corresponding coefficients β and γ set to 1. These choices yield $6^4 = 1296$ unique tuples of confounding parameters, with $0.40 \leq ICC_Z \leq 0.60$.

For each of the 1296 combinations, we simulate 100 datasets drawn from the DGP with the two confounders, U and V, included but effectively unknown to the analyst. We then apply our method to each simulated dataset, using $\tau = 1.5$ in the bounds calculation. By choosing a *small* τ, we potentially *underestimate* the bounds on γ^{yz} upon which we rely for constructing (bounded) line segments, so this is a conservative choice for the simulation study. Each dataset and model fit yields a single line segment in confounding space, and the set of 100 (all under the same DGP) form an ensemble of segments that cover a portion of the confounding space. The maximum and minimum coordinates from the line segment identify two diagonally opposite corners of a rectangle, establishing our marginal bounds for each confounding parameter across the corresponding axis.

In the presence of uncertainty, identifying a subset of the confounding space consistent with the data and model is important. Asymptotically, we expect the line to contain the true confounding parameters, but it is an approximation in a finite sample. Our simulation determines how often we successfully identify the range of the confounding parameters given sampling variability common to real data for an exemplary case in which confounding has a strong effect ($ICC_Z \approx 0.50$). We calculate this success rate in our simulations by computing the percent of simulated line segments for which the corresponding rectangle covers the true $(\delta^{yz}, \zeta^{yz})$ used in the DGP. This rate does not represent classical confidence bounds; rather, it is simply an assessment of the success of our method in covering the true confounding parameters.

Coverage as defined in this way is 100% for δ^{yz}, while it is between 90% and 100% for ζ^{yz}. Thus, our method is excellent at identifying a subset of the confounding space that contains the true value of δ^{yz}. Note as well that for 93% of the cases, coverage for ζ^{yz} is at least 95%, and recall that our choice of τ was conservative (small). If we had simulated using a larger value of τ, the line segment would be longer, and the enclosing rectangle larger, so we would cover the true within-group parameter more often. Upon closer inspection, we find that the specific draws from the DGP whose corresponding lines fail to cover the true parameters involve a system of linear equations that tend to have larger condition numbers.[10] A conservative strategy would thus include screening based on this condition number. When computing the line segment determined by the system of equations, we can report the condition number of the associated matrix. In these simulations, if we exclude cases in which the condition number exceeds 6000, then of those remaining, we fail to

[10] A large condition number implies that small perturbations in the terms in the linear equations yield large changes in the solution, which in our case is the line segment.

cover ζ^{yz} only 1.3% of the time.[11] The line segment in analyses with a smaller condition number is thus extremely likely to contain the true confounding parameters, which is a major step forward, in terms of evaluating the conditions required for the treatment effect to be unbiased.

The use of a bounding rectangle is a marginal analysis (each parameter taken separately), which is useful, but in practice, we will be generating line segments and we are interested in determining how often the space covered by an ensemble of segments "covers" points in the true segment; this is a more specific test of an oracle property of our method. We compute the line segment determined by the (true) parameters of the DGP, using the same choice for $\tau = 1.5$ in the bounds.[12] We note that the ensemble of line segments is, for the most part, quite precisely estimated. Variation from sample to sample is not tremendous, unless the line was determined from a system of equations with a very large condition number. For a two-dimensional test of coverage, we take the ensemble of line segments generated by a tuple, remove those generated from a linear system with condition number larger than 1000 (about 5% of lines across the full set of simulations), and then take the convex hull of the remaining line segments.[13] Across 1296 confounding scenarios, we have 100% coverage by this convex hull 98% of the time, with all but five scenarios with coverage at or above 95% (the lowest was 87% but the four remaining were at 94%).

We looked more closely at the five scenarios with line coverage below 95% and find them to be only slightly atypical, but all similar to each other. Their main characteristic is that the signs of ζ^y, ζ^z terms are opposing, as are the signs of the δ^y, δ^z terms, and those terms' magnitudes are relatively small. The ICC_Z is close to 0.60 in four out of five cases. The opposing signs could lead to partial masking of variance components in the Y process, and thus our bounds on the line would rely on the guess for τ in one of the bounding terms. In fact, upon visual inspection we see that the "true" line (based on the known DGP) extends slightly outside the space of the sampled lines, and this is where the lack of coverage occurs (it is not a matter of the slope being incorrect; rather, the ensemble of line segments are slightly too short as they are based on a conservative bound set by the choice of τ.).

Overall, these simulations suggest that in moderately sized samples the framework and resulting sensitivity analysis are fairly robust to different types of confounding. With smaller levels of confounding or smaller ICCs, our methods have equal or better coverage properties.

[11] A cutoff of 6000 excludes fewer than 0.8% of cases, and of these, 20% would have shown acceptible coverage of the true parameter.

[12] This idealized case and its corresponding line most closely resemble a situation in which we have a very large sample and group size.

[13] Removing outlying lines before taking the convex hull is again, a conservative approach. We used a condition number of 1000 to remove essentially all outliers, as these would inappropriately expand the convex hull.

4. Example using IEA Data

We apply this form of sensitivity analysis to a study conducted by the International Association for the Evaluation of Educational Achievement (IEA) in 1970-1971. The study adopted, as measures of verbal ability, a test of reading comprehension, a brief test of speed of reading, and a brief test of word knowledge. Student, teacher and school level covariates were collected. The test was undertaken on three levels, 10-year-old students, 14-year-old students, and students in the final grade of the secondary school. The analysis we present here is based on the final secondary school grade. A multinational study, participating educational systems for this outcome include: Australia, Belgium (Flemish), Belgium (French), Chile, England, Finland, France, Hungary, Iran, Israel, Sweden, Thailand, and USA, among others.[14] In this two level (students within schools) analysis, we used *number of books at home* as the treatment variable, the sex of student, hours spent on homework per week, the raw score of a word knowledge test as individual level controls, with the size of the school, and type of community the school serves (a measure of urbanicity) as group level controls. Within-school controls were group mean centered as required by our modeling framework. The dependent variable is the reading comprehension test raw score. All variables, including the outcome, were standardized to aid in interpretation. Each country is treated as a different dataset in our analysis. Results from three countries, Italy, Scotland and Sweden are presented. The countries chosen reflect a range of scenarios one might experience in practice. Three of the fifteen countries in our analysis dataset yielded condition numbers larger than 1000, so these were not considered in our selection.

