Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Cluster-Specific Estimands: Identification and Uncertainty in Multilevel Analysis

Jee-Seon Kim [*,†], Graham Buhrman [†], and Xiangyi Liao [‡]

†Department of Educational Psychology, University of Wisconsin, Madison, Wisconsin, USA
‡Educational and Counselling Psychology, and Special Education, University of British Columbia, Vancouver, Canada
*Corresponding author. Email: jeeseonkim@wisc.edu

Abstract

In multilevel studies, there is growing interest in cluster-specific effects, such as site- or school-specific treatment effects, to understand contextual variation and to inform local decision-making, implementation heterogeneity, and context-specific evaluation. Yet it is often unclear which estimands best support cluster-specific inference for particular aims and designs, under what assumptions they are identified, and how uncertainty should be reflected in their interpretation. This paper provides an estimand-centered clarification of cluster-specific causal quantities in multilevel analysis. We organize the argument around three components of inference. First, we define cluster-specific estimands and distinguish them from marginal and cluster-average targets, emphasizing that many common analytic workflows implicitly switch estimands while keeping interpretation fixed. Second, we discuss identification of cluster-specific estimands in randomized and observational multilevel designs, highlighting that while such estimands may be formally identified under both designs, the strength and interpretability of cluster-specific inference depend critically on within-cluster information and modeling assumptions. Third, we discuss inference quality and characterize uncertainty in cluster-specific estimands as arising from partial pooling, limited and uneven within-cluster information, and features of the variance structure, including between-cluster variability and within-cluster heteroskedasticity. Taken together, these clarifications aim to align estimand choice, identification assumptions, and uncertainty quantification, and to support more transparent and defensible cluster-level inference in multilevel research.

Keywords: multilevel causal inference, treatment effect heterogeneity, inference uncertainty, partial pooling, shrinkage

1. Introduction

Multilevel data structures are central to research in education, psychology, health, and the social sciences, where individuals are nested within schools, classrooms, therapists, clinics, programs, or geographic units (Gelman & Hill, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 2011). In these settings, analysts frequently move beyond a single population-average treatment effect and ask how effects vary across higher-level units. The motivation is often practical. Local decision-makers want to understand whether an intervention appears beneficial in a particular site, whether some sites appear to benefit more than others, and what contextual features might explain differences. These goals align with broader trends in causal inference and heterogeneous treatment effects that seek to characterize variation rather than only averages (Athey & Wager, 2019; Chernozhukov et al., 2018; Hahn et al., 2020; Künzel et al., 2019).

At the same time, cluster-specific effect reporting raises a recurring interpretational challenge. Analysts often present cluster-specific estimates as if they were well-identified causal effects for each cluster, and then use them for comparison, ranking, or decision-making. In practice, many

cluster-specific effects are inherently heterogeneous but only weakly supported by the data, leading reported estimates to be heavily regularized.

Even in randomized settings, the information available within many clusters is limited, so cluster-specific estimates depend strongly on pooling and modeling assumptions, which can dominate the signal. In observational settings, cluster-specific causal interpretations rest on identification assumptions that are typically stronger than those required for marginal effects, particularly when cluster-level confounding is present or within-cluster overlap is limited (Fuentes et al., 2021; Kim & Frees, 2007; Kim et al., 2015).

This paper argues that progress on cluster-specific causal claims requires an explicit focus on *estimands*, *identification*, and *uncertainty*. This emphasis is consistent with modern causal inference practice in which estimands and assumptions are treated as first-class scientific objects (Bang & Robins, 2005; Robins, 2000; Robins et al., 1994). Our goal is not to propose new estimators. Instead, we provide an inferential clarification that aims to make cluster-specific conclusions more transparent, more defensible, and better aligned with the information content of multilevel data. We also aim to separate two activities that are frequently conflated: (i) constructing stable *predictions* of cluster effects using partial pooling, and (ii) making *inferential statements* about cluster-specific causal estimands.

The central message is that *formal identification does not imply informative estimation*, particularly for cluster-specific estimands. Cluster-specific estimands may be formally identified under standard assumptions, yet still be weakly informed by the observed data, in the sense that estimation uncertainty is large and conclusions depend strongly on pooling and variance assumptions.

