Assessing Heterogeneous Causal Effects across Clusters in Partially Nested Designs

Xiao Liu

The University of Texas at Austin

Author Note

Correspondence should be addressed to Xiao Liu, Department of Educational Psychology, The University of Texas at Austin, Austin, TX 78712. Email: xiao.liu@austin.utexas.edu. This study was not preregistered.

Abstract

Intervention studies in psychology often have a partially nested design (PND): After individuals are assigned to study arms, individuals in a treatment arm are subsequently assigned to clusters (e.g., therapists/therapy-groups) to receive treatment, whereas individuals in a control arm are unclustered. Given the presence of clustering in the treatment arm, it can be of interest to examine heterogeneity of treatment effects across the clusters; but this is challenging in PNDs. First, in defining a causal effect of treatment for a specific cluster, it is unclear how the treatment and control outcomes should be compared, as the clustering is absent in the control arm. Although it may be tempting to compare outcomes between a specific cluster and the entire control arm, this crude comparison may not represent a causal effect even in PNDs with randomized treatment assignments, as the cluster assignment may be nonrandomized (elaborated in this study). In this study, extending the principal stratification and principal score approach, we develop methods to define, identify, and estimate the causal effects of treatment across specific clusters in a PND where the treatment and/or cluster assignment may be nonrandomized. Particularly, we obtain four alternative estimators, including a triply-robust estimator that can provide more robustness to parametric model misspecification. The triply-robust estimator also facilitates using data-adaptive techniques (e.g., machine learning) to assist in the causal effect estimation, and we implement this via the double machine learning procedure. We evaluate the estimators' performance through simulations, and illustrate the application using an empirical PND example.

Keywords: partially nested design, heterogeneous causal effects, principal score, principal stratification, multilevel modeling, multiply robust estimation, double machine learning

Assessing Heterogeneous Causal Effects across Clusters in Partially Nested Designs

In intervention studies in psychology, a common design scenario is the partially nested design (PND): after individuals are assigned to study arms, individuals in one (or some) of the study arms are assigned to clusters, whereas individuals in the other study arm(s) have no such clustering (e.g., Baldwin et al., 2011; Bauer et al., 2008; Cox & Kelcey, 2022; Hedges & Citkowicz, 2015; Lohr et al., 2014; Roberts & Roberts, 2005; Sanders, 2011; Sterba, 2017).

For example, PNDs occur in studies comparing a therapist-delivered treatment vs. a wait-list control arm (Lee & Thompson, 2005; Sterba, 2017; Walwyn & Roberts, 2010), where patients in the treatment arm are assigned to one of several therapists to receive treatment (i.e., patients clustered within therapists), whereas patients in the control arm have no such clustering. PNDs also appear in studies evaluating a group-administered treatment (e.g., participants are assigned to therapy groups) vs. a no-treatment (unclustered) control arm (e.g., Compas et al., 2009; Stice et al., 2006). In education, PNDs are common in assessing effects of summer-class or after-school programs (where students are assigned to teachers/classes) vs. individual as-usual activities out-of-school (i.e., unclustered; e.g., Lohr et al., 2014; Reed et al., 2019).

As shown in the above examples, a key feature of the PND is that for each individual, the cluster assignment in the treatment arm, by design, can only be observed after the treatment assignment[1]. In other words, in the PND, the observed cluster membership of an individual is a *posttreatment* variable. This temporal ordering of the treatment assignment and clustering differentiates the PND from designs with naturally existing clusters, such as multisite designs (site as cluster). In a multisite design, within each site/cluster individuals are assigned to the

[1] We use "treatment assignment" to refer to the assignment of individuals to the treatment and control arms at baseline; correspondingly, we use "posttreatment" to mean after the treatment assignment, and use "pretreatment" and "baseline" to express before (or right before) the treatment assignment.

treatment and control arms (e.g., Raudenbush & Bloom, 2015), and each individual's site/cluster membership is therefore a baseline variable (observed before the treatment assignment).

Because of this PND feature in the timing of clustering (observed after the treatment assignment), substantive research questions that involve assessing treatment effects across specific clusters can be challenging to investigate in a PND (described later); however, such research questions (e.g., assessing treatment effects across therapists [or therapy groups] in a PND of therapist-delivered [or group-therapy] treatment) can often be of interest. In extensive literature on studying causal effects of treatments when individuals are nested in clusters (e.g., teachers, therapists, sites in multisite designs; e.g., Angrist et al., 2017; Austin et al., 2001; Bloom et al., 2017; Chang & Stuart, 2022; Goldstein & Spiegelhalter, 1996; Normand et al., 2016; Raudenbush & Bloom, 2015; Weiss et al., 2017), it has been shown that assessing the cluster-specific causal effects can be substantively appealing in various ways, such as in providing effect estimates for specific cluster(s) or describing the effect distribution across clusters (e.g., effect variation or ranking); this is especially so when there are potential reasons for the effects of treatment to be heterogeneous across clusters.

Heterogeneous Cluster-specific Treatment Effects in PNDs

In the PND, the effects of treatment can be heterogeneous across clusters in the treatment arm, such as due to cluster characteristics in the implementation of treatment conditions (e.g., Bauer et al., 2008; Lohr et al., 2014; Roberts, 1999). For example, in a PND with a therapist-led treatment arm (patients clustered within therapists), the treatment effects for patients assigned to different therapists can be heterogeneous because of differences in therapists' skills (e.g., Baldwin & Imel, 2013; Wampold & Imel, 2015). In a PND evaluating a summer school program (students clustered within teachers), the program effects can be heterogeneous across teachers

due to heterogeneity in teacher experiences (e.g., Lohr et al., 2014).

Given potential heterogeneity in the cluster-specific treatment effects in a PND, it can be of substantive interest to assess such possibly heterogeneous effects (e.g., appealing in ways described earlier). However, there is currently limited research that provides methods for causally defining, identifying, and thereby estimating the effects of treatment assignment across specific clusters in PNDs.

Although for designs with naturally existing clusters (e.g., multisite [non]randomized designs) methods exist for assessing the causal effects across clusters (e.g., Bloom et al., 2017; Hong, 2004; Raudenbush & Bryk, 2002; Thoemmes & West, 2011), these methods are not applicable for PNDs—because in a PND, the cluster assignment is unobserved for every individual in the control arm (i.e., the study arm with no clustering). In the PND literature, there is substantial research on appropriately incorporating the partial nesting data structure into modeling the observed outcomes, often with the adapted multilevel models (or structural equation models; see e.g., Candlish et al., 2018; Lohr et al., 2014; Sterba, 2017 for recent reviews). For example, in a basic multilevel model adapted for the PND (see e.g., Sterba, 2017), the control arm is modeled by a single-level regression, with an intercept representing the mean observed outcome for control arm individuals; the treatment arm is modeled by a multilevel regression, with a cluster-specific intercept representing the mean observed outcome for treatment arm individuals in a specific cluster. While researchers can use such adapted multilevel models to describe the observed outcomes in a PND, a comparison of the observed outcomes—such as a comparison of observed outcomes between the treatment arm individuals in a specific cluster vs. the control arm individuals—may not represent a causal effect for the specific cluster.

In particular, if the cluster assignment in the treatment arm is nonrandomized, individuals

in a specific cluster could have different baseline characteristics compared to individuals in the control arm, even when the treatment assignment is randomized. Then, a crude comparison of observed outcomes between a specific cluster vs. the control arm would not provide a causal effect (Lohr et al., 2014). In PNDs, nonrandomized cluster assignments occur frequently. For example, as Lohr et al. (2014) discussed, reserachers may not want to randomize the cluster assignments in PNDs, so they could examine the treatment effects in typical implementation settings, which are policy-relevant. In an empirical example presented later (Reed et al., 2019), a PND was used to evaluate a summer reading program (in the treatment arm, students were clustered within teacher-led classes); the teacher/class (i.e., cluster) assignment for the treatment arm students were nonrandomized and could depend on students' instructional needs.

Furthermore, besides the cluster assignments, nonrandomized treatment assignments can occur in PNDs due to various ethical or practical reasons (e.g., Liu et al., 2023). For example, PNDs are common in studies comparing a summer school program vs. no-treatment (or wait-list) control arm. In this context, it is often infeasible to randomize children to the study arms (e.g., because a no-treatment control during summer is no instruction at all, which can be infeasible or unethical; e.g., Kim & Quinn, 2013; Reed et al., 2019). With nonrandomized treatment assignment added onto the nonrandomized cluster assignment, it can be even more challenging to assess the causal effects of treatment across specific clusters in PNDs.

The Current Study

This study aims to develop methods for defining, identifying, and estimating the causal effects of treatment assignment across specific clusters in PNDs. We achieve the aim as follows.

First, to define the causal effect of treatment assignment for a specific cluster in a PND—which we call the "cluster-specific treatment effect", we employ the principal stratification

framework in causal inference (Frangakis & Rubin, 2002). Principal stratification is a commonly used framework for defining subgroup causal effects when the subgroups are described by posttreatment variables (e.g., Ding & Lu, 2017; Imbens & Rubin, 2015; Jo, 2008; Page et al., 2015); and for the PND, the cluster assignment is a posttreatment variable (as described earlier).

Second, to identify the cluster-specific treatment effects in a PND, we extend the principal score approach for identifying causal effects defined by principal stratification (e.g., Ding & Lu, 2017; Jiang et al., 2022). The principal score approach has been popularly applied in causal analysis to overcome challenges posed by posttreatment variables (e.g., Follmann, 2000; Hill et al., 2002; Jo & Stuart, 2009). Based on the principal score literature (Jiang et al., 2022), we obtain four identification formulas for the cluster-specific treatment effect.

Third, with the four identification formulas obtained, we develop four corresponding estimators for estimating the cluster-specific treatment effects in a PND. Specifically, we provide three estimators that rely on modeling two out of the three "nuisance functions" (which we call the treatment probability, cluster assignment probability, and outcome mean). By combining all the three nuisance functions, we further provide a triply-robust estimator that is consistent so long as any two of the nuisance functions are estimated with correctly specified parametric models. Moreover, extending the literature on causal inference with machine learning (e.g., Athey & Imbens, 2019; Chernozhukov et al., 2018; Jiang et al., 2022), we also combine the triply-robust estimator with the double machine learning procedure (Chernozhukov et al., 2018), to facilitate using data-adaptive techniques (e.g., machine learning) to assist in the causal effect estimation. We develop an R package for applying all of our developed estimators.

Lastly, for the identification and estimation of the cluster-specific treatment effects, we describe simplified versions for PNDs with randomized treatment and/or cluster assignments.

In the rest of this article, we first describe the definition and identification of the cluster-specific treatment effects in a PND. We then obtain the estimators for these effects. A simulation study is conducted to evaluate the performance of the estimators. An empirical example is provided to illustrate the application of the developed methods. We conclude with a discussion of implications and limitations of the current study and future directions.

Definition and Identification of the Cluster-specific Treatment Effects in PNDs

In this section, we describe the causal definition and identification results for cluster-specific treatment effects in a PND. For notation simplicity, the subscript "i" for individual i is omitted when no confusion occurs.

We consider a two-arm PND, with a total of J clusters in the treatment arm, each denoted by a positive integer k in $\{1,2,\dots,J\}$. In the PND, individuals assigned to the treatment arm are subsequently assigned to one of the J clusters (while individuals assigned to the control arm are unclustered). Let T denote the treatment assignment for an individual, with $T = 1$ for individuals assigned to the treatment arm and $T = 0$ for the control arm. The observed cluster assignment, denoted by K, is $K = k$ for individuals assigned to the treatment arm ($T = 1$) and then to cluster k to receive the treatment condition (where $k \in \{1,2,\dots,J\}$), and is $K = 0$ (no cluster assignment) for individuals assigned to the control arm ($T = 0$). Finally, let Y be each individual's observed outcome.