For any country, the first row of Table 1 lists estimates of the treatment effect τ for the within, between and OLS estimators, respectively, for a model that excludes all predictors other than the treatment. The second row provides estimates for models that include the available predictors. Analysts typically privilege the second set of estimates over the first, as potential confounders have been controlled, increasing the plausibility of ignorability.[15] However, we use the difference in these two (third row, estimates of $\Delta_W, \Delta_B, \Delta_O$), which we have shown can be used to narrow the viable range of any remaining confounding. We also summarize the variance components estimated from the models for Z in Table 2. These are important inputs to our constraint equations. The terms we note as c_W and c_B, which contain within and between group variation, respectively, are further denoted by subscripts 0 and 1 to indicate models without and with controls, respectively, and are presented in Table 2. From these, we can estimate intra-class coefficients (ICC_Z), which succinctly

[14] See Peaker (1975) and the following link (http://ips.gu.se/english/Research/research_databases/compeat/Before_1995/Six_Subject_Survey/SSS_Sample) for a detailed description of the sampling design.

[15] For all three countries, the within and OLS treatment effect estimates are highly significant, at conventional levels; the between estimates are similarly significant, except for Italy, for which $p = 0.06$.

	Treatment Effect ($\hat{\tau}$)		
Country/Model	Within	Between	OLS
Italy			
without predictors	0.12	0.33	0.22
with predictors	0.06	0.11	0.08
difference (Δ_{bias})	0.06	0.23	0.14
Scotland			
without predictors	0.20	0.52	0.28
with predictors	0.08	0.52	0.15
difference (Δ_{bias})	0.12	0.00	0.14
Sweden			
without predictors	0.14	0.40	0.17
with predictors	0.08	0.28	0.09
difference (Δ_{bias})	0.06	0.12	0.07

Table 1: Multilevel model-based estimates of treatment effect for models excluding and including individual and school predictors.

Country	c_{W0}	c_{B0}	ICC_{Z_0}	c_{W1}	c_{B1}	ICC_{Z_1}
Italy	0.75	0.28	0.27	0.74	0.22	0.23
Scotland	0.81	0.19	0.19	0.78	0.14	0.15
Sweden	0.93	0.08	0.08	0.91	0.04	0.04

Table 2: Multilevel model-based within and between schools variance components and ICC_Z estimates for models for Z excluding and including individual and school predictors.

summarize the proportion of total variation in treatment that is between groups, net of controls.

In the system of equations given in (25), we examine a grid of values for open parameter $\eta = \gamma^{yz}$, bounding the range using the constraints derived in the last section. These bounds are not likely to be sharp, as we are choosing the maximum of three bounds to ensure that we are not misled by potentially biased estimates involved in the construction of the bounds. In practice, we use the minimum of this bound and a complementary bound based on setting the open parameter $\eta = \beta^{yz}$ rather than γ^{yz} and utilizing a correspondingly modified version of (25). This allows us to tighten the bounds somewhat in some instances.

We create a bias difference plot for the within versus the OLS estimators for the model including predictors (M_1^Y or M_3^Y), across a range of $(\delta^{yz}, \zeta^{yz})$ based on the constraints implied by the range for $\eta = \gamma^{yz}$. These plots initially provide us with two pieces of information: a *finite* range of $(\delta^{yz}, \zeta^{yz})$ consistent with the data and model; and the extent to which absolute bias is larger or smaller when using the within or OLS estimators.

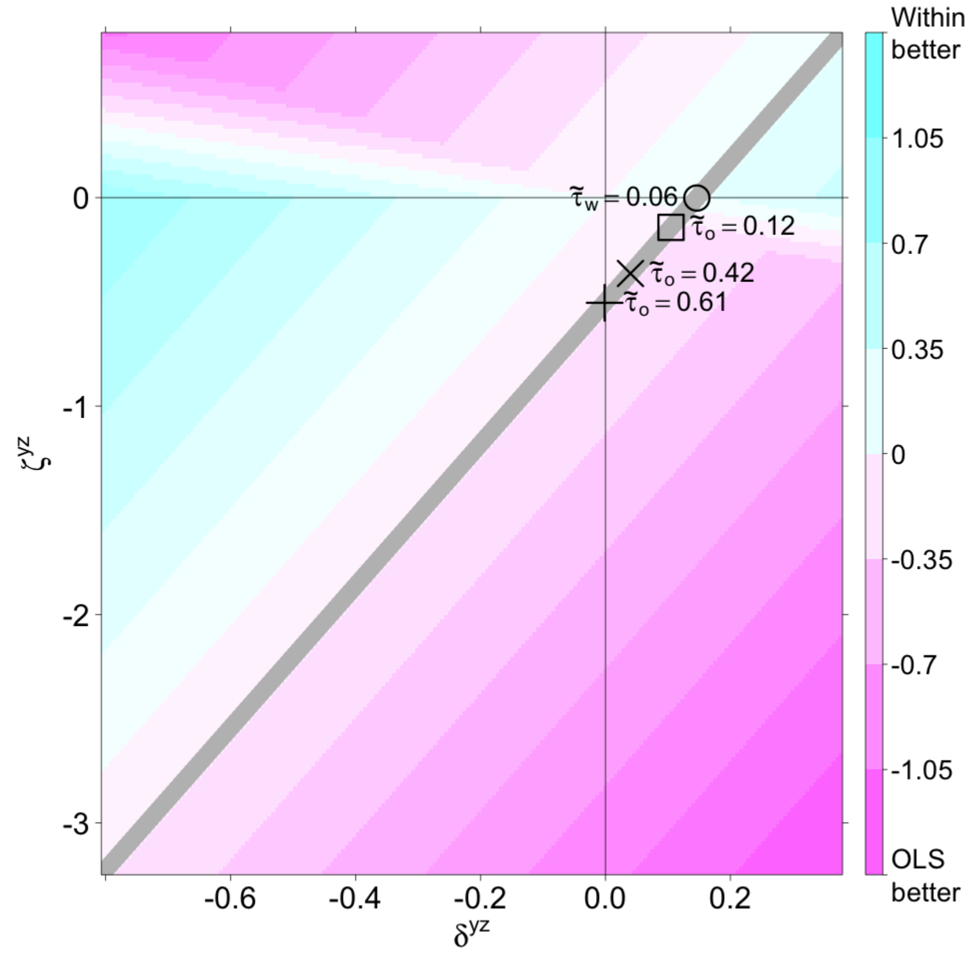

Figure 2: Difference in absolute bias for Italy: $|\text{bias}[\hat{\tau}_{OLS}]| - |\text{bias}[\hat{\tau}_W]|$ for a range of within and between group confounding (see text for more details)

