Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Beyond Averages: Portraying Treatment Effect Variation

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

This paper addresses distinct forms of treatment effect variation that commonly arise in social and behavioral science research. Differences in the nature of relationships among variables or in the contexts where treatments are implemented can lead to both quantitatively and qualitatively different patterns of treatment effect heterogeneity. Such variation may involve interactions between treatments and covariates, conditional average treatment effects defined by observed characteristics, random treatment coefficients across clusters, or differences in treatment effects across unobserved subpopulations. By highlighting distinctive features of treatment effect variation, this paper emphasizes the importance of addressing heterogeneity as an integral part of study design and research objectives, rather than treating it as a secondary or post hoc concern. This paper concludes by emphasizing the need for a structured and conceptually grounded framework to better identify, interpret, and apply heterogeneous treatment effects.

Keywords: Causal Inference, Heterogeneous Treatment Effects, Causal Machine Learning, Conditional Average Treatment Effect (CATE), Random Slope Models, Latent Class Analysis, Unobserved Heterogeneity

1. Introduction

1.1 Potential Outcomes Framework for Multilevel Data

Treatment effects are commonly defined using the potential outcomes framework (Neyman, 1923; Rubin, 1974), which provides a formal basis for causal inference. In this paper, we adopt an extension of the potential outcomes framework appropriate for multilevel data structures, where individuals are nested within clusters (Hong & Raudenbush, 2006; Lyu et al., 2022).

Let $i = 1, \ldots, n_j$ index individuals within cluster $j = 1, \ldots, M$, where n_j is the number of individuals in cluster j, and the total number of individuals across all clusters is $N = \sum_{j=1}^{M} n_j$. For each individual i in cluster j, let $T_{ij} \in \{0, 1\}$ denote the binary treatment indicator, where $T_{ij} = 1$ if the individual receives the treatment and $T_{ij} = 0$ otherwise. Each individual has two potential outcomes: $Y_{ij}(1)$, representing the outcome if the individual receives the treatment, and $Y_{ij}(0)$, representing the outcome under the control condition. The observed outcome Y_{ij} is determined by the treatment assignment and the corresponding potential outcomes:

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

This framework relies on the *stable unit treatment value assumption* (SUTVA) (Rubin, 1986), which includes two components: (1) no interference between units (i.e., one individual's potential outcomes do not depend on the treatment assignments of others), and (2) no hidden versions of treatment (i.e.,

treatment is consistently defined and delivered across units). Hong and Raudenbush (2006), Imbens and Rubin (2015), and Kim et al. (2015) are among those who extend SUTVA to multilevel contexts, noting that interactions among units within clusters may be more likely in such settings.

Since only one of the two potential outcomes is observed for each individual, the individual treatment effect $Y_{ij}(1) - Y_{ij}(0)$ is fundamentally unobservable. This is known as the *fundamental problem of causal inference* (Holland, 1986). However, when certain conditions or assumptions hold, researchers can estimate average treatment effects at various levels, for example, population average treatment effects, subgroup-specific effects, or cluster-level effects. These forms of treatment effect estimation will be discussed in the following sections.

1.2 Key Assumptions for Causal Inference

Causal inference from observational multilevel data requires several identifying assumptions to support valid estimation of treatment effects. These assumptions address the fundamental challenge that only one potential outcome is observed for each unit.

Let $\boldsymbol{X}_{ij}$ denote a vector of observed individual-level (Level-1) covariates for individual i in cluster j (e.g., demographic characteristics, prior achievement, baseline health status), and let $\boldsymbol{Z}_j$ denote a vector of observed cluster-level (Level-2) covariates for cluster j (e.g., school size, neighborhood poverty rate, average clinic staffing). Identification of causal effects under the potential outcomes framework relies on the following assumptions:

Unconfoundedness (Conditional Ignorability): Treatment assignment is assumed to be independent of the potential outcomes, conditional on observed covariates at both levels:

$$\{Y_{ij}(1), Y_{ij}(0)\} \perp T_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_j.$$

This assumption implies that, after adjusting for $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_j$, there are no unmeasured confounders that influence both treatment assignment and the potential outcomes.