Throughout this study, we do not consider noncompliance. That is, we assume perfect compliance such that individuals (i) comply with the treatment assignment and (ii) comply with the cluster assignment if assigned to the treatment arm.

Notations: Potential Outcomes and Potential Cluster Assignments

We employ the potential outcomes framework (Holland, 1986; Imbens & Rubin, 2015;

Rubin, 1974) to define causal effects. Following previous research (Liu et al., 2023; VanderWeele & Hernan, 2013), we use the expanded potential outcome notations to explicitly shown that an individual's outcome if assigned to the treatment arm may vary depending on the cluster assignment.

Specifically, let $K(1)$ be potential cluster assignment that an individual would have if assigned to the treatment arm. As we assume perfect compliance, $K(1)$ is a value in $\{1,2, \ldots, J\}$, and furthermore, $K(0) = 0$ (i.e., no cluster assignment if assigned to the control arm).[2] For the potential outcomes, let $Y(1, k)$ (where k = 1, …, J) be an individual's potential outcome if assigned to the treatment arm and then to cluster k to receive the treatment. Then, $Y(1, K(1))$ is the individual's potential outcome if assigned to the treatment arm; thus, we write it as $Y(1)$ for notation simplicity. Let $Y(0)$ be the individual's potential outcome if assigned to the control arm.

We make the stable unit treatment value assumption (SUTVA) for the potential cluster assignment and expanded potential outcome notations (Rubin, 1980, 1986; VanderWeele & Hernan, 2013). For the PND, this assumes that (i) there are no multiple versions of the treatment condition within each cluster of the treatment arm, and there are no multiple versions of the control condition within the control arm, and that (ii) there is no interference among individuals. In general, with clustered designs, plausibility of the no-interference assumption warrants consideration (e.g., Hong & Raudenbush, 2006; Liu et al., 2023; Thoemmes & West, 2011); for example, this could involve considering whether the assignment to treatment or control arm for an individual may affect other individuals in the PND; this no-interference assumption could also be more plausible when little interaction exists between the treatment and control arm

[2] This is similar to the "strong monotonicity" assumption in the causal inference literature (e.g., Angrist et al., 1996; Ding & Lu, 2017; Jin & Rubin, 2008).

individuals (see e.g., Hong & Raudenbush, 2006; Schochet, 2015 for more discussions).

Under the SUTVA, the notations for the potential quantities are connected to the observed quantities; that is, the observed cluster assignment is $K = K(1)$ for individuals assigned to $T = 1$ (treatment arm), and is $K = K(0) = 0$ for individuals assigned to $T = 0$ (control arm); the observed outcome is $Y = Y(0)$ for individuals assigned to $T = 0$, and is $Y = Y(1, k)$ for individuals assigned to $T = 1$ and then to $K = k$ (cluster k in the treatment arm); further, combining the potential cluster assignment and potential outcome, we have $Y = Y(1, K(1)) = Y(1)$ for individuals assigned to $T = 1$.

Definition of the Cluster-specific Treatment Effect

For cluster k in the treatment arm ($k \in \{1, \dots, J\}$), we define the cluster-specific treatment effect, ATE_k, as the average causal effect of assignment to treatment vs. control arm—for the subgroup of individuals who would be assigned to cluster k if they were assigned to the treatment arm (i.e., the subgroup with the potential cluster assignment being $K(1) = k$); that is:

$$ATE_k = \mathrm{E}[\, Y(1) - Y(0) \mid K(1) = k \,] \quad (1)$$

where k is a value in $\{1, \dots, J\}$ (denoting the cluster k in the treatment arm).

We obtain the above causal effect definition for ATE_k (Eq. 1) with the principal stratification framework in causal inference (Frangakis & Rubin, 2002). As mentioned in the introduction, this framework is useful for defining subgroup causal effects involving posttreatment variables. Specifically, using this framework entails conditioning on (i.e., stratifying on) the joint vector of (two) potential values that a posttreatment variable would have, one under the treatment arm, and the other under the control arm. The potential value of a posttreatment variable (under either treatment or control) is, by definition, unaffected by the treatment assignment, just like a baseline (i.e., pretreatment) covariate (Frangakis & Rubin,

2002). Hence, conditioning on the potential values provides a subgroup causal effect.

In our case with the PND, the observed cluster assignment K is the posttreatment variable. Using the principal stratification, we condition on the joint vector of potential cluster assignments under the treatment and control arms, namely $\{K(1), K(0)\}$; because for every individual, $K(0) = 0$ (no cluster assignment under the control arm), conditioning on the joint vector is equivalent to conditioning on only $K(1)$ (potential cluster assignment under the treatment arm). As with the potential value of any posttreatment variable (under either treatment or control), the potential cluster assignment that an individual would have, namely $K(1)$, is unaffected by the treatment assignment, just like a baseline covariate. Thus, by conditioning on $K(1)$, the ATE_k defined in Eq. (1) is just like a subgroup causal effect.

Broadly, the ATE_k in Eq. (1) defines a causal effect of the treatment assignment, because it compares the potential outcomes under different treatment assignments [i.e., $Y(1)$ vs. $Y(0)$] for the *same* set of individuals. The ATE_k also defines a causal effect specific for the cluster k, because the *same* set of individuals considered in ATE_k includes only the individuals who would be assigned to the cluster k to receive treatment (if they were assigned to the treatment arm; i.e., the individuals with $K(1) = k$).

For example, in a PND comparing a therapist-led therapy treatment vs. a wait-list control (i.e., the treatment arm participants are assigned to therapists, the control arm are unclustered), ATE_k—the therapist-specific treatment effect for therapist k—represents the average causal effect of assignment to the therapy treatment (vs. the wait-list control) for the participants who would be assigned to receive the therapy treatment with therapist k (if they were assigned to the therapy treatment arm; i.e., participants with the potential therapist assignment being $K(1) = k$).

Note that for every cluster $k \in \{1, \dots, J\}$ in the treatment arm of the PND, the subgroup of

individuals who would be assigned to the cluster, namely the subgroup with $K(1) = k$, is a latent subgroup (referred to as latent principal stratum in principal stratification; e.g., Ding & Lu, 2017; Frangakis & Rubin, 2002). It is a latent subgroup because for every individual who is assigned to the control arm in reality, we can never observe what cluster assignment the individual would have if the individual were (counterfactually) assigned to the treatment arm; that is, $K(1)$ is unobserved (i.e., missing, or, latent) for every individual with $T = 0$. Besides the unobserved $K(1)$ values, unobserved potential outcome values are involved in the causal effect ATE_k: the $Y(1)$ potential outcome value (or the $Y(0)$ potential outcome) is unobserved for every individual who is assigned to the control arm $T = 0$ (or to the treatment arm $T = 1$) in reality. Thus, to calculate the ATE_k (defined in Eq. 1), causal identification assumptions are required.

Identification of the Cluster-specific Treatment Effect

To identify the cluster-specific treatment effects ATE_k (k = 1,…, J) in the PND, we extend the identification assumptions made in the principal score approach[3] for principal stratification analyses (e.g., Hill et al., 2002; Jiang et al., 2022; Jo & Stuart, 2009; Stuart & Jo, 2015); for general introductions about this approach, see Feller et al. (2017), Ding and Lu (2017), among others.

Below, we present the identification results (assumptions and formulas) for a PND where both the treatment and cluster assignments may be nonrandomized. With the identification results obtained, we also describe the simplified versions for the special cases of PNDs where the treatment and/or cluster assignment is randomized.

[3] Other approaches exist for conducting principal stratification analyses (e.g., Hirano et al., 2000; Imbens & Rubin, 1997; Jiang & Ding, 2021; Jin & Rubin, 2008; Jo, Asparouhov, Muthén, et al., 2008); see e.g., Ding and Lu (2017) for a summary and comparison of the approaches. Our current study focuses on the principal score approach. We consider that in PNDs, researchers may have information about baseline covariates that might affect the cluster assignment and treatment assignment; such information would help strengthen plausibility of the ignorability assumptions required by the principal score approach (described next).

Identification Assumptions

Two ignorability assumptions (IA1 and IA2) are required to identify the cluster-specific treatment effects in a PND (ATE_k, k = 1, …, J), IA1 for the treatment assignment T and IA2 for the potential cluster assignment $K(1)$. Following terminology in the principal score literature, IA2 (below) is referred to as the "principal ignorability" assumption (e.g., Ding & Lu, 2017; Jiang et al., 2022), as the potential cluster assignment in a PND is parallel to the principal stratification variable in this literature.

Specifically, let **X** be a set of baseline covariates measured for each individual before the treatment assignment. The ignorability assumptions are written as follows:

IA1 ("Treatment assignment ignorability"). $T \perp\!\!\!\perp \{ K(1), Y(0), Y(1) \} \mid \mathbf{X}$.

IA2 ("Principal ignorability"). $K(1) \perp\!\!\!\perp Y(0) \mid \mathbf{X}$.

IA1 states that for each individual, the treatment assignment is "as good as randomized"—that is, independent of the potential outcomes and potential cluster assignment, given the baseline covariates **X**. IA1 would be satisfied if the treatment assignment is randomized; but even then, IA2 is required. IA2 states that for each individual, the potential cluster assignment if assigned to the treatment arm is "as good as randomized"—that is, independent[4] of the individual's potential outcome under the control arm, given the baseline covariates **X**. In the empirical illustration (presented later), we more substantively interpret IA1 and IA2 in the context of a PND evaluating a summer program.

The ignorability assumptions (IA1 and IA2) are empirically untestable. Hence, echoing

[4] The independence requirement in IA2 can be weakened as $\mathrm{E}(Y(0) \mid K(1) = k, \mathbf{X}) = \mathrm{E}(Y(0) \mid \mathbf{X})$ across all clusters k = 1, …, J, which states that given **X**, the conditional mean of the potential control outcome is constant across individuals with different potential cluster assignments. This weaker version of IA2, together with IA1, can also suffice for identifying the ATE_k's. In the main text, we use the independence requirement in IA2 for easy interpretation (following the principal score literature; e.g., Ding & Lu, 2017).

the literature (e.g., Ding & Lu, 2017; Feller et al., 2017), we emphasize that to investigate the cluster-specific treatment effects ATE_k's in PNDs, it is critical to have a rich set of baseline covariates ($\mathbf{X}$) that might affect the outcome, treatment assignment, and/or cluster assignment.

Identification Formulas

Extending the principal score literature (Jiang et al., 2022), we obtain four alternative formulas for identifying (nonparametrically calculating) the cluster-specific treatment effect ATE_k's in a PND. These identification formulas involve combinations of some "nuisance functions", which are functions of the baseline covariates $\mathbf{X}$ and can be calculated with data. Specifically, three nuisance functions are involved, referred to as (i) the treatment probability $\pi_t(\mathbf{X})$, (ii) the cluster assignment probability $p_k(\mathbf{X})$, and (iii) the outcome mean $\{\mu_{y|1,k}(\mathbf{X})$, $\mu_{y|0}(\mathbf{X})\}$, as described below.

The treatment probability, $\pi_t(\mathbf{X}) = p(T = 1 \mid \mathbf{X})$, is the conditional probability of assignment to the treatment arm given the covariates (also known as the propensity score in the literature; e.g., Imbens & Rubin, 2015; Rosenbaum & Rubin, 1983).