Refer to Figure 2 for Italy. Recall that axis scale and range are determined by the endpoints of the line segment. The range of plausible standardized ζ^{yz} is much larger than that of δ^{yz} when you focus on the scale. Being able to constrain the sensitivity analysis to this rectangle within the $\delta^{yz} - \zeta^{yz}$ plane is an important contribution of this approach. Note that the ICC_{Z_1} for Italy is moderate, at 0.23, so the potential for bias when comparing

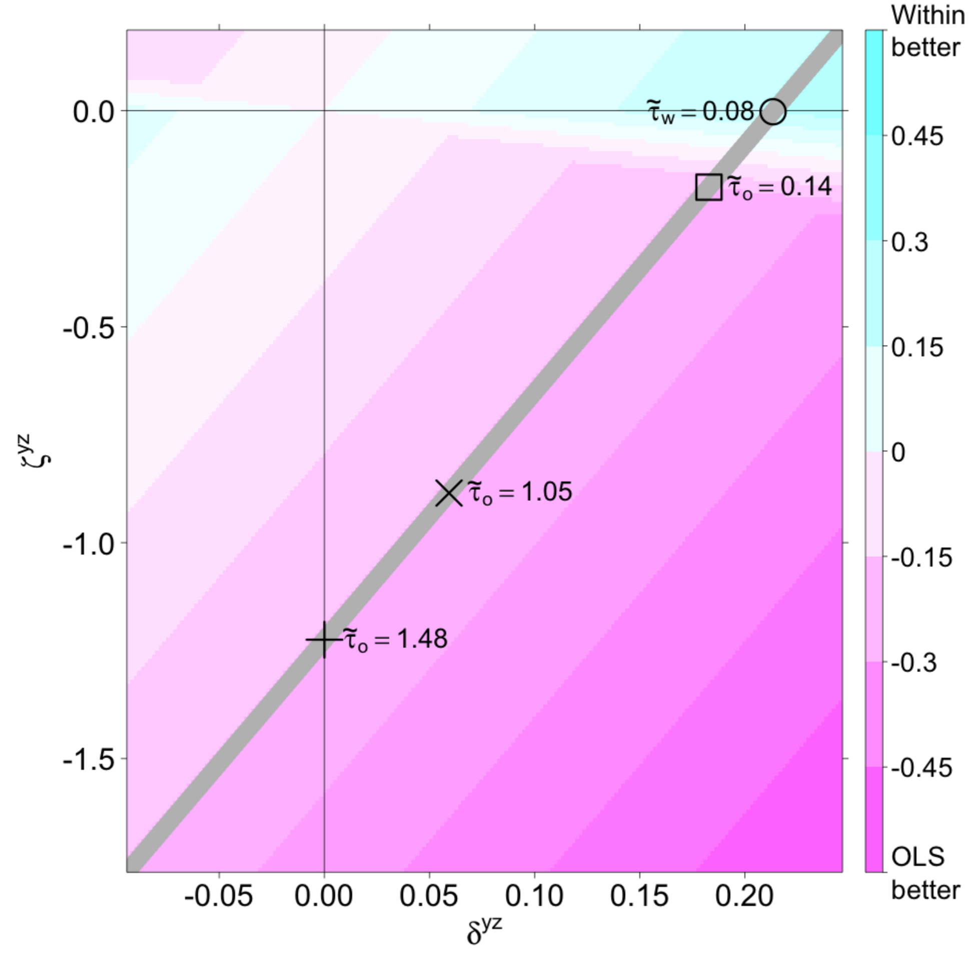

Figure 3: Difference in absolute bias for Scotland: $|\text{bias}[\hat{\tau}_{OLS}]| - |\text{bias}[\hat{\tau}_W]|$ for a range of within and between group confounding (see text for more details)

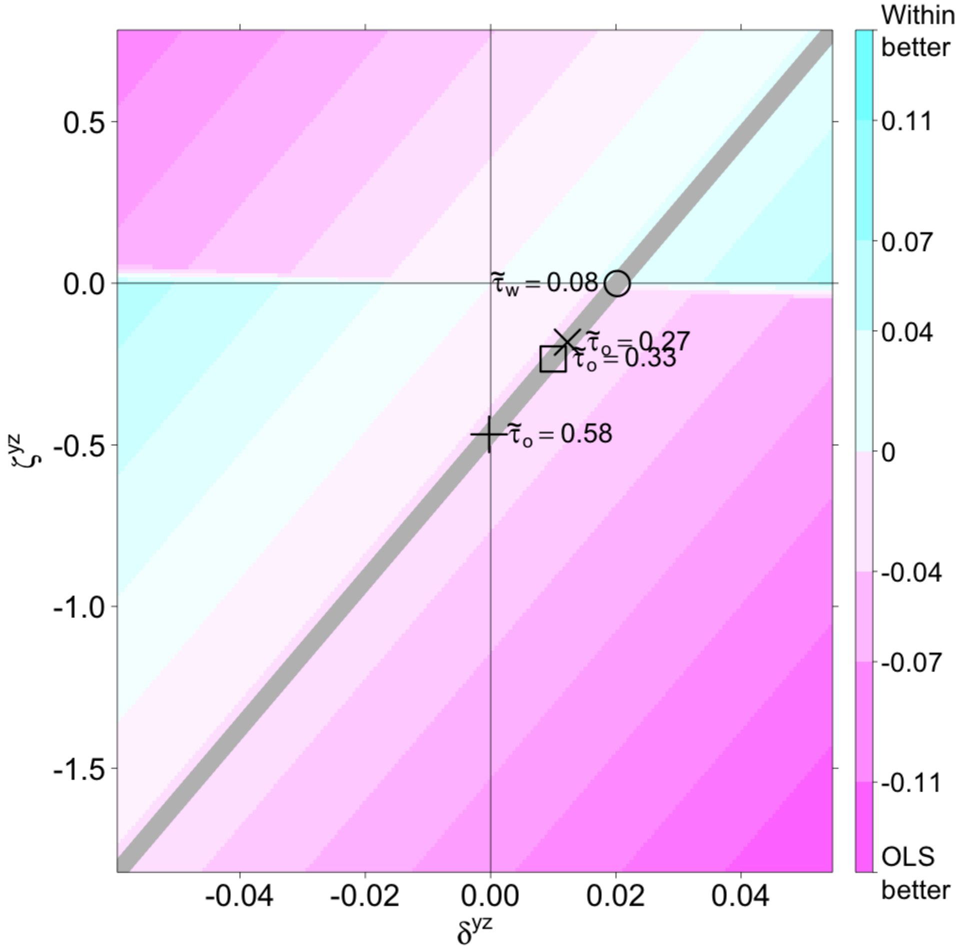

Figure 4: Difference in absolute bias for Sweden: $|\text{bias}[\hat{\tau}_{OLS}]| - |\text{bias}[\hat{\tau}_W]|$ for a range of within and between group confounding (see text for more details)

estimators is moderate, when you examine the color scale, which is in standard deviation units of the outcome.