Positivity (Overlap): Each individual must have a non-zero probability of receiving either treatment condition, given their observed covariates:

$$0 < \Pr(T_{ij} = 1 \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_j) < 1.$$

This ensures the existence of comparable treated and control individuals across the observed range of covariates.

When these assumptions are plausible, causal effects can be estimated using a range of methods tailored to multilevel data. These include multilevel propensity score modeling (Leite, 2016; Thoemmes & West, 2011), matching within or across clusters (Steiner et al., 2012; Stuart, 2010), and inverse probability weighting (Lunceford & Davidian, 2004). Such approaches aim to reduce confounding by balancing covariates across treatment groups, thereby supporting credible causal inference in clustered or hierarchical data structures.

2. Individual and Average Treatment Effects

Treatment effects in causal inference are often defined within the potential outcomes framework, which provides a formal structure for comparing outcomes under different treatment conditions for the same unit. As outlined above, this framework assumes that each unit (e.g., individual i in cluster j) has a pair of potential outcomes: one under treatment and one under control. The difference between these potential outcomes represents the causal effect of the treatment for that unit. However, because only one of these outcomes is observed for each unit, causal effects must be inferred under additional assumptions.

Two fundamental quantities emerge from this framework: the *Individual Treatment Effect* (ITE) and the *Average Treatment Effect* (ATE). These quantities differ in their conceptual focus, assumptions required for estimation, and practical interpretability in applied research. While the ITE targets unit-level causal impacts, the ATE provides a population-level summary of treatment effects.

2.1 Individual Treatment Effect (ITE)

The individual treatment effect for unit i in cluster j is defined as:

$$ITE_{ij} = Y_{ij}(1) - Y_{ij}(0).$$

This quantity reflects the unit-specific causal effect of treatment but is fundamentally unobservable due to the impossibility of observing both potential outcomes for the same unit. Estimation of the ITE requires strong modeling assumptions, such as functional form restrictions, ignorability, or repeated observations.

2.2 Average Treatment Effect (ATE)

The average treatment effect summarizes the expected causal impact of treatment across the population:

$$ATE = E[Y_{ij}(1) - Y_{ij}(0)].$$

In randomized experiments or well-designed observational studies satisfying unconfoundedness and positivity, the ATE is identifiable and can be estimated using regression, weighting, or matching methods. However, the ATE represents an average across potentially diverse units and may mask substantial variation in effects across individuals, subgroups, or clusters.

The distinction between the ITE and ATE is especially important in multilevel settings, where both within- and between-cluster sources of heterogeneity may influence treatment effects. Recognizing and accounting for this variation is essential for designing targeted interventions, supporting equitable policy decisions, and developing theories that reflect heterogeneity in treatment response.

3. Variation in Treatment Effects

In multilevel data structures, treatment effects may vary both within and between clusters, necessitating flexible modeling strategies to capture individual- and context-level sources of heterogeneity. Below, we describe several common approaches to modeling variation in treatment effects, including interactions with covariates, machine learning–based estimation of conditional average treatment effects (CATE), random slopes, and latent class models.

These approaches offer structured forms of heterogeneity that are less granular than ITEs but more detailed than the ATE. Each rests on distinct assumptions and serves different analytic goals. While many machine learning–based methods treat the ITE as the default estimand from which ATE and CATE are derived, not all approaches require estimation of treatment effects at the individual level. As a result, the variety of methods yields a spectrum of estimands situated between the ITE and the ATE, each offering a distinct summary of treatment effect variation.

3.1 Interaction with Covariates

In multilevel data, treatment effect heterogeneity can be modeled by allowing treatment effects to interact with covariates measured at both the individual (Level-1) and cluster (Level-2) levels. To account for unobserved differences in baseline outcomes across clusters, a random intercept is generally included.