The cluster assignment probability, $p_k(\mathbf{X}) = p(K = k \mid T = 1, \mathbf{X})$, is the conditional probability that the observed cluster assignment is $K = k$ for an individual in the treatment arm given the covariates. Immediately, under IA1 (treatment assignment ignorability), we can use this observed cluster assignment probability to identify $p(\,K(1) = k \mid \mathbf{X}\,)$, the conditional probability of the potential cluster assignment given the covariates (or, the "principal score"; e.g., Ding & Lu, 2017); that is, under IA1, $p(\,K(1) = k \mid \mathbf{X}\,) = p_k(\mathbf{X})$.

The outcome mean refers to $\{\mu_{y|1,k}(\mathbf{X}), \mu_{y|0}(\mathbf{X})\}$, where $\mu_{y|1,k}(\mathbf{X}) =$ $\mathrm{E}(\,Y \mid T = 1, K = k, \mathbf{X}\,)$ and $\mu_{y|0}(\mathbf{X}) = \mathrm{E}(\,Y \mid T = 0, \mathbf{X}\,)$ are two conditional mean outcomes given the covariates. $\mu_{y|1,k}(\mathbf{X})$ is for calculating (or, predicting) the outcome if assigned to the

treatment arm and then to cluster k; $\mu_{y|0}(\mathbf{X})$ is for calculating/predicting the outcome if assigned to the control arm.

By combining two or three of the above three nuisance functions, the ATE_k can be identified by four alternative identification formulas. The identification formulas require IA1 and IA2, as well as a positivity assumption, which states that given the covariates of each individual, each individual is probable to be assigned to either the treatment or control arm and is probable to be assigned to each cluster k ($k \in \{1, \ldots, J\}$) if assigned to the treatment arm (i.e., $p_k(\mathbf{X}) > 0$ for k = 1, …, J and $0 < \pi_t(\mathbf{X}) < 1$ for every individual's $\mathbf{X}$ value). Below we present each formula. Technical details and proofs are provided in the supplemental materials.

(a. "trt-cluster") With the treatment probability and cluster assignment probability, we can identify ATE_k as:

$$ATE_k = \frac{\mathrm{E}\left[\frac{1_{\{T=1\}}1_{\{K=k\}}}{\pi_t(\mathbf{X})}Y\right]}{\mathrm{E}[p_k(\mathbf{X})]} - \frac{\mathrm{E}\left[\frac{1_{\{T=0\}}p_k(\mathbf{X})}{1-\pi_t(\mathbf{X})}Y\right]}{\mathrm{E}[p_k(\mathbf{X})]}. \quad (2)$$

In the formula in Eq. (2) (and also the formulas in Eqs. 3-5 below), "$1_{\{.\}}$" is an indicator function; for example, $1_{\{T=1\}}$ is 1 for an individual in the treatment arm (i.e., with $T = 1$) and 0 otherwise; the expectation ("E[.]") is averaging over the covariates $\mathbf{X}$ of all individuals.

(b. "trt-y") With the treatment probability and outcome mean, we have:

$$ATE_k = \frac{\mathrm{E}\left[\frac{1_{\{T=1\}}1_{\{K=k\}}}{\pi_t(\mathbf{X})}\{\mu_{y|1,k}(\mathbf{X}) - \mu_{y|0}(\mathbf{X})\}\right]}{\mathrm{E}\left[\frac{1_{\{T=1\}}1_{\{K=k\}}}{\pi_t(\mathbf{X})}\right]}. \quad (3)$$

(c. "cluster-y") With the cluster assignment probability and outcome mean, we identify ATE_k as:

$$ATE_k = \frac{\mathrm{E}\left[p_k(\mathbf{X})\left\{\mu_{y|1,k}(\mathbf{X}) - \mu_{y|0}(\mathbf{X})\right\}\right]}{\mathrm{E}\left[p_k(\mathbf{X})\right]}. \quad (4)$$

(d. "triply-robust") Finally, with the treatment probability, cluster assignment probability, and outcome mean, we obtain the following triply-robust formula for ATE_k:

$$ATE_k = \frac{\mathrm{E}\left[\phi_{y|1,k}(\mathbf{X}) - \phi_{y|0}(\mathbf{X})\right]}{\mathrm{E}\left[p_k^{\mathrm{dr}}(\mathbf{X})\right]}, \text{where} \quad (5)$$

$$\phi_{y|1,k}(\mathbf{X}) = \frac{1_{\{T=1\}}1_{\{K=k\}}}{\pi_t(\mathbf{X})}\left\{Y - \mu_{y|1,k}(\mathbf{X})\right\} + \mu_{y|1,k}(\mathbf{X})p_k^{\mathrm{dr}}(\mathbf{X}),$$

$$\phi_{y|0}(\mathbf{X}) = \frac{p_k(\mathbf{X})\mathbf{1}_{(T=0)}}{1-\pi_t(\mathbf{X})}\left\{Y - \mu_{y|0}(\mathbf{X})\right\} + \mu_{y|0}(\mathbf{X})p_k^{\mathrm{dr}}(\mathbf{X}) \text{ , and}$$

$$p_k^{\mathrm{dr}}(\mathbf{X}) = \frac{1_{\{T=1\}}[1_{\{K=k\}} - p_k(\mathbf{X})]}{\pi_t(\mathbf{X})} + p_k(\mathbf{X}).$$

The superscript "dr" in $p_k^{\mathrm{dr}}(\mathbf{X})$ denotes doubly-robust, as $p_k^{\mathrm{dr}}(\mathbf{X})$ is constructed based on the doubly-robust strategy for causal effect estimation (see e.g., Bang & Robins, 2005; Schafer & Kang, 2008 for more on doubly-robust estimation). Here, by doubly-robust, it means that by taking the average of $p_k^{\mathrm{dr}}(\mathbf{X})$, we obtain a doubly-robust formula for identifying the potential proportion of individuals who would be assigned to cluster k, that is, $\mathrm{E}\left[1_{\{K(1)=k\}}\right] = \mathrm{E}\left[p_k^{\mathrm{dr}}(\mathbf{X})\right]$.[5]

Several comments on the above identification formulas (Eq. 2-Eq.5) are worth noting. As these formulas all identify the same causal effect ATE_k under the identification assumptions, the formulas are equivalent to each other nonparametrically. Note that their equivalence holds in a nonparametric sense, that is, when the nuisance functions are all calculated nonparametrically. In practice, however, the covariates **X** are likely contain continuous covariates and/or contain a not-

[5] Technically, this formula is named as doubly-robust, because it allows us to consistently estimate this potential proportion (i.e., $\mathrm{E}\left[1_{\{K(1)=k\}}\right]$), so long as we correctly specified the parametric model for estimating either $\pi_t(\mathbf{X})$ or $p_k(\mathbf{X})$ or both of the two (see the supplemental materials for additional details). Such robustness property (i.e., robust to misspecification of the parametric model for either one of the two involved nuisance functions) is often described as doubly-robust (see also e.g., Kang & Schafer, 2007).

small number of covariates, making nonparametric calculation infeasible, and thus estimation models (e.g., parametric models) need to be specified for estimating the nuisance functions. When the nuisance functions are estimated, the four identification formulas (Eq. 2-Eq.5) would then yield four distinct estimators of ATE_k.

Particularly, when the nuisance functions are estimated with parametric models (e.g., those in the next section), using the formula in Eq. (2), Eq. (3), or Eq. (4), which involves only two out of the three nuisance functions (i.e., "trt-cluster" in Eq. 2, "trt-y" in Eq. 3, or "cluster-y" in Eq. 4), the resulting estimator of ATE_k is consistent, when the parametric models are correctly specified for both of the involved nuisance functions. For example, using the formula "trt-cluster" in Eq. (2), the resulting estimator of ATE_k is consistent, when the parametric models for both $\pi_t(\mathbf{X})$ (the treatment probability) and $p_k(\mathbf{X})$ (the cluster assignment probability) are correctly specified, but would be generally inconsistent, if the parametric model for either $\pi_t(\mathbf{X})$ or $p_k(\mathbf{X})$ is misspecified.

In comparison, with the triply-robust formula in Eq. (5), which combines all three nuisance functions, the resulting estimator of ATE_k is consistent, so long as parametric models are correctly specified for at least two of the three nuisance functions. In other words, the resulting estimator is robust to misspecification of the parametric model for any one of the three nuisance functions (referred to as the "triply-robust" estimator). More specifically, the triply-robust estimator is consistent for ATE_k, when the parametric models are correctly specified for (a) the treatment probability and cluster assignment probability, or (b) the treatment probability and outcome mean, or (c) the cluster assignment probability and outcome mean, or (d) for all three nuisance functions. The supplemental materials provide additional technical details about the triply-robust property. See also the previous literature on multiply-robust estimation with the

principal score (Jiang et al., 2022).

Special Case of the Identification Results: Randomized Assignments

With the identification results obtained above, we discuss their simplified versions for identifying the cluster-specific treatment effects ATE_k's in PNDs where the treatment assignment and/or cluster assignment is randomized.

When the Treatment Assignment is Randomized

Suppose T is randomized. Then, IA1 also holds unconditional on the covariates, that is, $T \perp\!\!\!\perp \{Y(0), Y(1), K(1)\}$. The treatment probability for each individual becomes a known constant given by the proportion randomized to the treatment arm, $\pi_t = p(T = 1)$. Thus, the identification formulas in Eq. (2)-Eq. (5) also hold by substituting π_t for $\pi_t(\mathbf{X})$.

For example, the identification formula ("trt-y") in Eq. (3) can be simplified to $ATE_k = \mathrm{E}\left[\mu_{y|1,k}(\mathbf{X}) - \mu_{y|0}(\mathbf{X}) \mid K = k, T = 1\right]$. This suggests that in a PND with randomized treatment assignment, we can estimate the cluster-specific treatment effect ATE_k as follows. Fit a model for the nuisance functions involved, namely $\mu_{y|1,k}(\mathbf{X})$ and $\mu_{y|0}(\mathbf{X})$; for example, $\mu_{y|1,k}(\mathbf{X})$ can be an outcome regression fitted to the treatment arm data ($T = 1$) with the clusters K and covariates $\mathbf{X}$ included, and $\mu_{y|0}(\mathbf{X})$ can be an outcome regression fitted to the control arm data ($T = 0$) with $\mathbf{X}$ as predictors. Then, from the fitted models, obtain estimates (i.e., predicted values) of the nuisance functions for each individual, namely estimates of each individual's outcome under treatment in cluster k and outcome under control (i.e., $\hat{\mu}_{y|1,k}(\mathbf{X}_i)$ and $\hat{\mu}_{y|0}(\mathbf{X}_i)$ of each individual i). Finally, estimate ATE_k as $\widehat{ATE_k} = \mathrm{E}\left[\hat{\mu}_{y|1,k}(\mathbf{X}_i) - \hat{\mu}_{y|0}(\mathbf{X}_i) \mid K_i = k, T_i = 1\right]$, namely the cluster-specific mean difference between these estimated outcome values among the individuals in cluster k of the treatment arm (i.e., among individuals with $K_i = k, T_i = 1$).