We now explain the additional elements of these figures. First, a thick grey line is included in the plot. The data, model and asymptotic-expectation-based constraint equations imply this further restriction on confounders ζ^{yz} and δ^{yz}; namely, this line is our "point estimate" of their values, consistent with the observed data and assumed model.[16] Again, this is a dramatic restriction on the potential for confounding. Without bounds and the system of equations (25) based on multiple multilevel fits, we could only have plotted ζ^{yz} versus δ^{yz} based on c_{W1} and c_{B1}, but we could not have restricted the range of plausible scenarios, neither in terms of the range of the axes (the rectangle) nor to the line segment itself.

In addition to the line, we plot four points. A circle is situated at the intersection of $\zeta^{yz} = 0$ and the grey line. This represents the point of no within-group confounding that is consistent with the data and DGP model. It is a single point, and a corresponding estimate of τ can be calculated (to be discussed shortly). Next, symbol '+' marks the intersection of $\delta^{yz} = 0$ and the grey line. This represents the assumption of no remaining group-level confounding that is consistent with the data and DGP. It is a single point as well. Ideally, these points would be very close to each other and thus near the origin, representing minimal or no unobserved confounding. When these two points are far from the origin, the hypothesis of no unmeasured confounders is not supported in the following sense. Since the truth lies on the grey line, one that does not pass through or near the origin suggests that either there is group-level confounding or individual-level confounding present. *The data, relying on a model for confounding, informed us of this fact.* The estimates of τ may be "corrected" by subtracting the asymptotic bias derived in our formulas, plugging in the confounding parameters $(\delta^{yz}, \zeta^{yz})$ given by each location on the line. These are given for key points on the line segment as well. The "less biased" estimator (in an absolute sense, based on the asymptotic bias inequality (13)) is reported, and indicated by a subscript 'o' or 'w' (OLS or within, respectively).

These points and the line upon which they reside form a sensitivity analysis of the implications of group- and subject-level confounding in this multilevel setting. The range of estimates for τ for different confounding parameters indicates how sensitive findings are to the assumptions. In addition, the absolute bias difference between estimators indicates the level of the "danger zone"[17] and the trade-off between estimators. Note that rather than "zoom in" on these points, we show the restricted subset of the coordinate plane consistent

[16] Given that our results are based on asymptotic expectations, the line is what we would expect in very large samples, assuming the parameters estimated remain approximately the same.

[17] We use this term to refer to confounding scenarios in which one estimation method (under its assumptions) has much larger absolute bias than another.

with data and model. Arguably, when the data and model do not restrict the subset to be near the origin, one must allow that the potential for unobserved confounding may be large.

Next, a square symbol is placed on the grey line at a point based on the estimated value of β^{yz}. Under conditions derived previously ($\zeta^{yz} = 0$; see Section 3.6), our models will yield unbiased estimates of this parameter; if so, they provide the value of the fourth unknown in our linear system of equations, which implies a specific point (on the grey line) in the δ^{yz}–ζ^{yz} plane. The extent to which the square's location differs from the circle's is evidence against the unbiasedness with respect to subject-level confounding. Their distance suggests evidence opposing the assumption; i.e., if $\zeta^{yz} = 0$, then β^{yz} should be unbiased, so the two points should be (nearly) coincident. It is a thought experiment, the implications of which can be made explicit, with their discrepancy offering evidence against the original assumption. For Italy, the distance is not tremendous, and the two estimates of the treatment effect, although recommended for different estimators, are close. Both indicate a treatment effect close to zero.

In the same manner, symbol '×' is placed at a point based on the estimated value of γ^{yz}, and the extent to which this differs from the point marked '+' is evidence that group level confounding should be a concern.[18] In Figure 2, for Italy, the symbols '+' and '×', representing the assumption that $\delta^{yz} = 0$, are somewhat near each other, with corresponding τ values near 0.5. OLS estimates are less absolutely biased near these points, but the difference is not too large. Unfortunately, we cannot determine the direction of bias from the points on these plots (the true τ could be to either 'side' of the corrected estimates). Thus, each point represents a confounding scenario based on some restrictive assumption. To the extent that all points are close to one another, a consistent story is plausible. To the extent that they diverge, one must concede that further confounders may be lurking.

While the two stories (based on assumptions regarding within- or between-group confounding) are not identical, they are similar, in that the treatment effect size is small or zero. However, each is based on a strong assumption of the lack of one form of confounding, and we have no way of knowing whether the remaining points on the grey line are plausible. The range of plausible scenarios would suggest negative treatment effects in quadrant I, as well as much larger positive treatment effects in quadrant III. This range is large, but a substantive researcher could contend that a negative treatment effect is not consistent with theory, and thus one could further restrict the sensitivity analysis to reflect this. For Italy, this contention implies that there remains *individual* level confounding with $\zeta^{yz} < 0$.

[18] Note that if either $\zeta^{yz} \neq 0$ or $\delta^{yz} \neq 0$, γ^{yz} may be biased, but we wish to explore the extent to which the implications of this estimate are consistent with those that only assume the latter.

A cautious practitioner might choose to use OLS over a within-group estimator given this evidence.

Evaluating the implications of the lack of within or between group confounding is important even if one plans to use so-called econometric fixed effects (within group) estimators, which are unbiased in the presence of group confounding. Should the evaluation give the practitioner reason to doubt that within-group confounding has been removed, then the within-group estimator will be biased, and the extent of this bias ought to be examined.

In Figure 3, for Scotland, we first note that the range for both within and between group confounding (the rectangular portion of the $\delta^{yz} - \zeta^{yz}$ plane) is smaller than it was for Italy. Our bounds on the line are strong enough to restrict us to this smaller range or "strength" of confounding. In other words, the mixtures of confounding (between, within) consistent with the data and models cover a smaller portion of the coordinate plane, the majority being in quadrant IV. However, the differences between estimates under different assumptions are more dramatic in the sense that estimates of τ at the two intersections of the grey line and $\zeta^{yz} = 0$ and $\delta^{yz} = 0$ disagree by over 1 unit. The '+' and '×' symbols are somewhat near each other, as are the circle and square, so there is at least an internal consistency under different assumptions on the extent of confounding. However, these two different assumptions (within or between group confounding being negligible) yield quite different estimates of τ.