2. Cluster-Specific Estimands

2.1 Defining cluster-specific causal estimands

We adopt the potential outcomes framework (Neyman, 1923; Rubin, 1974) and its extension to multilevel data structures in which individuals are nested within clusters (Hong & Raudenbush, 2006; Lyu et al., 2023). Let clusters be indexed by $j = 1, \ldots, J$, and individuals within cluster j by $i = 1, \ldots, n_j$. Let $A_{ij} \in \{0, 1\}$ denote treatment assignment or exposure, where $A_{ij} = 1$ indicates receipt of the intervention and $A_{ij} = 0$ indicates control or baseline status; treatment may be assigned at either the cluster or individual level, and unless otherwise noted we focus on settings with individual-level assignment within clusters. Let $Y_{ij}(a)$ denote the potential outcome for individual i in cluster j under treatment level $a \in \{0, 1\}$.

A *cluster-specific estimand* is defined as

$$\tau_j = E\left[Y_{ij}(1) - Y_{ij}(0) \mid j\right], \quad j = 1, \ldots, J, \tag{1}$$

which represents the average causal contrast for individuals within cluster j.

In this paper, a *cluster-specific estimand* refers to a formally defined target quantity in the potential outcomes framework. The cluster-specific estimand is given by Eq. (1). We use the term *cluster-specific effect* as a substantive label for this estimand, as shorthand once the estimand has been defined. We distinguish the estimand itself from *model-based estimates or predictions*, such as empirical Bayes or posterior summaries produced by hierarchical models.

Because cluster-specific estimands are conditional on cluster membership, they constitute conditional average treatment effects, with the cluster indicator serving as the conditioning variable. Unlike individual-level conditional average treatment effects, cluster-specific estimands are often supported by limited within-cluster information in multilevel studies and therefore rely more heavily on pooling across clusters. They also differ fundamentally from marginal estimands that average over individuals or clusters. Cluster-specific estimands are most commonly targeted in multilevel random-slope models, where treatment effects are allowed to vary across clusters (Kim et al., 2024).

Two points are central for interpretation. First, the conditioning event matters: conditioning on membership in an observed cluster is conceptually distinct from conditioning on an unobserved

random effect representing latent cluster propensity, yet applied interpretations often blur these distinctions. Second, cluster-specific effects are used in two distinct ways: as scientific targets for describing site-to-site variation, and as operational targets for predicting local impacts. These uses can coincide, but they do not coincide automatically.

2.2 Marginal and cluster-level estimands

A central source of confusion in multilevel analyses is estimand mismatch. Applied papers often move implicitly among several causal targets while maintaining a fixed interpretation. In multilevel settings, three estimands are especially common.

Marginal individual-average effect (average treatment effect). The *marginal individual-average effect*, commonly referred to as the *average treatment effect*, is defined as

$$\tau^{\mathrm{ATE}} = E\left[Y_{ij}(1) - Y_{ij}(0)\right],$$

where the expectation is taken over individuals in the target population. This estimand averages individual-level causal contrasts and is therefore marginal with respect to cluster membership. In multilevel data, it implicitly weights clusters in proportion to their population or sample sizes.

In this paper, the marginal individual-average effect is identical to the average treatment effect because treatment is binary and the estimand is defined directly in terms of potential outcomes; this equivalence does not extend to model-based marginal effects derived from nonlinear specifications. As a result, marginal individual-average effects should not be interpreted interchangeably with regression-based marginal effects without additional assumptions.

Cluster-average effect. A *cluster-average effect* assigns equal weight to clusters rather than individuals:

$$\tau^{\mathrm{CA}} = \frac{1}{J}\sum_{j=1}^{J} E\left[Y_{ij}(1) - Y_{ij}(0) \mid j\right].$$

This estimand is marginal over clusters and yields a single population-level summary in which each cluster contributes equally. Although it is defined in terms of cluster-specific effects, it is not indexed by j because it averages over clusters rather than conditioning on them. Conditioning on a specific cluster instead yields the cluster-specific effect τ_j, which is a conditional average treatment effect.

The distinction between marginal individual-average and cluster-average estimands becomes consequential when cluster sizes vary or when cluster size is informative, as the two estimands correspond to different weighting schemes and may lead to different substantive conclusions. Table 1 summarizes these three estimands and their defining features.

2.3 What hierarchical models "estimate" and why that matters

Hierarchical outcome models are often used because they stabilize cluster-specific estimates by borrowing strength across clusters. This stabilization has two closely related consequences.