When the Cluster Assignment is Randomized

In some PND scenarios, researchers are unable to randomize the treatment assignment due to practical considerations, but may have control over the cluster (e.g., teacher) assignment for an individual if the individual choose to participate in the treatment arm (e.g., a treatment arm providing a summer-class program). When the cluster assignment would be set by randomization, that is, when $K(1)$ is randomized, then IA2 also holds without the covariates, namely $K(1) \perp\!\!\!\perp Y(0)$; additionally, IA1 needs not include the potential cluster assignment, and can be simplified as $T \perp\!\!\!\perp \{Y(0), Y(1)\} \mid \mathbf{X}$. The identification formulas can be simplified by replacing the cluster assignment probability $p_k(\mathbf{X})$ with the known constant, $p_k = p(K = k \mid T = 1)$, the proportion of treatment arm individuals in cluster k.

For example, the identification formula (c. "cluster-y") in Eq. (4) can be simplified as $ATE_k = \mathrm{E}[\mu_{y|1,k}(\mathbf{X}) - \mu_{y|0}(\mathbf{X})]$. This shows that with randomized cluster assignment, the cluster-specific treatment effect ATE_k can be estimated as $\widehat{ATE_k} = \mathrm{E}\big[\hat{\mu}_{y|1,k}(\mathbf{X}_i) - \hat{\mu}_{y|0}(\mathbf{X}_i)\big]$, the sample mean difference between the two estimated outcome values among all individuals (where $\hat{\mu}_{y|1,k}(\mathbf{X}_i)$ and $\hat{\mu}_{y|0}(\mathbf{X}_i)$ can be obtained as described above).

When Both the Treatment Assignment and Cluster Assignment are Randomized

In a PND with randomized treatment assignment and randomized cluster assignment if an individual were randomized to the treatment arm, both IA1 and IA2 hold without conditioning on the covariates, and IA1 needs not include the cluster assignment; that is, IA1 and IA2 can be simplified as $T \perp\!\!\!\perp \{Y(0), Y(1)\}$ and $K(1) \perp\!\!\!\perp Y(0)$. Accordingly, the identification formulas also hold without the covariates $\mathbf{X}$, as no covariate adjustment is needed.

For example, the identification formula (a. "trt-cluster") in Eq. (2) can be simplified to $ATE_k = \mathrm{E}[Y \mid K = k, T = 1] - \mathrm{E}[Y \mid T = 0]$. Thus, the ATE_k can be calculated as the

difference-in-means, that is, the difference in the observed mean outcome between the cluster k of the treatment arm vs the control arm. This difference-in-means estimator is the crude comparison mentioned in the introduction. Our results here show that in the PND where the treatment and cluster assignments are both randomized, this difference-in-means (i.e., $\mathrm{E}[Y \mid K = k, T = 1] - \mathrm{E}[Y \mid T = 0]$) can be causally interpreted as the cluster-specific treatment effect ATE_k, the average causal effect of being assigned to treatment (vs. control) for the subgroup of individuals who would be assigned to cluster k (i.e., individuals with $K(1) = k$). Furthermore, with the randomized cluster assignment (i.e., $K(1)$ is randomized), the subgroup of individuals who would be assigned to cluster k is, in expectation, the same as the entire sample with respect to baseline characteristics.

Estimation of the Cluster-specific Treatment Effects in PNDs

With the identification formulas (Eqs. 2, 3, 4, 5), estimating the cluster-specific treatment effects ATE_k's in a PND involves estimating the three nuisance functions, namely the treatment probability $\pi_t(\mathbf{X})$, cluster assignment probability $p_k(\mathbf{X})$, and outcome mean $\{\mu_{y|1,k}(\mathbf{X}), \mu_{y|0}(\mathbf{X})\}$. In this section, we first apply commonly used parametric models to estimate the nuisance functions, where we also describe the resulting estimators for ATE_k's in PNDs with randomized assignments. Then, we describe the use of the triply-robust formula with the double machine learning procedure (Chernozhukov et al., 2018).

Estimation with Parametric Models following the Identification Formulas

For estimating the treatment probability, $\pi_t(\mathbf{X}) = p(T = 1 \mid \mathbf{X})$, a logistic regression of treatment on the covariates can be specified, such as:

$$\text{logit}[p(T = 1 \mid \mathbf{X})] = \alpha_0 + \mathbf{X}'\alpha_1 \quad (6)$$

where "logit" denotes the logit link. From the fitted model, we obtain $\hat{\pi}_t(\mathbf{X}_i)$, the estimated

treatment probability of each individual i in the sample.

For estimating the cluster assignment probability, $p_k(\mathbf{X}) = p(K = k \mid T = 1, \mathbf{X})$ where $k = 1, \ldots, J$, a multinomial (logistic) regression can be fitted to the treatment arm data, such as:

$$\log\left[\frac{p(K = k \mid T = 1, \mathbf{X})}{p(K = 1 \mid T = 1, \mathbf{X})}\right] = \beta_{k,0} + \mathbf{X}'\beta_{k,1}, \tag{7}$$

for cluster $k = 2, \ldots, J$ in the treatment arm. From the fitted model, we obtain $\hat{p}_k(\mathbf{X}_i)$, the estimated cluster assignment probability for each individual i in the sample.

For estimating the outcome mean $\mu_{y|1,k}(\mathbf{X}) = \mathrm{E}(Y \mid T = 1, K = k, \mathbf{X})$ and $\mu_{y|0}(\mathbf{X}) = \mathrm{E}(Y \mid T = 0, \mathbf{X})$, we may use the multilevel random-effects models adapted for PNDs (Baldwin et al., 2011; Bauer et al., 2008; Roberts & Roberts, 2005; Sterba, 2017; Sterba et al., 2014). To describe the model, let Y_{ik} be the observed outcome of individual i in cluster k of the treatment arm. A multilevel random-intercept outcome model for the PND can be specified as:

	Treatment arm ($T_i = 1$):	Control arm ($T_i = 0$):	(8)
level-1:	$Y_{ik} = \gamma_{0k}^{(t)} + \mathbf{X}'^{\text{within}}_{ik}\gamma_{1,\text{within}}^{(t)} + \epsilon_{ik}^{(t)},\ k = 1, \ldots, J$	$Y_i = \gamma_0^{(c)} + \mathbf{X}_i'\gamma_1^{(c)} + \epsilon_i^{(c)}$	
level-2:	$\gamma_{0k}^{(t)} = \gamma_0^{(t)} + \mathbf{X}'^{\text{between}}_{k}\gamma_{1,\text{between}}^{(t)} + u_k^{(t)}$		
	$\epsilon_{ij}^{(t)} \sim N\left(0, \sigma_\epsilon^{(t)2}\right), u_k^{(t)} \sim N\left(0, \sigma_u^{(t)2}\right)$	$\epsilon_i^{(c)} \sim N\left(0, \sigma_\epsilon^{(c)2}\right)$	

where $\mathbf{X}^{\text{between}}_{k} = \frac{\sum_{i=1}^{n}\mathbf{X}_i 1_{\{K_i=k, T_i=1\}}}{\sum_{i=1}^{n} 1_{\{K_i=k, T_i=1\}}} = \mathrm{E}(\mathbf{X}_i \mid K_i = k, T_i = 1)$ contains the cluster-mean covariates (with the coefficients $\gamma_{1,\text{between}}^{(t)}$) and $\mathbf{X}^{\text{within}}_{ik} = \mathbf{X}_i - \mathbf{X}^{\text{between}}_{k}$ contains the cluster-mean centered covariates (with the coefficients $\gamma_{1,\text{within}}^{(t)}$). In the level-1 model for outcomes of the treatment arm, $\gamma_{0k}^{(t)}$ is the random cluster-specific intercept and $\epsilon_{ik}^{(t)}$ is the within-cluster residual; in the level-2 model, $\gamma_{0k}^{(t)}$ is predicted by $\mathbf{X}^{\text{between}}_{k}$, with $\gamma_0^{(t)}$ being the intercept and

$u_k^{(t)}$ being the level-2 residual. In the model for outcomes of the (unclustered) control arm, the covariates $\mathbf{X}_i$ have coefficients $\gamma_1^{(c)}$, $\gamma_0^{(c)}$ is the intercept, and $\epsilon_i^{(c)}$ is the residual.

While random-effects outcome modeling is common in clustered data analyses, it may result in inaccurate estimation of the outcome mean $\mu_{y|1,k}(\mathbf{X}_i)$ (i.e., the conditional mean outcome under treatment in cluster k), because the estimated random coefficients (such as $u_k^{(t)}$ in Eq. 8) can have the shrinkage issue when cluster sizes are not sufficiently large (see e.g., Raudenbush & Bryk, 2002). With inaccurate estimates of $\mu_{y|1,k}(\mathbf{X}_i)$, estimators of the ATE_k resulting from the formulas in Eq. (3) ("trt-y") or Eq. (4) ("cluster-y") can be biased, as these estimators require the outcome mean to be accurately estimated. Although the outcome mean is also involved in the triply-robust formula (Eq. 5), the resulting estimator of the ATE_k can be robust to estimation errors in any one of the three nuisance functions, and remain consistent so long as the other two nuisance functions are accurately estimated.

For estimating the outcome mean $\mu_{y|1,k}(\mathbf{X}_i)$, an alternative to random-effects modeling is fixed-effects modeling (e.g., Allison, 2009; Baldwin et al., 2011; McNeish & Kelley, 2019). A fixed-intercept model for outcomes of the treatment arm ($T_i = 1$) can be specified as:

$$\text{Treatment arm: } Y_{ik} = \sum_{k=1}^{J} \gamma_k^{(t)} 1_{\{K_i=k\}} + \mathbf{X}_i' \gamma_1^{(t)} + \epsilon_{ik}^{(t)} \tag{9}$$

where $\gamma_k^{(t)}$ is the fixed cluster-specific intercept of cluster k in the treatment arm and $1_{\{K_i=k\}}$ is the cluster indicator. The residual $\epsilon_{ik}^{(t)}$ is assumed normally distributed with mean zero and constant variance. The fixed-effects model has no distributional assumptions on the fixed cluster-specific intercept, and thus can avoid the shrinkage issue in estimating $\mu_{y|1,k}(\mathbf{X}_i)$ with the random-effects modeling (such as Eq. 8). In the simulation study, we examine the performance of using the random- vs. fixed-effects outcome models for the estimation of ATE_k's.

Special Case of the Estimation: Randomized Assignments

When the treatment and/or cluster assignments are randomized, the cluster-specific treatment effects ATE_k's in PNDs have simplified identification formulas (described in the previous section), and thus simplified estimators. In this subsection, we connect such simplified estimators of ATE_k to the multilevel outcome model in Eq. (8), an outcome model commonly used in analyses of PNDs (e.g., Sterba et al., 2014).

When the Treatment Assignment is Randomized

With randomized treatment assignment (thus IA1 is satisfied) and assuming IA2 is plausible (i.e., the cluster assignment is as good as randomized given the baseline covariates **X**), the cluster-specific treatment effect ATE_k can be estimated with the simplified identification formula, $ATE_k = \mathrm{E}\left[\mu_{y|1,k}(\mathbf{X}) - \mu_{y|0}(\mathbf{X}) \mid K = k, T = 1\right]$. Implementing this formula with the outcome model in Eq. (8), the estimator $\widehat{ATE_k}$ is:

$$\widehat{ATE_k} = (\gamma_0^{(t)} - \gamma_0^{(c)}) + \mathbf{X'}_k^{\text{between}}(\gamma_{1,\text{between}}^{(t)} - \gamma_1^{(c)}) + u_k^{(t)}, \tag{10}$$

where the second term, $\mathbf{X'}_k^{\text{between}}(\gamma_{1,\text{between}}^{(t)} - \gamma_1^{(c)})$, captures how much the cluster-specific treatment effect (ATE_k) is associated with the cluster-specific means of individuals' baseline characteristics ($\mathbf{X}_k^{\text{between}}$). The third term, $u_k^{(t)}$ (the level-2 residual in Eq. 8), captures how much the ATE_k is associated with the cluster-specific treatment implementation that occurs after the cluster assignment (e.g., therapist-specific skills in therapy implementation, in a PND with therapists as clusters in a therapy treatment arm).