Perhaps the discrepancy between all four estimates gives the practitioner reason to pause and reconsider the set of control variables, particularly within group, where the extent of confounding is likely to be large. The location of the bulk of the line segment suggests that OLS offers a "hedge" against the potential within group confounding. The OLS estimate of τ in Table 1 is 0.15, which is about twice the within-group estimate, but is somewhat robust to failure of that estimator's key assumption. Equally to the point, if one fit the within estimator and did not perform the above sensitivity analysis, one would be ignoring the bulk of *a posteriori* evidence that ignorability (at the individual level) was not satisfied.[19] Thus, blindly utilizing a within-estimator in this situation is inadvisable.

Figure 4, for Sweden, might at first appear to resemble the graphic for Italy but there are important differences, which we now highlight. While the estimates of τ assuming $\zeta^{yz} = 0$ or $\delta^{yz} = 0$ are similar to those of Italy, and differ by about 0.5, the pattern of the implied estimates of β^{yz} and γ^{yz} differs – the '×' should be closer to the '+', but the square is, suggesting a greater discrepancy in the implications of the assumptions. Such discrepancy should give the researcher reason to pause and reassess the sufficiency of the controls. Upon deeper inspection, however, it appears that the magnitude of both types of confounding is

[19] The proportion of the confounding space that is consistent with ignorable individual-level confounding is small. The plot provides the implications of various assumptions; one would have to have very strong priors to ignore so much of the potential confounding space.

smaller for Sweden (the portion of the plane is smaller upon examining the axes' scales). Further, the difference in estimates obtained by using the OLS versus within estimators is smaller than it was for Italy. The results are less sensitive to choice of estimator, yet about as sensitive to assumptions regarding the extent of remaining confounding. Perhaps bias amplification or unmasking (Middleton et al. 2016; Pearl 2012) is less of a concern given this lack of sensitivity to estimator. Again, this country is an example in which individual-level confounding dominates the sensitivity analysis, suggesting that ignorability is questionable. However, the choice of estimator cannot compensate for this, given the small difference in magnitude of OLS versus within estimates on the line segment determined by the analysis.

5. Discussion

This approach to multilevel sensitivity analysis provides the researcher with far more information about the range and impact of potential omitted confounders than the results obtained from using any of the individual estimators on its own. Furthermore it provides more information than traditional sensitivity analyses that explore the impact of a single confounder. Our approach quantifies the extent to which different estimators can be expected to disagree on treatment effect, allowing one to choose an estimator with smaller absolute bias, depending on the assumptions regarding remaining confounding. The framework builds on that established by Imbens (2003), provides a model-based assessment of potential bias amplification and unmasking as per Middleton et al. (2016) and Pearl (2012) and is a natural extension of Carnegie et al. (2016) to the multilevel setting. In contrast to much of the prior research on sensitivity analysis, however, this use of sequential multilevel models to partially identify a subspace of "viable" confounders limits the set of feasible confounding parameters using empirical evidence, rather than thought experiments.

The bolder implication of our modeling framework is that estimates from sequential multilevel models imply a highly constrained set of possible confounders, ultimately limited to a line segment in the $\delta^{yz} - \zeta^{yz}$ plane. If our models and assumptions are correct, and we have closely approximated the data generating process, the "truth" should be contained in a finite line segment. The points of intersection of this segment with the two axes, the extent to which estimates differ from one another, and the additional point estimates based on alternative confounding assumptions are quite revealing about the plausibility of and tradeoffs between methods. Moreover, the distance between these points and the origin suggest the degree of additional within and between group confounding. Prior to the development of this tool, researchers could learn very little about the plausibility of confounding scenarios through sensitivity analysis and instead were forced to examine the entire space of confounders.

Importantly, our methods offer some protection against drawing incorrect conclusions from the scenario in which changes in treatment effect estimates upon adding controls decreases the researchers confidence in the model. In this case, one has evidence via changes in $\hat{\tau}$ that an omitted confounder initially biased the treatment effect. It is safe to assume that variance components change with the addition of these new predictors. Given our change in bias formulas (22-24) and using equation (22) for the within estimator as an example, the change in bias implies that $\frac{\zeta^{yz}+\beta^{yz}}{c_{w_0}} \neq \frac{\zeta^{yz}}{c_{w_1}}$. When the difference in the denominators is small, β^{yz} must be changing, and we have evidence that we have removed the impact of a confounder. When the denominators differ, we have reduced the remaining unexplained variance, leaving less room, so to speak, in which the omitted confounders U and V can operate. This identifies the line in the confounding space and tightens the bounds we make on it as a segment.

Under a scenario in which additional predictors do not change the treatment effect estimate nor reduce variance components (between and within groups), we will not be able to impose a restriction on the confounding space consistent with the data and model. This should disabuse us of the notion that we have a robust estimate of the treatment. Thus, in very different scenarios, our multilevel sensitivity framework does more to unpack the meaning behind changes in $\hat{\tau}$ (or lack thereof).

We have shown how the assumptions of our DGP lead to the ability to identify a reduced set of feasible confounding. It is important to consider the extent to which the current findings rely on these assumptions particularly with regard to independence between confounders and observed predictors, as well as among the confounders themselves. Regarding the first consideration, if observed predictor X_{ij} were correlated with unobserved confounder U_{ij}, then by necessity, we would have to introduce a parameter that captured that relationship in addition to those already posited. While nothing precludes this type of formulation, we adopt the simpler assumption used by Imbens (2003), which is that U and in our case V are formulated *net* of observed variables. Put another way, we conceptualize U as that part of omitted confounder that is orthogonal to the rest of the confounders. This allows us to use as much existing software, particularly for MLMs, as possible, as these impose the traditional assumptions of independence of random components with each other and with predictors. If we were to approach this problem using a Bayesian inferential framework, additional parametrization would not pose a problem.

The assumption that group and subject random effects (α and ϵ terms) are also orthogonal to U and V is less restrictive than it might first appear. In fact, one could model the correlations between (α^y, α^z) and (ϵ^y, ϵ^z) across response and treatment models directly, and this is equivalent to our DGP under a renaming of parameters. Let $\alpha^{y'} = \alpha^y + \delta^y U$ and $\alpha^{z'} = \alpha^z + \delta^z U$. Let $\epsilon^{y'} = \epsilon^y + \zeta^y V$ and $\epsilon^{z'} = \epsilon^z + \zeta^z V$. This removes U and V from the model, and we have closed-form expressions for $cor(\alpha^{y'}, \alpha^{z'}) = \frac{\delta^{yz}}{\sqrt{(1+(\delta^y)^2)(1+(\delta^z)^2)}}$ and $cor(\epsilon^{y'}, \epsilon^{z'}) = \frac{\zeta^{yz}}{\sqrt{(1+(\zeta^y)^2)(1+(\zeta^z)^2)}}$ in terms of the original confounding parameters. While at first it might appear that we have only two relationships to model, the correlations implicitly depend, respectively, on two variances, $\sigma^2_{\alpha^y}$ and $\sigma^2_{\epsilon^y}$, neither of which is estimable unbiasedly in our DGP for Y (recall that group structure in Z can mask group structure in Y and thus yield underestimates of variance components in Y). This reparametrization could prove more intuitive to some researchers, and thus equivalent plots of confounding space may be made from these as well. The four equivalent parameters, two correlations and two variances, are only partially identified as well.