First, reported cluster-specific quantities, under standard hierarchical specifications, correspond to *predictions* (such as empirical Bayes posterior means or modes) of latent cluster effects under the assumed model, rather than unbiased estimators of the cluster-specific estimand τ_j in the frequentist sense. More precisely, empirical Bayes (EB) or posterior summaries target model-defined latent parameters (e.g., β_j) rather than the causal estimand τ_j. These quantities coincide with τ_j only under correct model specification and alignment between the model and the causal estimand. Thus, the issue is not bias of EB estimators per se, but a mismatch between model-defined parameters and causal estimands.

Table 1. Marginal, cluster-average, and cluster-specific estimands in multilevel designs.

Estimand and definition	Unit of averaging	Interpretation
Marginal individual-average effect (average treatment effect) $\tau^{\mathrm{ATE}} = E[Y_{ij}(1) - Y_{ij}(0)]$	Individuals	Average causal effect across individuals; clusters weighted by their contribution to the individual-level population
Cluster-average effect $\tau^{\mathrm{CA}} = \frac{1}{J}\sum_{j=1}^{J} E[Y_{ij}(1) - Y_{ij}(0) \mid j]$	Clusters	Average causal effect across clusters; each cluster contributes equally
Cluster-specific effect $\tau_j = E[Y_{ij}(1) - Y_{ij}(0) \mid j]$	Individuals within cluster j	Conditional average treatment effect for a specific cluster

Note. Cluster-specific effects are conditional average treatment effects with cluster membership as the conditioning variable. Cluster-average effects are marginal over clusters and therefore are not indexed by j. The marginal individual-average and cluster-average estimands coincide only under special conditions, such as equal cluster sizes.

Second, shrinkage is not incidental: it is the mechanism that produces stability when within-cluster information is limited. Interpretation of cluster-specific estimates must therefore treat pooling as a defining feature of the estimand–estimator relationship, not as a minor technical detail. This view aligns with the classical empirical Bayes perspective on shrinkage, in which stabilization arises from borrowing strength across related units under a shared distributional structure rather than from design-based information alone (Efron & Morris, 1977; Morris, 1983).

The distinction between cluster-specific causal estimands and their model-based predictions has direct empirical implications. In multisite datasets with moderate cluster sizes, which are commonly encountered in practice, unpooled site-by-site estimates of cluster-specific effects often exhibit substantial variability, particularly in clusters with small sample sizes or high outcome variance. Hierarchical models stabilize these quantities through partial pooling across clusters, with the strongest regularization occurring where within-cluster information is most limited.

The following section examines how these distinctions shape identification in randomized and observational multilevel designs.

3. Identification of Cluster-Specific Estimands

Identification concerns whether a causal estimand is uniquely determined by the observed data distribution under a stated set of assumptions. It is therefore a property of the estimand–assumption pair, not of an estimator or model.

This estimand-first framing aligns with modern causal inference treatments that emphasize explicit causal targets and assumptions in observational studies (Hernán & Robins, 2020). A related distinction appears in econometrics, where parameters may be formally identified yet only weakly informative in finite samples, motivating partial identification and sensitivity-based approaches (Manski, 2007). Throughout this section, we distinguish formal identification from the practical informativeness of the data for estimating an estimand.

3.1 Randomized multisite and cluster-randomized designs

In randomized experiments, identification of the marginal individual-average effect (the average treatment effect) follows directly from randomization under standard assumptions (Shadish & Steiner, 2010; Thoemmes & West, 2011). Cluster-randomized trials and multisite randomized designs extend these principles to multilevel settings. Design-based analyses of cluster-randomized experiments formalize this identification logic and clarify which causal estimands are supported by randomization alone (Imai et al., 2009).

Under cluster randomization, the cluster-specific estimand τ_j defined in Eq. (1) is identified in the sense that, under an asymptotic regime with infinitely many units per cluster, each τ_j could be estimated consistently from within-cluster contrasts alone. This notion of identification is theoretical and asymptotic, and it does not guarantee that the data observed in finite-sample applications contain sufficient information to support precise or stable estimation of cluster-specific effects.

In finite samples, many clusters contain relatively few observed units, even in randomized settings. As a result, the effective information available for estimating cluster-specific estimands can be limited, and uncertainty can be substantial. Stable reporting of cluster-specific effects therefore often relies on partial pooling across clusters through hierarchical models (Raudenbush & Bryk, 2002; Snijders & Bosker, 2011). Thus, while identification holds under randomization, the practical inferential question is which cluster-level conclusions are justified given limited within-cluster information and features of the variance structure.