For illustration, the following model includes two Level-1 covariates (X_{1ij}, X_{2ij}), two Level-2 covariates (Z_{1j}, Z_{2j}), and a single interaction between treatment and one individual-level covariate:

$$Y_{ij} = (\gamma_0 + U_{0j}) + \gamma_1 T_{ij} + \gamma_2 X_{1ij} + \gamma_3 X_{2ij} + \gamma_4 (T_{ij} X_{1ij}) + \gamma_5 Z_{1j} + \gamma_6 Z_{2j} + \epsilon_{ij},$$

where:

- Y_{ij} is the outcome for individual i in cluster j,
- T_{ij} is a binary treatment indicator,
- X_{1ij} and X_{2ij} are Level-1 covariates (e.g., individual characteristics),
- Z_{1j} and Z_{2j} are Level-2 covariates (e.g., contextual or institutional features),
- $U_{0j} \sim \mathcal{N}(0, \tau_{00})$ is the random intercept capturing unobserved cluster-level differences,
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the individual-level residual error term, assumed to be independent across individuals and clusters,
- γ_0 through γ_6 are fixed-effect coefficients.

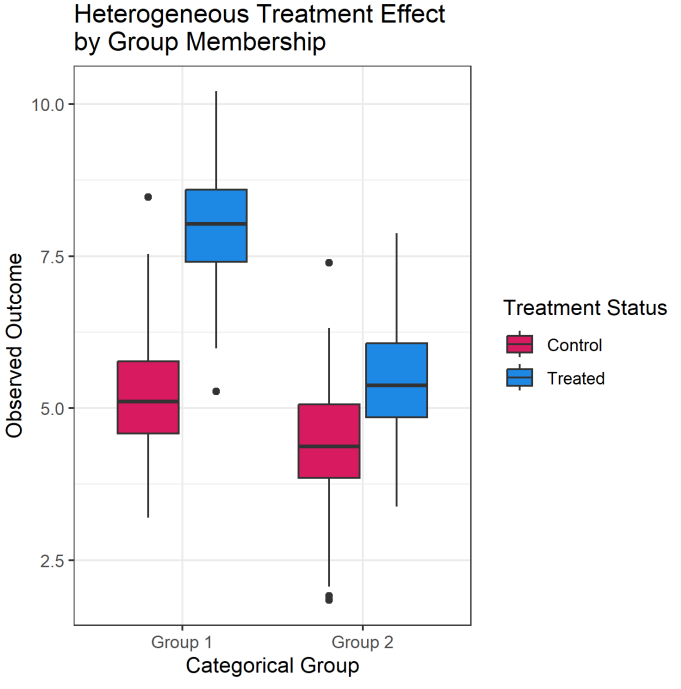

Figure 1. Treatment effect differences between two groups. Box plots show outcome distributions for treated and control conditions, separately by two subgroups defined by a binary covariate.

In this specification, the interaction term $T_{ij} X_{1ij}$ tests whether the treatment effect systematically varies with the Level-1 covariate X_{1ij}. A significant coefficient γ_4 indicates that this covariate moderates the treatment effect, accounting for at least some of the observed heterogeneity.

While this example includes only two covariates per level and a single interaction term, both the number of covariates and the choice of interactions are flexible and should be guided by theoretical considerations, empirical context, and data availability. Interactions may be included with any individual- or cluster-level covariates and can involve more than one moderator when appropriate.

To illustrate how treatment effect heterogeneity can arise through an interaction with a binary covariate, Figure 1 presents a simulated example. The figure shows box plots of outcomes for treated and control groups, separated by a binary covariate (X_{1ij}) that defines two subgroups: Group 1 and Group 2. The difference in treatment effects across these subgroups is visually apparent—Group 1 exhibits a clear positive shift in outcomes under treatment, whereas Group 2 shows little to no treatment effect. This simulated example demonstrates how an interaction term can reveal subgroup-specific treatment responses that would otherwise be masked by the overall average effect.