While we believe that this sensitivity analysis greatly extends the concrete tools an analyst can apply to causal inference problems, there are some limitations to the methodology. First, it relies on asymptotic bias results; to the extent that our sample is not large enough for these to be a good approximation to the bias, we may have unstable results. Our use of the condition number as a screen should protect against this somewhat. Related to this is the lack of an inferential framework for the estimates at different stages of the analysis. For example, we do not have a sampling distribution for parameters defining the line segment. This is due indirectly to our desire to build on already existing R software libraries. Although (asymptotic) inference is readily available for estimated parameters of MLMs, the inference is incorrect under model misspecification, and our framework relies on model building exploiting a set of highly-controlled model misspecifications. Inference cannot be easily corrected, for example, to account for relationships across treatment and response models. Bootstrap standard errors have been considered, but these are unstable and formulation is non-trivial in realistic MLM scenarios. Future work will instead explore a Bayesian framework for specifying the unobserved heterogeneity, re-establishing our sensitivity analysis as a set simultaneous (linked) equations similar to seemingly unrelated regressions (SUR; see (Greene 2003)). Preliminary findings for this extension of the framework are quite promising.

Acknowledgments

This research was partially supported by Institute of Education Sciences grants R305D110037 and R305B120017. The authors thank Jeremy Yang for research assistance.

Appendices

A. Table of Notation

Model Parameters	
τ	Treatment effect
β^y, β^z	Subject-level predictor effects for response (Y) and treatment (Z)
γ^y, γ^z	Group-level predictor effects for response (Y) and treatment (Z)
ζ^y, ζ^z	Effects for subject-level confounder (U) on response and treatment
δ^y, δ^z	Effects for group-level confounder (V) on response and treatment
Model Random Effects	
α^y, α^z	Group random effects for response and treatment
ϵ^y, ϵ^z	Subject level random effects (error) for response and treatment
Model Variance Parameters	
σ_y^2, σ_z^2	Subject-level error variance for response and treatment
ψ^y, ψ^z	Group-level error variance for response and treatment
Composite Parameters	
$c_W = \sigma_z^2 + (\zeta^z)^2$	Total subject-level variance for treatment model
$c_B = \psi^z + (\delta^z)^2\sigma_y^2, \sigma_z^2$	Total group-level variance for treatment model
$\beta^{yz} = \beta^y\beta^z$	Strength of subject-level confounding due to observables (X)
$\gamma^{yz} = \gamma^y\gamma^z$	Strength of group-level confounding due to observables (W)
$\zeta^{yz} = \zeta^y\zeta^z$	Strength of subject-level confounding due to unobservable (U)
$\delta^{yz} = \delta^y\delta^z$	Strength of group-level confounding due to unobservable (V)
Estimated Model-Specific Parameters	
$\hat{c}_{W0}, \hat{c}_{W1}$	Total subject-level variance for treatment models M_0^Z, M_1^Z, resp.
$\hat{c}_{B0}, \hat{c}_{B1}$	Total group-level variance for treatment models M_0^Z, M_1^Z, resp.
$\hat{\tau}_{W0}, \hat{\tau}_{W1}$	Within treatment effect estimate for outcome models M_0^Y, M_1^Y, resp.
$\hat{\tau}_{B0}, \hat{\tau}_{B1}$	Between treatment effect estimate for outcome models M_0^Y, M_1^Y, resp.
$\hat{\tau}_{O2}, \hat{\tau}_{O3}$	Between treatment effect estimate for outcome models M_2^Y, M_3^Y, resp.

B. Bias Derivations

From Middleton et al. (2016), based on Greene (2003), we begin with a general regression model with treatment variable Z, with specified covariates S_* independent of these, and a set of omitted covariates labeled O. The DGP for outcome Y is

$$Y = S_*\beta^S + O\beta^O + \tau Z + \varepsilon', \tag{B.1}$$

but we fit this model:

$$Y = S_*\beta^{S'} + \tau' Z + \varepsilon', \tag{B.2}$$

inducing bias through the omission of O. Then the bias on τ (the treatment effect) can be written

$$\mathrm{Bias}[\hat{\tau}] = \mathrm{E}\left[\left(Z'Z - Z'S_*[S_*'S_*]^{-1}S_*'Z\right)^{-1} Z'O\beta^O\right], \tag{B.3}$$

where β^O are the coefficient(s) on omitted variable(s) O in the DGP. This expectation will be shown to simplify into ratios of estimators of covariances and conditional variances. We then apply the continuous mapping (CM) and Slutsky theorems, letting the sample size N grow. It is in this sense that we are deriving and utilizing asymptotic expectations. Typically, these asymptotics are the limit, as the number of individuals goes to infinity; whether we hold the number of groups constant will depend on the context. In what follows, assume (wlog) that the variables are mean-centered, so that all marginal expectations are zero.

Bias for the Within or Fixed Effects Estimator

When estimating using fixed effects, $O = [U]$ and $S_* = [X\ W\ \mathbf{D}]$ ($\mathbf{D}$ are separate indicators for each group; refer to Section 2.3 for remaining definitions). Substituting into Equation (B.3), we get:

$$\mathrm{Bias}[\hat{\tau}_W] = \mathrm{E}\left[\left(Z'Z - Z'S_*[S_*'S_*]^{-1}S_*'Z\right)^{-1}\left(Z'U\zeta^y\right)\right] \tag{B.4}$$

We recognize estimators within the expression. With implicit sample size indexing the terms, $\left\{\frac{1}{N}\left[Z'Z - Z'S_*[S_*'S_*]^{-1}S_*'Z\right]\right\}^{-1} \to_p \mathrm{var}(Z \mid X, W, \mathbf{D})^{-1}$, using the CM theorem twice (the expectation operator is the the first continuous mapping, while the reciprocal is the second, defined on the positive real numbers). Similarly, $\frac{1}{N}Z'U\zeta^y \to_p \zeta^y \mathrm{cov}(Z, U)$, applying the CM theorem with the expectation operator. Lastly, we apply the expectation operator and Slutsky's theorem to show that the product in the expression converges to the product of the two limits:

$$\begin{aligned}\lim_{N\to\infty} \mathrm{Bias}[\hat{\tau}_W] &= \mathrm{E}\left[\frac{\zeta^y \mathrm{cov}(Z,U)}{\mathrm{var}(Z \mid X, W, \mathbf{D})}\right] \\ &= \frac{\zeta^y \zeta^z \sigma_u^2}{\sigma_z^2 + (\zeta^z)^2\sigma_u^2}\end{aligned} \tag{B.5}$$

Bias for the between estimator

For this estimator, we set Z in Equation (B.3) to reflect the between estimator only, so that if there are J groups with common group size N_J, we set $\{Z\}_{ij} = \bar{Z}_{\cdot j}$, for each subject i in group j ($\bar{Z}_{\cdot j}$ is the mean treatment for group j and is repeated for each member of the group).