3.2 Observational multilevel data

In observational multilevel settings, identification of causal estimands requires assumptions such as ignorability given observed covariates and sufficient overlap. Design-based perspectives emphasize that the choice of estimand determines which assumptions are required for identification and how sensitive conclusions are to violations of those assumptions (Aronow & Samii, 2017). When treatment varies within clusters, within-cluster strategies and fixed-effects approaches can mitigate bias from time-invariant cluster-level confounding. This includes settings where treatment is assigned at the individual level but is highly imbalanced across clusters in the observed data, so that effective variation is limited within some clusters. These approaches require adequate within-cluster treatment variation and overlap to support identification of cluster-specific effects (Fuentes et al., 2021; Steiner et al., 2012).

When treatment assignment occurs primarily between clusters, either by design or effectively due to imbalance in individual-level assignment across clusters, cluster membership and treatment are closely intertwined. In such settings, cluster-specific causal interpretations become more model dependent, because within-cluster counterfactual contrasts are weakly supported by the data. The result is that identification of cluster-specific estimands relies more heavily on modeling assumptions than identification of marginal individual-average effects. In such cases, even with individual-level assignment, the effective variation relevant for identifying cluster-specific effects may be predominantly between clusters.

Modern doubly robust and semiparametric methods provide tools for identifying and estimating marginal individual-average effects under weaker modeling dependence (Bang & Robins, 2005; Robins, 2000; Robins et al., 1994). In contrast, cluster-specific estimands reintroduce stronger dependence on modeling assumptions because they require estimation of cluster-specific causal functions and their distribution. Methods developed for heterogeneous treatment effects face analogous challenges: conditional average treatment effects can be defined, but reliable estimation depends on adequate effective sample size and overlap in the relevant subpopulations (Athey & Wager, 2019; Chernozhukov et al., 2018; Künzel et al., 2019).

The following section examines how these constraints on within-cluster information interact with pooling and variance structure to shape uncertainty in cluster-specific inference.

3.3 Limited and uneven within-cluster information

A central constraint for cluster-specific inference is the limited information available within clusters. This can arise from small cluster sizes, imbalance in treatment assignment within clusters, attrition patterns that differ by cluster, and high within-cluster outcome variance.

Under such conditions, estimates of cluster-specific effects are often driven more by the model's pooling structure than by within-cluster contrasts alone. For example, a standardized effect size derived from a heavily shrunk empirical Bayes estimate may appear precise and comparable across clusters, even when the underlying within-cluster evidence is sparse and the numerator is largely model-driven.

This limitation is estimand-relative. The same dataset may provide strong information for identifying and estimating a marginal individual-average effect while providing weak support for cluster-specific estimands. An estimand-centered perspective clarifies this distinction and helps avoid overinterpretation of cluster-specific reporting that exceeds what the study design and data can support.

4. Uncertainty in Cluster-Specific Inference

Uncertainty in cluster-specific inference arises from multiple sources beyond sampling variability alone (Frees & Kim, 2006; Kim & Frees, 2007). In this section, we focus on three interacting components: pooling, within-cluster information, and variance structure. Together, these components determine how stable cluster-specific estimates are and how sensitive they are to modeling choices.

4.1 Pooling and shrinkage as inferential structure

Pooling refers to the sharing of information across clusters when estimating cluster-specific effects. Partial pooling occupies an intermediate position between no pooling, in which clusters are analyzed independently, and complete pooling, in which all clusters are assumed to share a common effect.

In hierarchical models, partial pooling is implemented through model structure rather than an explicit penalty. Cluster-specific effects are treated as arising from a shared distribution, and estimates are *regularized* toward that distribution to an extent determined by the data. Here, regularization refers to the model-imposed constraint that limits how far cluster-specific estimates can deviate from values implied by the cross-cluster distribution, particularly when within-cluster information is limited.

From an inferential perspective, pooling is more than variance reduction. It replaces missing or weak within-cluster information with cross-cluster regularity. Clusters with limited information are shrunk more strongly, while clusters with substantial information are shrunk less. As a result, reported cluster-specific quantities are model-based summaries rather than direct within-cluster contrasts. Related distinctions arise in individual-level heterogeneous treatment effect estimation, where regularized learners can produce stable predictions even when local causal effects are weakly identified (Wager & Athey, 2018).