Models with an interaction between a treatment and a covariate are widely used in applied research settings where investigators seek to understand for whom and under what conditions treatments are most effective (Gelman & Hill, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 2011). They offer interpretable estimates and can be readily extended to include additional predictors or cross-level interactions. However, linear interaction models may not adequately capture more complex or nonlinear forms of heterogeneity. In such cases, more flexible modeling strategies, such as decision trees, splines, or machine learning methods, may provide valuable alternatives while preserving the underlying goal of identifying meaningful variation in treatment effects.

3.2 Causal Machine Learning for Conditional Average Treatment Effects

Conditional average treatment effects (CATEs) describe how treatment effects vary with observed individual- and cluster-level covariates, $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_j$, respectively. Causal machine learning methods allow for the estimation of heterogeneous treatment effects without requiring manual specification of interaction terms or functional forms. Formally, CATE can be defined as:

$$\tau_{ij} = [Y_{ij}(1) - Y_{ij}(0) \mid \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \boldsymbol{Z}_j = \boldsymbol{z}_j].$$

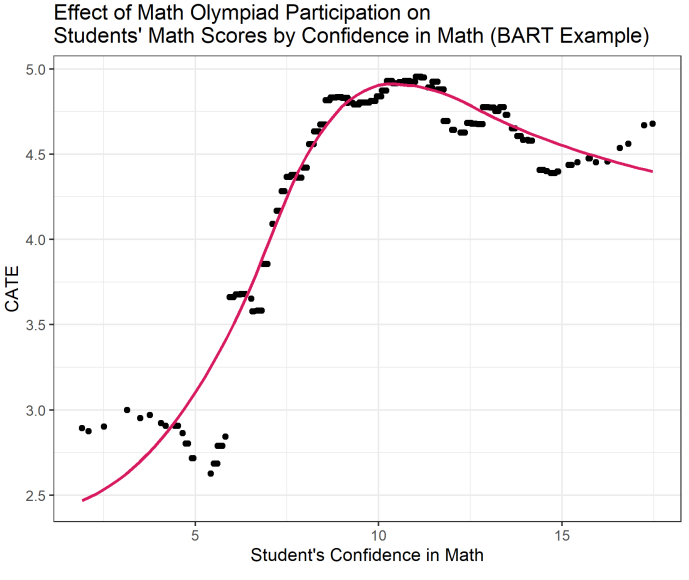

Figure 2. Illustration of CATE variation across a continuous covariate. The estimated CATE is plotted against student confidence in math. While treatment effects increase with confidence at low to moderate levels, the effect plateaus for higher-confidence students.

Unlike traditional regression models that depend on explicit interaction terms to detect effect heterogeneity, machine learning approaches—such as generalized random forests (Athey et al., 2019), Bayesian additive regression trees (BART) (Hill, 2011), and targeted maximum likelihood estimation

(TMLE) (van der Laan & Rose, 2011)—are well-suited for flexibly estimating CATE in complex, high-dimensional settings. These approaches rely on assumptions such as strong ignorability and SUTVA, and typically require careful regularization and sufficient sample size to perform well. While they may yield less interpretable models compared to traditional approaches, they can uncover subtle and nonlinear patterns of treatment effect variation that may otherwise go unnoticed.

Figure 2 provides an illustration of how CATE can vary nonlinearly with a continuous covariate. Using simulated data, treatment effects were estimated at the individual level using BART, and then smoothed to visualize the relationship between the treatment effect and the covariate. In this example, the treatment is binary and corresponds to whether a student participated in a math Olympiad. The outcome is student math performance. The horizontal axis represents students' confidence in math, while the vertical axis shows estimated conditional average treatment effects.

The figure shows that the conditional average treatment effect increases with student confidence at lower and moderate levels, but levels off at higher levels, suggesting that the treatment is more beneficial as students' confidence increases, but only up to a certain point at which additional confidence is no longer associated with additional benefit. Although the overall pattern is captured, the local variation in the individual treatment effect is not fully recovered, which highlights both the strengths and the limitations of machine learning estimators in detecting fine-grained heterogeneity.