Setting $O = [U\ V]$ and $S_* = [X\ W]$, substituting into Equation (B.3), and using a similar limit argument, we get:

$$\lim_{N\to\infty} \text{Bias}[\hat{\tau}_B] = \text{E}\left[\frac{\zeta^y \text{cov}(Z,U) + \delta^y \text{cov}(Z,V)}{\text{var}(Z \mid X, W)}\right[\tag{B.6}$$

Then substituting our group mean version of Z and evaluating the asymptotic expectation, as $J \to \infty$, holding group size constant, the bias is:

$$\begin{aligned}\lim_{J=N/N_J\to\infty} \text{Bias}[\hat{\tau}_B] &= \text{E}\left[\frac{\zeta^y \text{cov}(\bar{Z}_{\cdot j}, U) + \delta^y \text{cov}(\bar{Z}_{\cdot j}, V)}{\text{var}(\bar{Z}_{\cdot j} \mid X, W)}\right] \\ &= \frac{\zeta^y \zeta^z \sigma_u^2 / N_J + \delta^y \delta^z \sigma_v^2}{(\sigma_z^2 + (\zeta^z)^2 \sigma_u^2)/N_J + \psi^z + (\delta^z)^2 \sigma_v^2} \end{aligned} \tag{B.7}$$

The N_J terms arise due to the independence of all but one term in the mean $\bar{Z}_{\cdot j}$ with U.

Bias for GLS or random effects

While we do not utilize the GLS or random effects estimator in our sensitivity analysis, it is a compromise between the two prior estimators, taking a weighted average of both based on the ICC and group size. This yields a weighted bias in the treatment effect, as follows (here, we show the bias as an approximation, rather than repeat the limit argument).

$$\text{Bias}[\hat{\tau}_{GLS}] \approx \frac{\zeta^y \zeta^z \sigma_u^2 + \lambda \delta^y \delta^z \sigma_v^2}{\sigma_z^2 + (\zeta^z)^2 \sigma_u^2 + \lambda \left(\psi^z + (\delta^z)^2 \sigma_v^2\right)} \tag{B.8}$$

where $\lambda = \frac{\sigma_y^2}{\sigma_y^2 + N_J \psi^y}$. Given our embedded unobserved confounding, λ can not be estimated unbiasedly. The introduction of an additional parameter is one reason that we do not utilize the GLS estimator in our framework.

Bias under OLS

When estimating using OLS, $O = [U \ V \ \mathbf{D}]$ and $S_* = [X \ W]$. Substituting into Equation (B.3), we get:

$$\lim_{N\to\infty} \text{Bias}[\hat{\tau}_{OLS}] = \text{E}\left[\left(Z'Z - Z'S_*[S_*'S_*]^{-1}S_*'Z\right)^{-1}\left(Z'U\zeta^y + Z'V\delta^y + \sum_j Z'D_j\alpha_j^y\right)\right] \tag{B.9}$$

The term, $\sum_j Z'D_j\alpha_j^y$, has expectation zero under our DGP assumptions. Thus, for the OLS estimator, the bias is:

$$\lim_{N\to\infty} \text{Bias}[\hat{\tau}_{OLS}] = \text{E}\frac{\zeta^y \text{cov}(Z,U) + \delta^y \text{cov}(Z,V)}{\text{var}(Z \mid X, W)} = \frac{\zeta^y \zeta^z \sigma_u^2 + \delta^y \delta^z \sigma_v^2}{\sigma_z^2 + (\zeta^z)^2 \sigma_u^2 + \psi^z + (\delta^z)^2 \sigma_v^2} \tag{B.10}$$

We do not repeat the limit derivation, as it similarly applies CM and Slutsky's theorems and the expectation operator.

Bias in models with observed covariates excluded

It is common in the multilevel setting to estimate models with fewer predictors first and then include additional predictors, learning from the changes in variance components. Before we can provide estimates of bias differences, we need an expression for the bias under the model with no predictors. In terms the DGP defined in (B.1), we are omitting both S_* and O from the estimation model, and deriving the bias associated with that estimator. These omissions lead to additional terms in the expressions for bias, as follows.

$$\lim_{N\to\infty} \text{Bias}[\hat{\tau}_{W0}] = \frac{\zeta^y\zeta^z + \beta^{y\prime}V(X)\beta^z}{\sigma_z^2 + (\zeta^z)^2 + \beta^{z\prime}V(X)\beta^z} \tag{B.11}$$

$$\lim_{J=N/N_J\to\infty} \text{Bias}[\hat{\tau}_{B0}] = \frac{(\zeta^y\zeta^z + \beta^{y\prime}V(X)\beta^z)/N_j + \delta^y\delta^z + \gamma^{y\prime}V(W)\gamma^z}{(\sigma_z^2 + (\zeta^z)^2 + \beta^{z\prime}V(X)\beta^z)/N_j + \psi^z + (\delta^z)^2 + \gamma^{z\prime}V(W)\gamma^z} \tag{B.12}$$

The expression for the OLS estimator includes a similar set of additional terms:

$$\lim_{N\to\infty} \text{Bias}[\hat{\tau}_{O2}] = \frac{\zeta^y\zeta^z + \beta^{y\prime}V(X)\beta^z + \delta^y\delta^z + \gamma^{y\prime}V(W)\gamma^z}{\sigma_z^2 + (\zeta^z)^2 + \beta^{z\prime}V(X)\beta^z + \psi^z + (\delta^z)^2 + \gamma^{z\prime}V(W)\gamma^z} \tag{B.13}$$

C. Characteristics of the line describing plausible confounders

We first show that the line-segment solution for non-degenerate cases is always a positively sloped line in the $\delta^{yz} - \zeta^{yz}$ plane.