This has direct implications for interpretation. Differences among partially pooled estimates should not be interpreted as direct or unqualified evidence of heterogeneity without accounting for the role of pooling and the uncertainty it induces. Shrinkage both stabilizes and compresses variation, and apparent patterns can be sensitive to assumptions about the variance structure (Herrmann et al., 2016; Hui et al., 2021). Similar issues arise in education accountability and value-added modeling, where ranking and inference can be sensitive to the extent of pooling and to variance assumptions (Kane & Staiger, 2008).

4.2 Variance structure and heteroskedasticity

Variance structure plays a central role in cluster-specific inference because it determines the effective signal available to distinguish clusters. Between-cluster variability governs the scale of heterogeneity,

while within-cluster variability determines precision. When within-cluster variances differ across clusters, heteroskedasticity implies that some clusters contain substantially less information than others, leading to uneven shrinkage and uneven uncertainty (Leckie et al., 2024).

Misspecification of the variance structure is particularly consequential when clusters are few or when within-cluster information is limited. Under such conditions, variance components themselves are estimated imprecisely, and uncertainty about variance components propagates directly to cluster-specific estimates. Although this issue is implicit in multilevel modeling, it is often omitted from applied interpretations that treat variance estimates as fixed and known.

4.3 Weak identification and unstable rankings

When cluster-specific estimands are weakly identified or weakly supported by the data, instability in estimated effects arises from two interacting sources: (i) limited within-cluster information, which increases sampling variability and induces stronger shrinkage, and (ii) dependence on modeling and variance assumptions, which can alter the degree and direction of pooling. As a result, rankings based on estimated effects can be difficult to interpret, even in the presence of genuine underlying heterogeneity. Rankings amplify noise and can be highly unstable under small perturbations of modeling assumptions, variance structure, or pooling strength. Even with moderate sample sizes, alternative variance structures or pooling assumptions can lead to different rankings. This concern has a long history in statistics, particularly in critiques of league tables and institutional comparisons under uncertainty (Goldstein & Spiegelhalter, 1996), and parallels more recent discussions of instability in ranking and targeting based on estimated effects (Kleinberg et al., 2015).

The inferential lesson is that cluster-specific reporting should emphasize uncertainty, robustness, and interpretability, and should be cautious about league-table style conclusions when cluster-specific effects are heavily regularized. Rankings and comparative statements are most fragile precisely where inferential support is weakest—in small clusters and in the tails of the effect distribution—underscoring that instability is an expected consequence of limited information.

Taken together, randomized and observational multilevel designs support cluster-specific estimands under different identification regimes, but neither guarantees informative cluster-specific inference. Randomization removes confounding by design but often leaves cluster-specific effects weakly supported by within-cluster data. Observational designs introduce additional sensitivity to confounding and model specification. In both cases, the inferential challenge is not whether cluster-specific estimands can be defined, but whether they can be interpreted beyond heavily regularized summaries. Table 2 summarizes how identification, dominant sources of uncertainty, and interpretational risks differ for cluster-specific estimands across randomized and observational multilevel designs.

Table 2. Cluster-specific estimands, identification, and uncertainty by design.

	Randomized multilevel designs	**Observational multilevel designs**
Estimand definition	Well-defined by design	Well-defined conceptually
Identification	By randomization	By ignorability and overlap
Primary limitation	Limited within-cluster information	Confounding and limited within-cluster information
Role of pooling	Central for stabilization	Central and often more influential
Variance structure	Governs shrinkage and signal	Governs shrinkage and model sensitivity
Interpretational risk	Overinterpreting regularized summaries	Overinterpreting model-dependent summaries

5. Implications for Practice and Reporting

An estimand-centered perspective on cluster-specific inference has several implications for applied research practice and for how results are reported and interpreted. These implications do not prescribe specific estimation methods. Rather, they clarify how inferential claims about cluster-specific effects should be aligned with the information content of multilevel data and the assumptions under which analyses are conducted.

5.1 Clarifying the inferential target

First, analysts should state explicitly which causal estimand is being targeted. In multilevel settings, this requires distinguishing among marginal individual-average effects, cluster-average effects, and cluster-specific effects, and clarifying how clusters are weighted when sizes vary. Because hierarchical models can yield quantities corresponding to different estimands depending on specification and interpretation, failure to articulate the target estimand invites implicit shifts in interpretation. Explicit estimand statements help align scientific aims, modeling choices, and substantive conclusions.