3.3 Random Slopes of Treatment Effects

Multilevel or hierarchical linear models provide a flexible framework for modeling treatment effect heterogeneity across clusters by introducing random effects. In particular, random slope models allow the effect of treatment to vary across higher-level units (e.g., schools, clinics, or regions), capturing the possibility that normal contexts may amplify or diminish treatment impacts.

A general specification that includes both a random intercept and a random slope for treatment, along with covariates at both levels, is:

$$Y_{ij} = (\gamma_0 + U_{0j}) + (\gamma_1 + U_{1j})T_{ij} + \boldsymbol{\gamma}_2^\top \boldsymbol{X}_{ij} + \boldsymbol{\gamma}_3^\top \boldsymbol{Z}_j + \epsilon_{ij},$$

where:

- U_{0j} is the random intercept for cluster j,
- U_{1j} is the random slope for the treatment effect in cluster j,
- $\gamma_0, \gamma_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ are fixed-effect coefficients.

The random effects (U_{0j}, U_{1j}) are assumed to follow a bivariate normal distribution:

$$\begin{pmatrix} U_{0j} \\ U_{1j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\tau} \right), \quad \text{where} \quad \boldsymbol{\tau} = \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix}.$$

This formulation allows both the baseline outcome level and the treatment effect to vary across clusters. The covariance term τ_{01} captures the association between cluster-specific intercepts and treatment effects, such as whether clusters with higher baseline outcomes tend to exhibit stronger or weaker treatment effects.

Fixed effects for Level-1 and Level-2 covariates ($\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_j$) control for observed characteristics, but the key contribution of this model lies in its ability to capture unobserved between-cluster heterogeneity through U_{1j}. These models are especially useful when clusters represent meaningful social or institutional environments—such as schools, clinics, or neighborhoods—and when there is theoretical or empirical motivation to expect treatment effects to vary across these settings.

Random slope models are widely applied in education and health research, where understanding context-specific effectiveness is central to evaluating and implementing interventions (Gelman &

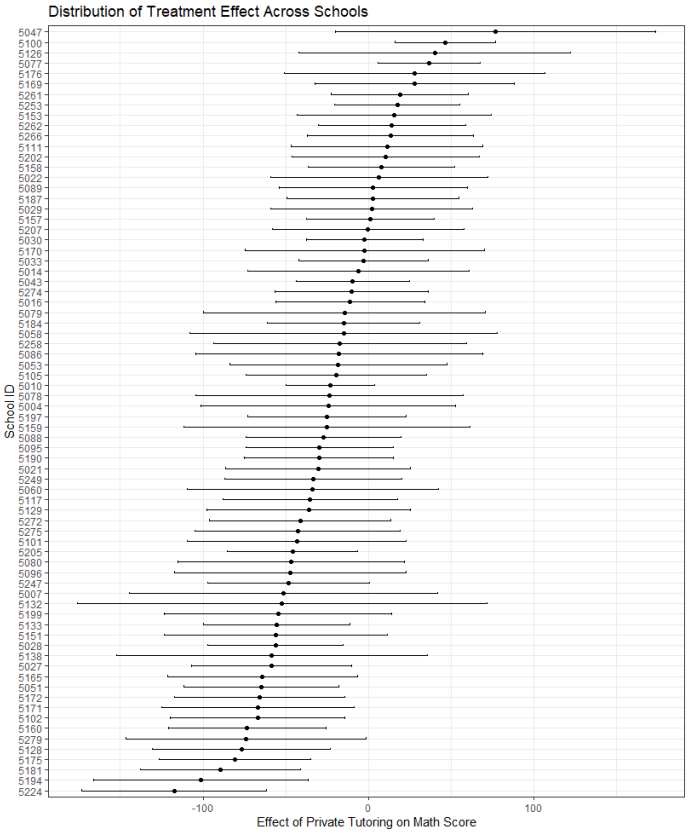

Figure 3. Caterpillar plot of school-level treatment effects (random slopes) of tutoring on math performance using 2019 U.S. TIMSS data. Schools are ordered by the magnitude of the predicted treatment effect, from smallest to largest. Horizontal lines represent 95% uncertainty intervals for predicted cluster-level effects.