Additionally, the estimator in Eq. (10) helps explicate the variability of the cluster-specific treatment effect ATE_k in the randomized PND. From Eq. (10), $\text{var}\left(\widehat{ATE_k}\right) = \text{var}[\mathbf{X'}_k^{\text{between}}(\gamma_{1,\text{between}}^{(t)} - \gamma_1^{(c)})] + \text{var}(u_k^{(t)})$. This shows that in the PND, the between-cluster

treatment effect variability can come from two sources: (1) $\text{var}(u_k^{(t)})$, the between-cluster variability in treatment implementation (e.g., the between-therapist variability in therapy implementation skills), and (2) $\text{var}[\mathbf{X'}_k^{\text{between}}(\gamma_{1,\text{between}}^{(t)} - \gamma_1^{(c)})]$, the between-cluster variability in individuals' baseline characteristics. This second source would vanish (a) if $\gamma_{1,\text{between}}^{(t)} = \gamma_1^{(c)}$, meaning that for every baseline covariate in **X**, the covariate's between-cluster association with the treatment outcome is equal to the covariate's association with the control outcome, or (b) if the cluster assignment is perfectly randomized so that $\text{var}\left[\mathbf{X'}_k^{\text{between}}\right] = 0$.

When the Cluster Assignment is Randomized

With randomized cluster assignment (thus IA2 is satisfied) and assuming IA1 is plausible, the simplified identification formula can be used to calculate the cluster-specific treatment effect as $ATE_k = \text{E}[\mu_{y|1,k}(\mathbf{X}) - \mu_{y|0}(\mathbf{X})]$. Let the grand-mean of the covariates be $\text{E}(\mathbf{X}_i') = \bar{\mathbf{x}}'$. Then, using the outcome model in Eq. (8), we obtain the estimator $\widehat{ATE_k}$:

$$\widehat{ATE_k} = [\gamma_0^{(t)} - \gamma_0^{(c)} + \bar{\mathbf{x}}'(\gamma_{1,\text{within}}^{(t)} - \gamma_1^{(c)})] + \mathbf{X'}_k^{\text{between}}(\gamma_{1,\text{between}}^{(t)} - \gamma_{1,\text{within}}^{(t)}) + u_k^{(t)}, \quad (11)$$

where the first term is a constant. The second term, $\mathbf{X'}_k^{\text{between}}(\gamma_{1,\text{between}}^{(t)} - \gamma_{1,\text{within}}^{(t)})$, is nearly a constant (as the cluster assignment is randomized). Furthermore, in expectation, the second term would equal the constant $\text{E}(\mathbf{X}_i' \mid T_i = 1)(\gamma_{1,\text{between}}^{(t)} - \gamma_{1,\text{within}}^{(t)})$. This shows that in a PND where the cluster assignment is randomized but the treatment assignment is not, assessing the cluster-specific treatment effect ATE_k needs to account for the differences in individual baseline covariates ($\mathbf{X}_i$) between the treatment and control arms. The third term $u_k^{(t)}$, as described above (under Eq. 10), captures heterogeneity in the treatment implementation across clusters.

The estimator in Eq. (11) also shows that in a PND where the cluster assignment is randomized, regardless of whether the treatment assignment is randomized or not, the variability

of the cluster-specific treatment effect ATE_k would only come from $\text{var}(u_k^{(t)})$, the between-cluster variability in treatment implementation, which may be related to heterogeneous cluster characteristics (e.g., therapists' skills, in a PND with therapists as clusters).

When Both rhe Treatment Assignment and Cluster Assignment are Randomized

With randomized treatment assignment and randomized cluster assignment, the cluster-specific treatment effect can be estimated following the difference-in-means formula, $ATE_k = \text{E}[Y \mid K = k, T = 1] - \text{E}[Y \mid T = 0]$. As both IA1 and IA2 are satisfied by the randomization, there is no need to adjust for covariates for confounding control. Nonetheless, suppose we still utilize the covariates (e.g., for purposes of improving precision) and implement this formula with the outcome model in Eq. (8). With the randomized assignments, in a large sample, the covariates' cluster-means $\mathbf{X}_k^{\text{between}}$ would approximately equal their grand-means $\bar{\mathbf{x}}$; given this, the estimator $\widehat{ATE_k}$ can be written as:

$$\widehat{ATE_k} = \left(\gamma_0^{(t)} - \gamma_0^{(c)}\right) + \bar{\mathbf{x}}'\left(\gamma_{1,\text{between}}^{(t)} - \gamma_1^{(c)}\right) + u_k^{(t)}. \tag{12}$$

As in the previous special case with randomized cluster assignment, the variability of ATE_k would only come from heterogeneous treatment implementation across the clusters.

Estimation with Double Machine Learning following the Triply-Robust Formula

In the previous subsections, the nuisance functions involved in estimating the ATE_k are estimated using the specified parametric models. In practice, however, correctly specifying a parametric model can be difficult especially with more than a few covariates. With misspecified parametric models for the nuisance functions, the resulting estimators of the causal effect ATE_k can be biased even when the identification assumptions are satisfied. Generally, for estimating nuisance functions in causal effect estimation, an alternative approach is using data-adaptive techniques (e.g., machine learning), and using machine learning to assist in estimating causal

effects has been increasingly popular in diverse fields (e.g., Athey & Imbens, 2019; Brand et al., 2023; Chernozhukov et al., 2018; Huber, 2023; Vowels, 2023). Appropriate use of machine learning in causal effect estimation can be appealing in various ways, such as flexibly and data-adaptively estimating nuisance functions, relaxing linearity assumptions, and reducing analyst degrees of freedom (e.g., Dorie et al., 2019; Linero & Zhang, 2022; Vowels, 2023).

Extending the use of machine learning in the principal score literature (Jiang et al., 2022), in this subsection, we provide a machine learning based triply-robust estimator for the cluster-specific treatment effect (ATE_k). Particularly, the triply-robust formula is suitable for incorporating data-adaptive techniques (e.g., machine learning) into estimating the nuisance functions. The other formulas ("trt-cluster", "cluster-y", "trt-y" in Eqs. 2-4; which only involve two of the three nuisance functions) are not suitable, because they require accurate estimation of both of the involved nuisance functions, but the nuisance functions estimated with machine learning techniques (e.g., lasso) contain errors due to the regularization in these techniques; thus, using machine learning with these formulas (Eqs. 2-4) would generally yield biased estimates of the causal effect ATE_k (known as the regularization bias, e.g., Chernozhukov et al., 2018). In contrast, with the triply-robust formula (Eq. 5), the resulting estimator of ATE_k can remain consistent even when the nuisance functions are estimated with small errors (such as when they are estimated with machine learning), so long as the product of the errors vanishes sufficiently fast.[6] As such, the triply-robust formula facilitates incorporating machine learning to assist in the estimation of the causal effect ATE_k.

[6] Technically, the product of errors in the estimates for any two of the three nuisance functions needs to converge to zero at an $n^{1/2}$ rate (where n is the sample size); this is satisfied when the error of each estimated nuisance function converges to zero at an $n^{1/4}$ rate, a rate achievable by many machine learning methods (e.g., random forests). The supplemental materials provide additional details on the statistical assumptions under which the triply-robust estimator implemented with machine learning techniques can be consistent. See also the principal score literature on relevant statistical assumptions (Jiang et al., 2022).

Specifically, the triply-robust formula (Eq. 5) can readily be implemented with the double machine learning procedure (DML; Chernozhukov et al., 2018), a generic procedure for incorporating supervised machine learning techniques in causal effect estimation (see e.g., Huber, 2023; Knaus, 2022 for detailed introduction). To implement the triply-robust formula with this procedure, we replace the three nuisance functions (i.e., $\pi_t(\mathbf{X})$, $p_k(\mathbf{X})$, and $\{\mu_{y|1,k}(\mathbf{X})$, $\mu_{y|0}(\mathbf{X})\}$) with their cross-fitted estimates, namely their estimates produced via the cross-fitting procedure (e.g., two-fold cross-fitting) suggested by the DML (Chernozhukov et al., 2018). In the two-fold cross-fitting procedure, (a) the sample is randomly split into two folds; (b) the nuisance functions are fitted (e.g., via machine learning techniques) using the first fold; (c) these fitted nuisance functions are then used to obtain estimates (i.e., predicted values) of the nuisance functions for the second fold (i.e., obtain estimates $\hat{\pi}_t(\mathbf{X}_i)$, $\hat{p}_k(\mathbf{X}_i)$, $\hat{\mu}_{y|1,k}(\mathbf{X}_i)$, $\hat{\mu}_{y|0}(\mathbf{X}_i)$ for each individual i in the second fold); and (d) swap the roles of the two folds, and repeat the steps in (b) and (c) (i.e., use the second fold to obtain the fitted nuisance functions, and then use them to estimate/predict the nuisance functions for the first fold). This cross-fitting procedure aims to remove bias due to overfitting (e.g., Athey & Imbens, 2019; Chernozhukov et al., 2018). The resulting DML-based triply-robust estimator is consistent, under the main requirement that the estimators of the nuisance functions are sufficiently accurate (in the sense of Footnote 6), which can be satisfied by multiple machine learning techniques that are commonly available (e.g., random forests [Wager & Walther, 2015]; see e.g., Bach et al., 2023; Huber, 2023; Knaus, 2022 on other techniques).

To obtain statistical inference (e.g., confidence intervals) with the triply-robust estimator resulting from Eq. (5), $\widehat{ATE_k}^{\text{triply}} = \frac{\mathrm{E}[\hat{\phi}_{y|1,k}(\mathbf{X}_i) - \hat{\phi}_{y|0}(\mathbf{X}_i)]}{\mathrm{E}[\hat{p}_k^{\text{dr}}(\mathbf{X}_i)]}$, we exploit the results in the literature on doubly/multiply robust causal effect estimation and double machine learning (e.g.,

Chernozhukov et al., 2018; Knaus, 2022; Jiang et al., 2022). This literature indicates that the numerator and denominator of this estimator (which are formulated via the doubly-robust estimation strategy; see the supplemental materials) would be asymptotically normally distributed, provided that the nuisance functions are estimated accurately (such as with correctly specified parametric models or machine learning techniques accurate in the sense of Footnote 6). Thus, to obtain the confidence interval of the triply-robust estimator ($\widehat{ATE_k}^{\text{triply}}$), a ratio of two asymptotically normal variables, we draw bootstrap samples from estimates of the functions in its numerator and denominator (i.e., bootstrap samples from $\{\hat{\phi}_{y|1,k}(\mathbf{X}_i), \hat{\phi}_{y|0}(\mathbf{X}_i), \hat{p}_k^{\text{dr}}(\mathbf{X}_i)\}_{i=1,\ldots,n}$). For example, in the simulation study (below), we draw 1000 bootstrap samples to obtain the 95% percentile intervals for the triply-robust estimator.

We developed an R package "PND.heter.cluster" to facilitate applying our proposed methods to estimate the cluster-specific treatment effects in PNDs (ATE_k's). The R package is open-source and available at https://github.com/xliu12/PND.heter/tree/main/PND.hetercluster. Specifically, the R function "cluster.specific.ate()" can be used to estimate the ATE_k's with all of the estimators described above.