5.2 Separating prediction from causal inference

Second, analysts should distinguish clearly between predictive summaries and inferential claims. Cluster-specific predictions produced by hierarchical models, such as empirical Bayes or posterior mean estimates, can be valuable for descriptive monitoring, exploratory analysis, and forecasting under the assumed model. However, these quantities should not automatically be interpreted as precise estimates of cluster-specific causal effects. When cluster-specific estimands are weakly supported by the data, predictive stability may coexist with inferential fragility.

Predictive stability arises from borrowing strength across clusters through partial pooling, whereas inferential strength about cluster-specific causal estimands depends on the amount and balance of within-cluster information. Conflating these roles risks overstating the evidentiary strength of cluster-specific conclusions. This distinction does not diminish the value of hierarchical models for prediction or description; rather, it clarifies the inferential claims such models can and cannot support.

5.3 Reporting uncertainty consistent with the estimand

Third, uncertainty should be reported in a manner that reflects the estimand and inferential task. For cluster-specific effects, uncertainty arises not only from sampling variability but also from partial pooling, limited and uneven within-cluster information, and imprecision in variance components. Reporting practices that emphasize point estimates or rankings without conveying these sources of uncertainty can be misleading. Interval estimates, posterior distributions, and sensitivity analyses provide more appropriate summaries when cluster-specific inference is of interest.

In practice, cluster-specific effects are often reported as standardized effect sizes. When estimates are heavily regularized or weakly supported by within-cluster information, such standardization can obscure the extent of uncertainty and may induce overinterpretation of differences across clusters. For example, a standardized effect size derived from a heavily shrunk empirical Bayes estimate may appear precise and comparable across clusters, even when the underlying within-cluster evidence is sparse and the numerator is largely comparing cluster-specific effects. Interpretation should therefore consider both scale and uncertainty when comparing cluster-specific effects.

5.4 Caution in ranking and comparative use

Fourth, cluster rankings warrant particular caution: rankings are unstable when cluster-specific effects are weakly identified, particularly in the tails of the distribution. In these regions, increased distance from the center and limited availability of neighboring clusters reduce local information,

strengthening the influence of regularization. Small changes in pooling strength, variance assumptions, or model specification can lead to substantial reordering. As a result, rankings are best treated as exploratory summaries and should be interpreted with care when inferential support is limited.

5.5 Design-aware interpretation

Finally, interpretation should be explicitly design-aware. In randomized multilevel designs, randomization supports identification of cluster-specific estimands but does not eliminate uncertainty due to limited within-cluster information. In observational designs, additional uncertainty arises from reliance on ignorability and overlap assumptions, particularly when treatment variation is primarily between clusters. Recognizing how design features shape both identification and uncertainty helps prevent overgeneralization of cluster-specific findings. As summarized in Table 2, both randomized and observational multilevel designs can support cluster-specific estimands, but neither guarantees informative cluster-specific inference, and the dominant sources of uncertainty differ by design.

6. Discussion and Path Forward

This paper has argued that cluster-specific effects are best understood as causal estimands whose inferential meaning depends jointly on identification assumptions and multiple sources of uncertainty intrinsic to multilevel data. By organizing the analysis around estimands, identification, and uncertainty, we have emphasized that the central challenge in cluster-specific inference is not whether heterogeneity exists—it often does—but how much of that heterogeneity can be learned from the data under a given design, taking uncertainty and pooling into account. The design-specific tradeoffs underlying this challenge are summarized in Table 2, which highlights that cluster-specific estimands face distinct but related inferential limitations in randomized and observational multilevel designs.

A central message is that cluster-specific estimands pose similar inferential challenges across randomized and observational multilevel designs, even though the sources of those challenges differ. Randomized designs eliminate confounding by construction, yet often provide limited within-cluster information for estimating cluster-specific effects with precision. Observational designs introduce additional sensitivity to confounding and model specification, making cluster-specific estimands more dependent on assumptions. In both settings, partial pooling and variance structure play defining roles in shaping what can be inferred.

Importantly, emphasizing these limitations does not imply that hierarchical models or cluster-specific analyses are uninformative or misguided. Hierarchical models remain powerful tools for stabilizing estimates, describing patterns of variation, and supporting prediction under clearly stated modeling assumptions. The contribution of this paper is not to question the utility of these methods, but to clarify the inferential claims that cluster-specific quantities can reasonably support. Cluster-specific estimates are most informative when interpreted as regularized summaries with explicit uncertainty, rather than as precise causal contrasts for individual clusters.