Hill, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 2011). Extensions of this framework include modeling random slopes as a function of cluster-level variables or incorporating cross-level interactions to explore sources of treatment effect variation (Kim et al., 2023).

Figure 3 presents a caterpillar plot of predicted random slopes from a multilevel model applied to the 2019 U.S. Trends in International Mathematics and Science Study (TIMSS) data (National Center for Education Statistics & International Association for the Evaluation of Educational Achievement, 2020). The treatment of interest is math tutoring, and the outcome is student math performance. Each point represents the predicted treatment effect (random slope) for a specific school, with schools ordered by the magnitude of their estimated effects. The horizontal lines represent 95% uncertainty intervals, constructed under the assumption of normality for the random coefficients.

These intervals tend to be wider for schools with smaller student sample sizes, reflecting greater estimation uncertainty. Additional contributors to interval width include the intraclass correlation, the informativeness of covariates, and the effects of partial pooling in multilevel models. Partial pooling refers to the way multilevel models balance school-specific estimates with information from the overall distribution of effects: estimates for schools with limited data are "shrunk" toward the overall average, reducing variance but potentially smoothing out extreme values. This approach improves stability in estimation, especially when cluster sizes vary or data are sparse.

3.4 Latent Class Analysis and Finite Mixture Models for Unobserved Heterogeneity

The CATE framework relies on observed covariates. When relevant covariates are unmeasured or subgroup membership is not directly observable, methods such as multilevel models, causal forests, or BART cannot directly estimate effect heterogeneity. In such cases, latent class analysis (LCA) or finite mixture models (FMM) can be used to capture unobserved heterogeneity (Kim & Steiner, 2015; Kim et al., 2016; Loh & Kim, 2022; Lyu et al., 2022; Suk et al., 2021).

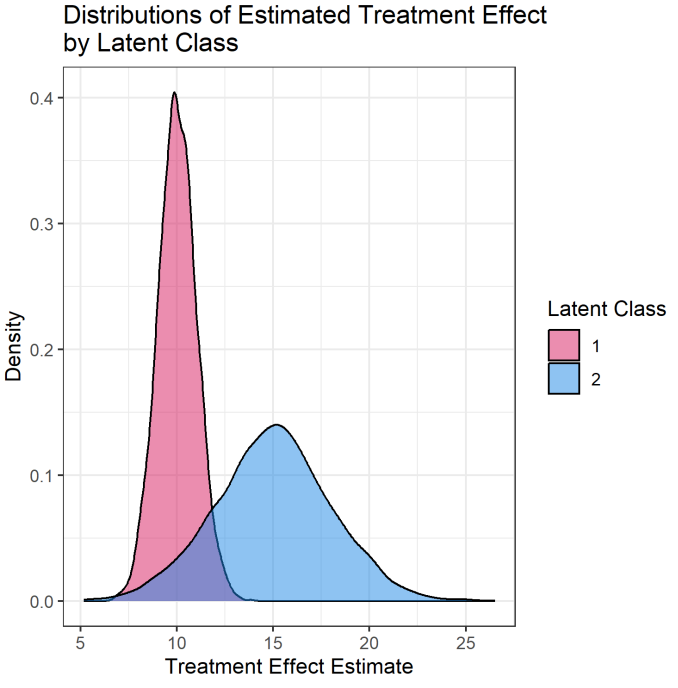

Figure 4. Illustrative example of a mixture of two latent classes with distinct treatment effects. The distributions differ in their means, variances, and class proportions. Class 1 has a larger average treatment effect with lower variance, while Class 2 shows a smaller effect with greater spread.