Simulation Study

To evaluate the performance of the developed estimators for the cluster-specific treatment effects in PNDs, we conduct a simulation study with two scenarios (described below).

Simulation Design

In Scenario 1, we examine the performance in PNDs with randomized treatment assignment. The functional forms of covariates **X** in the data-generation models are all linear.

In Scenario 2, we examine the performance in nonrandomized PNDs where the functional forms of covariates **X** in the data-generation models may be nonlinear (quadratic).

Specifically, we varied the functional forms of covariates **X** as either quadratic or linear in the data-generation models for the treatment assignments, cluster assignments, and outcomes. This would allow us to investigate the influences of misspecifying a parametric model(s) fitted for the nuisance functions on the estimators' performance (e.g., fitting models with only linear terms of **X** when the data-generation models had quadratic terms of **X**).

In both Scenario 1 and Scenario 2, four covariates $X_{i(q)}, q = 1, \dots, 4$ were generated, each from the standard normal distribution. The sample size n and the number of clusters in the treatment arm J was varied as (1) n = 300 with J = 20, (2) n = 400 with J = 10, (3) n = 600 with J = 30, (3) n = 800 with J = 20 or 40, and (4) n = 1600 with J = 40. These J and n values were examined considering previous simulation studies and sample size planning literature for PNDs with randomized assignments (e.g., Baldwin et al., 2013; Candlish et al., 2018; Lohr et al., 2014; Cox et al., 2022). Specifically, in a PND where the treatment and cluster assignments are both randomized with equal allocations of the sample size, the J and n combinations in (1)—(4) would lead to the average cluster size in the treatment arm being 7.5 (i.e., $\frac{300/2}{20}$), 10 (i.e., $\frac{600/2}{30}$ and $\frac{800/2}{40}$), or 20 (i.e., $\frac{400/2}{10}$, $\frac{800/2}{20}$, and $\frac{1600/2}{40}$), representing PNDs with moderate to relatively large cluster sizes. Additionally, the cluster size 10 and number of clusters J = 30 mimic our empirical example (next section).

Below, we describe the data generation for the two scenarios and the data analysis methods. The R code for the data generation is available at https://github.com/xliu12/PND.heter/tree/main/simulations_and_examples; the R code for implementing the data analysis methods is the same as that in our R package (https://github.com/xliu12/PND.heter/tree/main/PND.hetercluster).

Scenario 1. Randomized Treatment Assignments, Linear Data-Generation Models

The randomized treatment assignment T_i was generated with treatment probability π = 0.5 from $\text{Bernnoulli}(\pi = 0.5)$, with $T_i = 1$ indicating assignment to the treatment arm.

The cluster assignment in the treatment arm was generated as $K_i = K_i(1)T_i + 0 \cdot (1 - T_i)$, and the potential cluster assignment $K_i(1)$ was generated using a latent probit model for multinomial variable (e.g., Enders et al., 2016; Angrist et al, 2011): $K_i(1) = k$ if $K_i^* = \sum_{q=1}^{4} \beta X_{i(q)} + \epsilon_i < \tau_k$, where $k = 1, \ldots, J-1$, and $K_i(1) = J$ otherwise. ϵ_i was generated from the standard normal distribution, with τ_k being the k/J quantile of this distribution. The covariates $X_{(q)i}$'s together had a proportion of explained variance (R-squared) being 0.3 (relatively large based on Cohen 1988; set with β around 0.31).

The outcome was generated as $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$. For the potential outcome under the treatment arm $Y_i(1)$, or equivalently written as $Y_i(1, K_i(1))$, the data generation used a model similar to the multilevel model in Eq. (8). Specifically, for individual i who would be assigned to cluster k (i.e., $K_i(1) = k$), the individual's potential outcome was $Y_i\big(1, K_i(1)\big) = Y_i(1, k)$, where $Y_i(1, k)$ was generated with the cluster-specific intercept $u_k^{(t)}$, the cluster-mean covariates $X_{k(q)}^b = \frac{\sum_{i=1}^{n} X_{i(q)} 1_{\{K_i(1)=k\}}}{\sum_{i=1}^{n} 1_{\{K_i(1)=k\}}}$, and the individual's covariates; that is, $Y_i(1, k) = \gamma_0^{(t)} + \sum_{q=1}^{4} \gamma_b^{(t)} X_{k(q)}^b + \sum_{q=1}^{4} \gamma_1^{(t)} X_{i(q)} + u_k^{(t)} + \epsilon_{ik}^{(t)}$. The cluster-specific intercept $u_k^{(t)}$ and residual $\epsilon_{ik}^{(t)}$ were generated from normal distributions with means 0 and variances 0.2 and 0.8, respectively, so that the residual intra-cluster correlation was 0.2. The potential outcome under the control arm was generated as $Y_i(0) = \gamma_0^{(c)} + \sum_{q=1}^{4} \gamma_1^{(c)} X_{i(q)} + \epsilon_i^{(c)}$. The residual $\epsilon_i^{(c)}$ was normally distributed with mean 0 and variance 0.8 (the same as the variance of $\epsilon_{ik}^{(t)}$). The

average treatment effect had a Cohen's d (standardized mean difference) around 0.5 (set with $\gamma_0^{(t)}$ and $\gamma_0^{(c)}$ being 0.66 and 0, respectively). The coefficients of the covariates ($\gamma_b^{(t)}$, $\gamma_1^{(t)}$, and $\gamma_1^{(c)}$, around 0.91, 0.41, and 0.48, respectively) were set based on the proportion of variance explained (R-squared; Cohen, 1988; Rights & Sterba, 2019) when the cluster assignment was randomized. Specifically, for the potential treatment outcome $Y_i(1)$, the covariates explained around 0.5 of the total variance and explained 0.3 of the between-cluster variance; for the potential control outcome $Y_i(0)$, the covariates explained 0.5 of the total variance; these proportions of explained variances were relatively large (Cohen, 1988) and were set considering the R-squared values used in the literature on sample size planning for PNDs and clustered designs (e.g., Bloom et al., 2007; Cox et al., 2022; Lohr et al., 2014).

Scenario 2. Nonrandomized Treatment Assignments, Linear/Quadratic Data-Generation Models

In Scenario 2, we varied the functional forms of covariates $X_{i(q)}$'s in the data-generation models as either linear or quadratic, to examine the estimators' performance when the nuisance functions may be estimated with misspecified models. Specifically, we simulated five "true model" combinations. For combination (i), the true (i.e., data-generation) models for T_i, K_i, or Y_i all involved only linear covariate terms (i.e., no quadratic $X_{i(q)}$ terms); for combination (ii), (iii), or (iv), only the true model for T_i, for K_i, or for Y_i involved quadratic $X_{i(q)}$ terms; and for combination (v), the true models for T_i, K_i, and Y_i all involved quadratic $X_{i(q)}$ terms.

When the true model for T_i involved only linear covariate terms (the "linear scenario"), T_i (nonrandomized) was generated as $T_i = 1$ if $T_i^* = \sum_{q=1}^{4} \alpha X_{i(q)} + \varepsilon_i > 0$, where ε_i followed the standard normal distribution and the covariates explained 0.13 proportion of variance of T_i^*

(around medium; set with α around 0.19). When the true model for K_i (or Y_i) involved only linear covariate terms, the data generation for K_i (or Y_i) was the same as in Scenario 1 (the "linear scenario").

When quadratic $X_{i(q)}$ terms were involved in the true model for T_i (or K_i), we modified the data-generation model for T_i (or K_i) in the linear scenario (described above) by replacing each $X_{i(q)}$ (q = 1, …, 4) with the quadratic term $\frac{X_{i(q)}^2-1}{\sqrt{2}}$ (which had mean 0 and variance 1, the same as $X_{i(q)}$).

When quadratic $X_{i(q)}$ terms were involved in the data-generation for Y_i, we similarly modified the data-generation model for $Y_i(0)$ in the linear scenario by replacing each $X_{i(q)}$ (q = 1, …, 4) with the quadratic term $\frac{X_{i(q)}^2-1}{\sqrt{2}}$. For $Y_i(1)$, we modified its data-generation model in the linear scenario by replacing each $X_{i(q)}$ with the quadratic term $\frac{X_{i(q)}^2-1}{\sqrt{2}}$, including an interaction between one covariate and its cluster-mean (i.e., $X_{i(1)}X_{k(1)}^b$), and including cluster-specific slopes of two covariates ($v_{k(q)}^{(t)}X_{i(q)}$, q = 1, 2). That is, for individual i with the potential cluster assignment being $K_i(1) = k$, we generated $Y_i(1) = Y_i(1, K_i(1)) = Y_i(1, k)$ as: $Y_i(1,k) = \gamma_0^{(t)} + \sum_{q=1}^4 \gamma_b^{(t)} X_{k(q)}^b + \sum_{q=1}^4 \gamma_1^{(t)} \frac{X_{i(q)}^2-1}{\sqrt{2}} + u_k^{(t)} + \gamma_2^{(t)} X_{i(1)}X_{k(1)}^b + \sum_{q=1}^2 v_{k(q)}^{(t)} X_{i(q)} + \epsilon_{ik}^{(t)}$; the cluster-specific slopes were each simulated from a normal distribution with mean 0 and variance 0.1. With these terms added ($\gamma_2^{(t)}$ around 1.7), the covariates explained 0.5 of the total variance of $Y_i(1)$ and explained 0.7 of its between-cluster variance (relatively large based on Cohen [1988]; such relatively large proportions of explained variances have been considered in the sample size calculation literature, e.g., Lohr et al., 2014).

Analysis Methods

Under each data-generation condition, 1000 datasets were simulated using the R software (R Core Team, 2018). With each dataset, the cluster-specific treatment effects (ATE_k's) were estimated using five estimators resulting from the formulas in Eq. (2)—Eq. (5), labeled as: (i) "trt-cluster (linear)", (ii) "trt-y (linear)", (iii) "cluster-y (linear)", (iv) "triply-robust (linear)", and (v) "triply-robust (DML)".

For the estimators (i)—(iv) (labeled with "linear"), the nuisance functions were estimated using the parametric models with only linear covariates terms. Specifically, to estimate $\pi_t(\mathbf{X})$ and $p_k(\mathbf{X})$, we fitted the logistic regression in Eq. (6) (via the "glm()" R function) and the multinomial regression in Eq. (7) (via the "multinom()" R function in the "nnet" R package [Ripley & Venables, 2023]), respectively. To estimate $\mu_{y|0}(\mathbf{X})$, we fitted the (single-level) outcome regression in Eq. (8) (shown under "Control arm"). To estimate $\mu_{y|1,k}(\mathbf{X})$, we fitted the random-intercept outcome regression in Eq. (8) (shown under "Treatment arm"; via the "lme4" R package [Bates et al., 2011]); for comparison, we also estimated $\mu_{y|1,k}(\mathbf{X})$ using the fixed-intercept outcome regression in Eq. (9) (via the "lm()" R function). In presenting the simulation results, we label an estimator with "_yRE" (or "_yFE") if it involved the random-intercept (or fixed-intercept) outcome regression.