Several directions for future work follow from this analysis. One direction is the development of diagnostic tools that help assess when cluster-specific estimands are likely to be weakly supported by the data and how sensitive conclusions are to pooling and variance assumptions. Another direction is the integration of estimand-based reasoning into reporting standards for multilevel studies, particularly in applied fields where cluster-specific findings are increasingly used for accountability and decision-making. Finally, further work is needed to bridge inferential, predictive, and decision-theoretic perspectives, clarifying when and how cluster-specific predictions can inform policy choices under uncertainty. This includes further development of flexible learners for heterogeneous treatment effects, such as Bayesian additive regression trees and causal forests, which offer a natural framework for integrating prediction and causal inference under complex heterogeneity (Hahn et al., 2020; Kleinberg et al., 2015; Wager & Athey, 2018). These directions connect to a broader taxonomy of

treatment effect heterogeneity that distinguishes among types of variation by observability, structural form, and inferential role (Kim et al., 2025; Lyu et al., 2023).

In sum, cluster-specific effects occupy an important but delicate position in multilevel causal analysis. Treating them explicitly as estimands, rather than as automatically interpretable quantities, helps align analytic practice with inferential goals and supports more transparent and defensible use of multilevel evidence.

Acknowledgement

Funding Statement Support for this research was provided by Graduate Education at the University of Wisconsin– Madison with funding from the Wisconsin Alumni Research Foundation.

Competing Interests None.

References

Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference. *Annals of Applied Statistics*, *11*(4), 1912–1947.

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, *5*(2), 37–51.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127.

Frees, E. W., & Kim, J.-S. (2006). Multilevel model prediction. *Psychometrika*, *71*(1), 79–104.

Fuentes, A., Lüdtke, O., & Robitzsch, A. (2021). Causal inference with multilevel data: A comparison of different propensity score weighting approaches. *Multivariate Behavioral Research*, 1–24.

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

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*(3), 385–443.

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, *15*(3), 965–1056. https://doi.org/10.1214/19-BA1195

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

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy*, *3*(1), 1–10.

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

Hui, F. K., Müller, S., & Welsh, A. H. (2021). Random effects misspecification can have severe consequences for random effects inference in linear mixed models. *International Statistical Review*, *89*(1), 186–206.

Imai, K., King, G., & Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, *24*(1), 29–53.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. In D. N. Harris & T. R. Sass (Eds.), *The econometrics of education* (pp. 187–224). Elsevier.

Kim, J.-S., Buhrman, G., & Liao, X. (2025). Beyond averages: Portraying treatment effect variation. *Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic, 2024.* https://doi.org/10.64028/xvhr927961

Kim, J.-S., & Frees, E. W. (2007). Multilevel modeling with correlated effects. *Psychometrika*, *72*(4), 505–533.

Kim, J.-S., Liao, X., & Loh, W. W. (2024). Assessing cross-level interactions in clustered data using CATE estimation methods. *Quantitative Psychology: The 88th Annual Meeting of the Psychometric Society, Maryland, USA*, 87–97. https://doi.org/10.1007/978-3-031-55548-0_10

Kim, J.-S., Steiner, P. M., & Lim, W.-C. (2015). Mixture modeling methods for causal inference with multilevel data. *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*, 335–359.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–495.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Leckie, G., Parker, R., Goldstein, H., & Tilling, K. (2024). Mixed-effects location scale models for joint modeling school value-added effects on the mean and variance of student achievement. *Journal of Educational and Behavioral Statistics*, *49*(6), 879–911.

Lyu, W., Kim, J.-S., & Suk, Y. (2023). Estimating heterogeneous treatment effects within latent class multilevel models: A bayesian approach. *Journal of Educational and Behavioral Statistics*, *48*(1), 3–36.

Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.

Morris, C. N. (1983). Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*(381), 47–55.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science*, 465–472.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association*, *1999*, 6–10.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846–866.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis [Quantitative Research Methodology]. *Newborn and Infant Nursing Reviews*, *10*(1), 19–26. https://doi.org/10.1053/j.nainr.2009.12.010

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Steiner, P. M., Kim, J.-S., & Thoemmes, F. (2012). Matching strategies for observational multilevel data. *Joint Statistical Meeting Proceedings, Social Statistics Section*, 5020–5032.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*(3), 514–543.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.