These models assume the existence of C latent subgroups within the population, each associated with a distinct treatment effect. Let $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_j$ denote observed individual- and cluster-level covariates. The outcome model can be expressed as a mixture:

$$f(Y_{ij} \mid T_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_j) = \sum_{c=1}^{C} \pi_c \, \phi\left(Y_{ij}; \mu_{cij}, \sigma_c^2\right),$$

where π_c is the proportion of units in latent class c, and $\phi(Y_{ij}; \mu_{cij}, \sigma_c^2)$ denotes the normal density with mean μ_{cij} and variance σ_c^2. Although other distributions could be used depending on the context and nature of the outcome variable, we use the normal density here for simplicity. In each class c, the conditional mean is specified using a multilevel model:

$$\mu_{cij} = \gamma_{0c} + U_{0cj} + \gamma_{1c} T_{ij} + \boldsymbol{\gamma}_{2c}^{\top} \boldsymbol{X}_{ij} + \boldsymbol{\gamma}_{3c}^{\top} \boldsymbol{Z}_j,$$

where:

- μ_{cij} is the conditional mean of the outcome Y_{ij} for individual i in cluster j and latent class c,

- γ_{0c} is the class-specific fixed intercept,
- U_{0cj} is the class-specific random intercept for class c, assumed to follow $U_{0cj} \sim \mathcal{N}(0, \tau_c^2)$,
- γ_{1c} is the class-specific fixed effect for the treatment variable T_{ij},
- $\boldsymbol{\gamma}_{2c}$ is a vector of class-specific fixed-effect coefficients for the individual-level covariates $\boldsymbol{X}_{ij}$,
- $\boldsymbol{\gamma}_{3c}$ is a vector of class-specific fixed-effect coefficients for the cluster-level covariates $\boldsymbol{Z}_j$.

Class membership is unobserved but inferred from the joint distribution of the outcomes and covariates. These models are particularly useful when treatment effect heterogeneity is driven by latent traits, diagnostic subtypes, or behavioral profiles that are not captured by observed covariates (Jo, 2002; Linzer & Lewis, 2011). They are commonly applied in education, psychology, and health research, where unmeasured heterogeneity may be of substantive importance.

Figure 4 illustrates a simulated example of treatment effect heterogeneity arising from unobserved latent classes. The plot displays a mixture of two treatment effect distributions corresponding to two subgroups. These subgroups differ in several respects: Class 1 has a lower treatment effect with smaller variance, while Class 2 shows a higher treatment effect with greater variance. The two classes are of comparable size in this example. In practice, the number of latent classes may exceed two, and their relative proportions can vary substantially. Some or all classes may exhibit markedly distinct treatment effects, implying qualitative differences in causal impacts across subpopulations.

To determine the number of latent classes in applied settings, researchers often use information criteria such as BIC or AIC, along with considerations related to theoretical grounding and clarity of interpretation. In causal modeling contexts, it is also important to assess the stability of the identified classes and whether they meaningfully distinguish subgroups with different treatment responses. Substantive theory or practical considerations may also guide the final choice.

4. Discussion and Concluding Remarks

Understanding how treatment effects vary across individuals and contexts is central to advancing causal inference, particularly in multilevel settings where individuals are nested within clusters such as schools, clinics, or communities. While the average treatment effect (ATE) provides a useful summary of the overall impact of an intervention, it may conceal meaningful heterogeneity in treatment responses. The individual treatment effect (ITE), which captures person-specific causal impacts, is theoretically appealing but generally unidentifiable without strong and often untestable assumptions.

This paper has reviewed several frameworks for modeling treatment effect heterogeneity, including interactions with covariates, conditional average treatment effects (CATEs), random slope/coefficient models, and latent class models. These approaches differ in their assumptions, estimation strategies, and interpretive goals, yet all aim to uncover structured variation in treatment effects that holds both theoretical and practical significance.