For the estimator (v) "triply-robust (DML)", the nuisance functions were obtained using the double machine learning procedure. Specifically, we used two-fold cross-fitting. For estimating $p_k(\mathbf{X})$, we used boosted trees via the "xgboost" R package (Chen et al., 2023; Chen & Guestrin, 2016). For estimating $\pi_t(\mathbf{X})$, as well as the outcome mean $\mu_{y|1,k}(\mathbf{X})$ and $\mu_{y|0}(\mathbf{X})$, we used the super learner ensemble procedure (via the "SuperLearner" R package; Polley & van der Laan, 2017; van Der Laan et al., 2007) with an ensemble of algorithms including boosted trees (via the "xgboost" R package), random forest (via the "ranger" R package; Wright & Ziegler,

2017), and generalized additive model (via the "gam" R package; T. Hastie, 2017, 2023). For either of the triply-robust estimators, we used 1000 bootstrap samples to obtain the 95% percentile interval (as described in the section on estimation).

Performance Evaluation

We evaluated the estimators' performance with respect to a fixed set of clusters. Specifically, in a PND, clusters in the treatment arm (e.g., teachers or therapists) may be considered fixed or as randomly drawn from a population of clusters (Baldwin et al., 2011; Serlin et al., 2003). From the random-cluster perspective, the cluster-specific treatment effects are also random. In previous simulation studies evaluating estimators of such random cluster effects, the clusters were treated as fixed in the simulation (i.e., a fixed set of cluster-specific intercepts/slopes were simulated), because the estimators' performance are inherently conditional on the specific clusters in the sample (e.g., Austin, 2005; Austin & Leckie, 2020; Candel & Winkens, 2003; Farrell et al., 1997). Following the previous simulation studies, we fixed the clusters in the data simulation (i.e., we generated the cluster-specific intercepts $u_k^{(t)}$'s and/or covariate slopes $v_{k(q)}^{(t)}$'s [where $k = 1,\ldots, J$] and then fixed them across the 1000 simulated datasets); thus, we evaluate the performance of each estimator for estimating the true sample values of cluster-specific treatment effects ATE_k's ($k = 1,\ldots, J$).

The performance measures include the relative bias and mean squared error (MSE); for the triply-robust estimators, we also evaluated the coverage rate of the 95% confidence interval. Specifically, we calculated the bias as the average difference between the estimate $\widehat{ATE_k}$ and the true sample value of ATE_k, that is, $Bias = \frac{\sum_{rep=1}^{N^{rep}} \widehat{ATE_k}^{rep} - ATE_k^{rep}}{N^{rep}}$, where superscript "^{rep}" represents a simulated dataset and N_{rep} is the number of simulated datasets (i.e., 1000). The

relative bias ($RBias$) was then calculated as the bias relative to the average of the true sample values, $RBias = \frac{Bias}{\sum_{rep=1}^{N_{rep}} ATE_k^{rep}/N^{rep}}$. The MSE was calculated as $MSE = \frac{\sum_{rep=1}^{N^{rep}} \left(\widehat{ATE_k}^{rep} - ATE_k^{rep}\right)^2}{N^{rep}}$. The coverage rate was the proportion of simulated datasets for which the true sample value falls into the confidence interval. Based on previous simulation research (e.g., Muthén & Muthén, 2002), the relative bias (or coverage rate) is considered as satisfactory if the relative biases (or coverage rates) across the J cluster-specific treatment effects had lower and upper quartiles that fall between −0.10 and 0.10 (or fall between 0.91 and 0.98).

Results

Scenario 1. Randomized Treatment Assignments, Linear Data-Generation Models

In Scenario 1 (Figures 1-2 show the results), where the treatment assignment was randomized and the data-generation models involved only linear terms of the covariates, the estimators performed satisfactorily overall.

In terms of bias (Figure 1), for the estimators that rely on accurate estimation of the outcome mean ("trt-y" and "cluster-y"), the estimators using the random-intercept outcome regression (labeled with "_yRE) had larger bias than the estimators using the fixed-intercept regression (labeled with "_yFE). This is consistent with the shrinkage issue of random-effect modeling (e.g., Raudenbush & Bryk, 2002).

In comparison, while the outcome mean was also involved, the triply-robust estimators performed similarly well in general, regardless of whether the outcome mean was estimated by the random- or fixed-intercept outcome regression, or by the double machine learning procedure; this is consistent with the robustness property of estimation with the triply-robust formula. Additionally, the confidence intervals of the triply-robust estimators (Figure 2) had generally

satisfactory coverage rates.

In terms of MSE (Figure S1 in the supplemental materials), the estimator "trt-cluster" (which does not involve the outcome mean) had larger MSE than the other estimators. For the other estimators (which involve the outcome mean), the MSEs were generally similar.

Scenario 2. Nonrandomized Treatment Assignments, Linear/Quadratic Data-Generation Models

In Scenario 2 (Figures 3-10 show the results), when all of the true models involved only linear covariates terms (under panel "True: Trt, Cluster, Y all linear" in Figures 3-4), the parametric models for all of the nuisance functions had correct specifications, and the estimators performed satisfactorily in terms of bias.

When one of the true models involved nonlinear (quadratic) terms of covariates (Figures 3-4), the triply-robust estimators had smaller bias than the other estimators overall. The coverage rates (Figures 5-6) of their confidence intervals were generally satisfactory.

Among the other estimators, the estimator "trt-cluster" (obtained with Eq. 2) had larger bias, when the fitted model was misspecified for the treatment probability or cluster assignment probability (under panels "True: Trt quadratic" and "True: Cluster quadratic", respectively). The estimators "trt-y" and "cluster-y" (both of which involve the outcome mean) were more biased, when the model for the outcome mean was misspecified (under panel "True: Y quadratic"). The patterns of the MSE were generally similar to the relative bias (Figures S2-S3 in the supplemental materials).

When all the true models involved nonlinear (quadratic) terms of covariates (Figures 7-10), the triply-robust estimator implemented with the double machine learning ("triply robust (DML)") had smaller bias than the other estimators, and its bias decreased with larger samples.

To further examine the bias decreasing pattern of this estimator ("triply robust (DML)"), we conducted additional simulations with J = 40 clusters and larger total sample sizes of n = 4,000 and 20,000. The results (Figure 8) showed that with the other estimators, the bias remained large as the sample size increased. In contrast, with the triply-robust estimator implemented with the double machine learning ("triply robust (DML)"), the bias vanished with larger sample sizes; this shows a consistent pattern with its theoretical asymptotic behavior. For the MSEs (see Figures S3-S5 of the supplemental materials), the patterns were generally similar to the biases. For the confidence interval with the triply-robust estimator (Figures 9-10), the coverage rates were too low when it was implemented using the misspecified parametric models (labeled with "_linear"), and were more satisfactory when it was implemented with the double machine learning procedure (although slight undercoverage still occurred).

Empirical Illustration

To illustrate the developed methods, we use a simulated-data example. The data were simulated based on a real-life intervention study (Reed et al., 2019), in which a PND was used to examine the effects of participation in a summer reading program (T_i) on students' reading performance (Y_i). In evaluating summer programs, randomized assignments are often infeasible (e.g., Kim & Quinn, 2013; Reed et al., 2019). In Reed et al. (2019), the summer program participation (T_i) was nonrandomized. The treatment arm included 470 students who chose to participate ($T_i = 1$) and the control arm included 823 students who did not participate ($T_i = 0$). The control arm had a no-treatment control condition (i.e., unclustered). In the treatment arm, the summer program was provided in J = 34 teachers/classes ("clusters" in this illustration; one teacher was paired with one class of students); the class sizes from 8 to 15 students with a median of 12. In assigning the treatment arm students to the teachers/classes (i.e., the cluster

assignment K_i), the assignment was nonrandomized, and students with similar instructional needs tended to be assigned to the same clusters.

For illustration purposes, we simulated example data to mimic these design aspects of the PND in Reed et al. (2019). Specifically, using data-generation models similar to those in Scenario 1 of the simulation study, we simulated each student's summer-program participation status (T_i), teacher/class assignment (K_i), and reading outcome score (Y_i). The outcome score was standardized to facilitate interpretation (e.g., Weiss et al., 2017). For the baseline covariates ($\mathbf{X}_i$), to better mimic real-life covariate data, we used the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K: 1998; there were also kindergarten students in Reed et al., 2019). From the base year data of the ECLS-K, we selected covariates that might influence a student's participation in summer program (T_i), instructional needs (factors influence K_i), and reading performance (Y_i); a total of 25 covariates were selected, including students' math and reading test scores, demographics (e.g., sex, race), learning behaviors and psychological traits (e.g., approaches to learning, self-control, internal and external problem behaviors), parents' and household characteristics (e.g., parents' education, income, books at home).

To demonstrate the heterogeneity (vs. homogeneity) of the estimated ATE_k's when the cluster-specific treatment effects are truly heterogenous vs. homogeneous, we created two example datasets. In the first dataset, the true values of ATE_k, $k = 1, \ldots, J$ were heterogeneous, whereas in the second dataset the true values were the same. The example data and R code used in this illustration are available at https://github.com/xliu12/PND.heter/tree/main/simulations_and_examples.

With each example dataset, we applied the triply-robust estimator to estimate the cluster-specific treatment effects (i.e., ATE_k, $k = 1, \ldots, J$, where J = 34, the number of teachers/classes

["clusters"] in the treatment arm), considering that it showed relatively satisfactory performance through the simulations. The estimator was implemented it with the double machine learning procedure in the same way as the simulation study.

The results from the two example datasets are presented in Figures 11-12. For each dataset, the histogram (Figure 11) shows the distribution of the estimated effects of summer program participation across the teachers/classes in the treatment arm of the PND (i.e., the estimated ATE_k's). In the caterpillar plot (Figure 12), the teacher/class-specific effect estimates are plotted in order from lowest to highest, each with its 95% confidence interval; thus, the more heterogeneous these effects are, the steeper the slope will be.

For illustration, we interpret the results using the substantive context of the PND in Reed et al. (2019). In this context, the causal effect ATE_k for teacher/class (cluster) k represents the effect of participating in the summer program (vs. no-participation) for the students who would be assigned to the teacher/class k (the class with teacher k, i.e., cluster k) if they chose to participate in the program (i.e., with $K_i(1) = k$). Identifying the ATE_k's requires IA1 and IA2; IA1 states that a student's participation in the program (vs. no-participation; T_i) is independent of the student's potential teacher/class assignment and potential outcomes, given the baseline covariates ($\mathbf{X}_i$); IA2 states that the student's potential teacher/class assignment [i.e., the potential cluster assignment $K_i(1)$] is independent of the student's potential outcome if not participating in the program [i.e., the potential outcome, $Y_i(0)$], given the baseline covariates ($\mathbf{X}_i$). Under IA1 and IA2, the estimation results of the ATE_k's can provide empirical information on the teacher/class-specific (i.e., cluster-specific) program participation effects in the PND.

With the second dataset (truly homogeneous ATE_k's), the estimated program participation effects ($\widehat{ATE_k}$'s) have relatively little variation across the teachers/classes (clusters)

in the treatment arm (standard deviation: 0.12; range: -0.20 to 0.30). The histogram of these effects appears concentrated (Figure 11), and the caterpillar plot (Figure 12) has a relatively flat slope, indicating relatively low effect variation.

With the first dataset (truly heterogeneous ATE_k's), the estimated program participation effects ($\widehat{ATE_k}$'s) appear more heterogeneous (standard deviation: 1.07; range: -1.63 to 3.28). The histogram is less concentrated, and the caterpillar plot shows a steeper slope. These results can indicate that the effect of a student's participation in the summer program varied appreciably, depending on the teacher/class to which the student would be assigned. Despite the effect variation, with the simulated example dataset, only a small proportion (9/34) of the teachers/classes in the treatment arm were associated with negative effects (i.e., for students who would be assigned to these teachers/classes if they chose to participate in the program, the average outcome if participation would be lower than if no-participation).