To see this, we solve the system for either δ^{yz} or ζ^{yz} in terms of η and the estimable parameters $c_{W0}, c_{B0}, c_{W1}, c_{B1}$. We use the simpler expression for between estimator bias in which group size $N_J \to \infty$. We find that $\zeta^{yz} = \kappa_W + \frac{c_{W1}}{c_{B0}-c_{B1}}\eta$ and $\delta^{yz} = \kappa_B + \frac{c_{B1}}{c_{B0}-c_{B1}}\eta$, with κ_W and κ_B representing two constants that do not depend on η. It is natural to assume $c_{B0} > c_{B1}$ (from what we know about sequential variance components models). Eliminating η from the full expressions yields

$$\zeta^{yz} = \kappa_W + \frac{c_{W1}}{c_{B1}}(\delta^{yz} - \kappa_B).$$

Since all variance parameters are non-negative, the slope of the relationship is always positive.

We can also simplify the expression for the intercepts. That for ζ^{yz} is $\kappa_W - \frac{c_{W1}}{c_{B1}}\kappa_B$ while the δ^{yz} intercept is $\kappa_B - \frac{c_{B1}}{c_{W1}}\kappa_W$. The point (κ_B, κ_W) is the "center" of the line, corresponding to $\eta = 0$, which also identifies the quadrant in the plane from which the line of plausible confounders emanates. The expression for κ_B can be simplified.

$$\kappa_B = -\frac{c_{B0}c_{B1}\Delta_B}{c_{B0} - c_{B1}}.$$

We can assume $c_{B0} > c_{B1}$, in which case $\text{sign}(\kappa_B) = -\text{sign}(\Delta_B)$.

The expression for κ_W is more complicated.

$$\kappa_W = \kappa^* \left\{ (c_{B1} + c_{W1})((c_{B0} + c_{W0})\Delta_O - c_{W0}\Delta_W) - \frac{c_{B0}c_{B1}(c_{B0} - c_{B1} + c_{W0} - c_{W1})}{c_{B0} - c_{B1}}\Delta_B \right\}.$$

where $\kappa^* = \frac{c_{W1}}{c_{B0}c_{W1} - c_{B1}c_{W0}}$. We can assume $c_{W0} > c_{W1}$ and $c_{B0} > c_{B1}$, in which case differences between variance parameters, such as $c_{B0} - c_{B1}$ will all be positive. While the sign of the main expression is not a simple expression, it is determined by the signs and relative magnitudes of the bias differences. The sign of κ^* will be positive when $\frac{c_{W1}}{c_{W0}} > \frac{c_{B1}}{c_{B0}}$, otherwise the sign flips. The equation tracks the relative change in within and between group variation with the addition of predictors, with the more common situation being relatively less within variation being explained. Some insight may be gained by noting that the terms preceding Δ_O are likely to be large in magnitude in comparison to the "weights" for Δ_W and Δ_B, so the sign of κ_W is largely a function of the sign of the κ^* and Δ_O. A practical implication of this is that when κ_B is negative, implying that $\Delta_B > 0$, only a very constrained set of situations would allow κ_W to be positive.

References

Allison, P. D. (2006). *Fixed effects regression methods in SAS*. SAS Institute.

Altonji, J. G. and Mansfield, R. K. (2011). The role of family, school, and community characteristics in inequality in education and labor market outcomes. In Duncan, G. J. and Murnane, R. J., editors, *Whither opportunity?: Rising inequality, schools, and children's life chances*, pages 339–358. Russell Sage Foundation.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press, Princeton.

Carnegie, N. B., Harada, M., and Hill, J. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9:395–420.

Clark, T. S. and Linzer, D. A. (2012). Should I use fixed or random effects. *Working Paper #1315. Washington University, St Louis.*

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203.

Donner, A. and Koval, J. J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics*, pages 19–25.

Dorie, V., Carnegie, N. B., Harada, M., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics and Medicine*, 35:3453–3470.

Enders, C. K. and Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological methods*, 12(2):121.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85(4):907–920.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Greene, W. H. (2003). *Econometric analysis*. Prentice Hall.

Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, 25(6):1107–1116.

Griliches, Z. and Hausman, J. A. (1986). Errors in variables in panel data. *Journal of econometrics*, 31(1):93–118.

Gustafson, P. (2015). *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*, volume 140. CRC Press.

Harada, M. (2013). Generalized sensitivity analysis. Technical report, New York University, New York, NY.

Hedges, L. V. and Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1):60–87.

Hedges, L. V., Hedberg, E. C., et al. (2007). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education*, 22(10):1–15.

Hill, J. (2013). Causal inference and multilevel models. In Scott, M. A., Simonoff, J. S., and Marx, B. D., editors, *The SAGE Handbook of Multilevel Modeling*, chapter 12, pages 201–219. SAGE, London.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, pages 126–132.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, pages 948–963.

McCandless, L. C., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in medicine*, 26(11):2331–2347.

Middleton, J. A., Scott, M. A., Diakow, R., and Hill, J. L. (2016). Bias amplification and bias unmasking. *Political Analysis*, 24:307–323.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54:726–731.

Peaker, G. F. (1975). An empirical study of education in twenty-one countries: A technical report. International Studies in Evaluation VIII.

Pearl, J. (2012). On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*.

Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Education Finance and Policy*, 4(4):468–491.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. (2010). Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association*, 105:692–702.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 212–218.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.*, 6(1):34–58. Available from: https://doi.org/10.1214/aos/1176344064.

Schisterman, E. F., Cole, S. R., and Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4):488–495.

Scott, M. A., Shrout, P. E., and Weinberg, S. L. (2013). Multilevel model notation - establishing the commonalities. In Scott, M. A., Simonoff, J. S., and Marx, B. D., editors, *The SAGE Handbook of Multilevel Modeling*, chapter 2, pages 21–38. SAGE, London.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Singer, J. D. (1998). Using sas proc mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of educational and behavioral statistics*, 23(4):323–355.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford University Press, New York.

Thompson, D. M., Fernald, D. H., and Mold, J. W. (2012). Intraclass correlation coefficients typical of cluster-randomized studies: estimates from the robert wood johnson prescription for health projects. *The Annals of Family Medicine*, 10(3):235–240.

Townsend, Z., Buckley, J., Harada, M., and Scott, M. A. (2013). The choice between fixed and random effects. In Scott, M. A., Simonoff, J. S., and Marx, B. D., editors, *The SAGE Handbook of Multilevel Modeling*, chapter 5, pages 73–88. SAGE, London.

VanderWeele, T. J. and Arah, O. A. (2011). Unmeasured confounding for general outcomes, treatments, and confounders: Bias formulas for sensitivity analysis. *Epidemiology*, 22(1):42–52.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.