Researchers can implement the methods discussed in this paper using a variety of software tools and R packages. For multilevel or mixed-effects models with moderation effects and random slopes, the `lme4` and `nlme` packages are commonly used, with `brms` offering a Bayesian implementation. CATE can be estimated using several R packages, including `grf` and `causalTree`, as well as BART implementations such as `bartCause`, `dbarts`, and `BART`. Ensemble methods such as `xgboost` are also applied in practice. Buhrman et al. (2023) compares several CATE estimators, and Kim et al. (2023) illustrates the estimation of cross-level interactions using CATE methods. For mixture models and latent class analysis with multilevel observational data, we refer readers to Lyu et al. (2022), Suk et al. (2021), Kim et al. (2016), and Kim and Steiner (2015). While our focus is on R implementations, we note that `EconML`, a Python-based library, also offers a suite of tools for CATE estimation, including doubly robust learners and meta-learners.

The examples in this paper illustrate the practical value of these methods. Incorporating treatment effect variation into analysis can support more effective targeting of interventions, guide theory

development, and improve equity and efficiency in policy implementation. Methods that accommodate observed and unobserved sources of heterogeneity, such as causal machine learning and latent variable models, offer useful complements to conventional regression-based approaches and contribute to more robust causal analyses.

There is a growing need for a systematic approach to treatment effect heterogeneity, particularly in complex data settings involving multiple levels, time points, and sources of variability. We advocate for a more unified and conceptually grounded framework for understanding, estimating, and applying treatment effect heterogeneity. Such a framework can help clarify how different forms of heterogeneity arise, how they relate to substantive theory and design choices, and how they can guide decision-making in applied research. Advancing this area also requires greater attention to identification conditions, theoretical motivations, and the practical implications of heterogeneous effects. Progress in this direction will depend on continued integration across methodological traditions, including multilevel modeling, causal machine learning methods, and latent variable approaches, and will support the development of a more comprehensive understanding of causal inference in the social and behavioral sciences.

Acknowledgement

We thank Wen Wei Loh for valuable feedback during the early stages of this study, and proceedings editor Gabriel Wallin for helpful comments.

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

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Buhrman, G., Liao, X., & Kim, J.-S. (2023). Exploring conceptual differences among nonparametric estimators of treatment heterogeneity in the context of clustered data. *The Annual Meeting of the Psychometric Society*, 261–274.

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

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

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

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

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

Jo, B. (2002). Modeling of principal stratification using latent class models: Characterizing treatment effects on latent classes. *Journal of the American Statistical Association*, *97*(459), 136–147. https://doi.org/10.1198/016214502388618900

Kim, J.-S., Liao, X., & Loh, W. W. (2023). 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., Lim, W.-C., & Steiner, P. M. (2016). Causal inference with observational multilevel data: Investigating selection and outcome heterogeneity. *The Annual Meeting of the Psychometric Society*, 287–308. https://doi.org/https://doi.org/10.1007/978-3-319-56294-0_26

Kim, J.-S., & Steiner, P. M. (2015). Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy. In *Quantitative psychology research* (pp. 293–306). Springer. https://doi.org/https://doi.org/10.1007/978-3-319-19977-1_21

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.

Leite, W. (2016). *Practical propensity score methods using r*. Sage Publications. https://doi.org/https://doi.org/10.4135/9781071802854

Linzer, D. A., & Lewis, J. B. (2011). Polca: An r package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1–29. https://doi.org/10.18637/jss.v042.i10

Loh, W. W., & Kim, J.-S. (2022). Evaluating sensitivity to classification uncertainty in latent subgroup effect analyses. *BMC Medical Research Methodology*, *22*(1), 247.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*(19), 2937–2960.

Lyu, W., Kim, J.-S., & Suk, Y. (2022). Estimating heterogeneous treatment effects within latent class multilevel models: A bayesian approach. *Journal of Educational and Behavioral Statistics*, 10769986221115446.

National Center for Education Statistics & International Association for the Evaluation of Educational Achievement. (2020). Trends in international mathematics and science study (timss) 2019: U.s. data [U.S. dataset from the 2019 TIMSS].

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 465–472.

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

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

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

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.

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

Suk, Y., Kim, J.-S., & Kang, H. (2021). Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes. *Journal of Educational and Behavioral Statistics*, *46*(3), 323–347.

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.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer. https://doi.org/10.1007/978-1-4419-9782-8