Discussion

In recent years, there is a growing attention to PNDs and to heterogenous causal effects (e.g., Imbens, 2024; Lohr et al., 2014; Raudenbush & Bloom, 2015; Sterba, 2017). For PNDs, despite the possible heterogeneity in the causal effects of treatment across clusters, there had been limited research on methods to assess such (possibly heterogenous) causal effects across clusters. To reduce the research gap, in this study, we developed methods to define, identify, and estimate the cluster-specific treatment effects in PNDs. Our study adds to the literature on PNDs in the following ways.

First, extending the principal stratification framework (Frangakis & Rubin, 2002), we obtained the causal effect definition for the cluster-specific treatment effect (ATE_k). Our definition overcomes the challenge that for the PND, the observed cluster assignment (K) is

observed after the treatment assignment (a posttreatment variable) and cannot be conditioned on in defining a treatment effect. By instead conditioning on the potential cluster assignment $K(1)$, the ATE_k definition represents the causal effect of assignment to treatment (vs. control) for individuals who would be assigned to the cluster k (if they were assigned to the treatment arm).

Second, based on the principal score approach (e.g., Ding & Lu, 2017; Jiang et al., 2022), we obtained identification formulas for the cluster-specific treatment effects (ATE_k's). The identification requires the ignorability of the treatment assignment and ignorability of the potential cluster assignment given measured baseline covariates (IA1 and IA2). Thus, for empirically studying the heterogeneous cluster-specific treatment effects in PNDs where randomization is infeasible, it is important to collect a rich set of baseline covariates that might affect the treatment assignment, cluster assignment, and outcome.

Third, the identification formulas led to various estimators of the cluster-specific treatment effect (ATE_k), obtained by combining two or three out of the three nuisance functions. In particular, extending the principal score literature (Jiang et al., 2022), we obtained the triply-robust estimator of the ATE_k. When implemented with parametric models, the triply-robust estimator is robust to misspecification of the model for estimating any one of the nuisance functions, so long as the models are correctly specified for the other two nuisance functions. Furthermore, triply-robust estimator facilitates using data-adaptive techniques (e.g., machine learning) to assist in estimating the causal effect (ATE_k), such as via the double machine learning procedure (e.g., (Chernozhukov et al., 2018).

Lastly, for PNDs where the treatment and/or cluster assignment is randomized, we described the simplified identification formulas and estimators. Particularly, we showed how the cluster-specific treatment effect (ATE_k), as well as its between-cluster variability, can be

connected to parameters of a multilevel outcome model commonly used in PNDs (e.g., Bauer et al., 2008; Sterba, 2017). With these results, we help clarify the causal interpretation of the multilevel model parameters.

Limitations and Future Directions

The identification formulas require IA1 and IA2. Thus, it is critical to examine the sensitivity of the ATE_k estimation to potential violations of these assumptions (IA1 and IA2). Sensitivity analysis methods have been available in the principal score literature (e.g., Ding & Lu, 2017; Jiang et al., 2022; Nguyen et al., 2023); extending them to sensitivity analysis for assessing ATE_k's in PNDs is an important future direction.

In addition, the current study assumes the absence of noncompliance and missing data issues; but these issues occur in practice and warrant future research. Recent progress has been made on handling these issues in randomized PNDs and other types of clustered designs (e.g., (Enders, 2022; Jo, Asparouhov, & Muthén, 2008; Lüdtke et al., 2017; Roberts, 2021; Schochet & Chiang, 2011; Schweig & Pane, 2016, 2016; van Buuren, 2011; Yang & Gaskin, 2023). It could be important to extend the previous research to nonrandomized PNDs and to the inference of cluster-specific treatment effects in PNDs.

The simulation study could be expanded. For example, the design conditions could cover more levels of sample sizes and numbers of clusters, other levels of intra-cluster correlations, different ratios of outcome residual variances between study arms (e.g., Baldwin et al., 2011; Candlish et al., 2018; Sanders, 2011). The performance evaluation measures could include measures used in previous simulation studies on heterogeneous treatment effects and on PNDs (e.g., Baldwin et al., 2011; Kennedy, 2023; Lyu et al., 2023; Wager & Athey, 2018). The data-generation models could cover more complex functional forms (e.g., cubic covariate terms) and

noncontinuous (e.g., binary) outcome (e.g., Roberts et al., 2016).

Besides the principal score approach we considered, there are other possible ways for assessing causal effects defined with principal stratification (Imbens & Rubin, 1997, 2015; Jin & Rubin, 2008; Page et al., 2015; Zhang et al., 2009). Exploring the feasibility of other approaches and adapting them for studying cluster-specific treatment effects in PNDs could be helpful.

There are other research questions on effect heterogeneity in PNDs that await future studies. For example, with the cluster-specific treatment effects ATE_k's we examined, it could be interesting to study methods to assess moderators of these effects. There were previous studies examining moderators of treatment arm outcomes in randomized PNDs (e.g., Cox et al., 2022; Sterba et al., 2014); it could be helpful to extend them to investigate moderators of the ATE_k's in PNDs. Furthermore, we considered a basic PND structure, and extensions to more complex PND structures could be worthwhile (e.g., PNDs with three or more levels of nesting; e.g., Lohr et al., 2014; Luo et al., 2015; Roberts & Walwyn, 2013; Sterba, 2017).

Conclusion

In conclusion, echoing the growing emphasis on effect heterogeneity and growing attention to treatment-induced partial nesting, we hope that our study can provide useful insights and methods for assessing heterogeneous treatment effects across clusters in PNDs.

Reference

Allison, P. D. (2009). *Fixed Effects Regression Models*. SAGE Publications.

Angrist, J. D., Hull, P. D., Pathak, P. A., & Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, *132*(2), 871–919. https://doi.org/10.1093/qje/qjx001

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. http://www.jstor.org/stable/2291629

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Austin, P. C. (2005). Bias in penalized quasi-likelihood estimation in random effects logistic regression models when the random effects are not normally distributed. *Communications in Statistics - Simulation and Computation*, *34*(3), 549–565. https://doi.org/10.1081/SAC-200068364

Austin, P. C., & Leckie, G. (2020). Bootstrapped inference for variance parameters, measures of heterogeneity and random effects in multilevel logistic regression models. *Journal of Statistical Computation and Simulation*, *90*(17), 3175–3199. https://doi.org/10.1080/00949655.2020.1797738

Austin, P. C., Naylor, C. D., & Tu, J. V. (2001). A comparison of a Bayesian vs. A frequentist method for profiling hospital performance. *Journal of Evaluation in Clinical Practice*, *7*(1), 35–45. https://doi.org/10.1046/j.1365-2753.2001.00261.x

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2023). *DoubleML -- An object-oriented implementation of double machine learning in R* [Paper]. arXiv.org. https://econpapers.repec.org/paper/arxpapers/2103.09603.htm

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods*, *16*(2), 149–165. https://doi.org/10.1037/a0023464

Baldwin, S. A., & Imel, Z. (2013). Therapist effects: Findings and methods. In *Bergin and Garfield's handbook of psychotherapy and behavior change* (pp. 258–297).

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., & Grothendieck, G. (2011). Package `lme4'. *Linear Mixed-Effects Models Using S4 Classes. R Package Version*, *1*(6).

Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research*, *43*(2), 210–236. https://doi.org/10.1080/00273170802034810

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842. https://doi.org/10.1080/19345747.2016.1264518

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59. https://www.jstor.org/stable/30128044

Brand, J. E., Zhou, X., & Xie, Y. (2023). Recent developments in causal inference and machine learning. *Annual Review of Sociology*, *49*(1), 81–110. https://doi.org/10.1146/annurev-soc-030420-015345

Candel, M. J. J. M., & Winkens, B. (2003). Performance of empirical bayes estimators of level-2 random parameters in multilevel analysis: A monte carlo study for longitudinal designs. *Journal of Educational and Behavioral Statistics*, *28*(2), 169–194. https://doi.org/10.3102/10769986028002169

Candlish, J., Teare, M. D., Dimairo, M., Flight, L., Mandefield, L., & Walters, S. J. (2018). Appropriate statistical methods for analysing partially nested randomised controlled trials with continuous outcomes: A simulation study. *BMC Medical Research Methodology*, *18*(105), 1–17. https://doi.org/10.1186/s12874-018-0559-x

Chang, T.-H., & Stuart, E. A. (2022). Propensity score methods for observational studies with clustered data: A review. *Statistics in Medicine*, *41*(18), 3612–3626. https://doi.org/10.1002/sim.9437

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., & implementation), Xgb. contributors (base Xgb. (2023). *xgboost: Extreme Gradient Boosting* (1.7.5.1) [Computer software]. https://cran.r-project.org/web/packages/xgboost/index.html

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. https://doi.org/10.1111/ectj.12097

Compas, B. E., Forehand, R., Keller, G., Champion, J. E., Rakow, A., Reeslund, K. L., McKee, L., Fear, J. M., Colletti, C. J., Hardcastle, E., & others. (2009). Randomized controlled trial of a family cognitive-behavioral preventive intervention for children of depressed parents. *Journal of Consulting and Clinical Psychology*, *77*(6), 1007–1020. https://doi.org/10.1037/a0016930

Cox, K., & Kelcey, B. (2022). Statistical power for detecting moderation in partially nested designs. *American Journal of Evaluation*, *44*(1), 133–152. https://doi.org/10.1177/1098214020977692

Cox, K., Kelcey, B., & Luce, H. (2022). Power to detect moderated effects in studies with three-level partially nested data. *The Journal of Experimental Education*, *0*(0), 1–24. https://doi.org/10.1080/00220973.2022.2130130

Díaz, I. (2020). Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics (Oxford, England)*, *21*(2), 353–358. https://doi.org/10.1093/biostatistics/kxz042

Ding, P., & Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(3), 757–777. https://doi.org/10.1111/rssb.12191

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*(1), 43–68. https://doi.org/10.1214/18-STS667

Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.

Farrell, P. J., MacGibbon, B., & Tomberlin, T. J. (1997). Empirical bayes estimators of small area proportions in multistage designs. *Statistica Sinica*, *7*(4), 1065–1083.

Feller, A., Grindal, T., Miratrix, L., & Page, L. C. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics*, *10*(3), 1245–1285. https://doi.org/10.1214/16-AOAS910

Feller, A., Mealli, F., & Miratrix, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics*, *42*(6), 726–758. https://doi.org/10.3102/1076998617719726

Follmann, D. A. (2000). On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*, *95*(452), 1101–1109. https://doi.org/10.1080/01621459.2000.10474306

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29. https://doi.org/10.1111/j.0006-341x.2002.00021.x

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), 385–409. https://doi.org/10.2307/2983325

Hastie, T. (2017). Generalized additive models. In *Statistical models in S* (pp. 249–307). Routledge. https://www.taylorfrancis.com/chapters/edit/10.1201/9780203738535-7/generalized-additive-models-trevor-hastie

Hastie, T. (2023). *gam: Generalized Additive Models* (1.22-2) [Computer software]. https://cran.r-project.org/web/packages/gam/index.html

Hedges, L. V., & Citkowicz, M. (2015). Estimating effect size when there is clustering in one treatment group. *Behavior Research Methods*, *47*(4), 1295–1308. https://doi.org/10.3758/s13428-014-0538-z

Hill, J., Waldfogel, J., & Brooks-Gunn, J. (2002). Differential effects of high-quality child care. *Journal of Policy Analysis and Management*, *21*(4), 601–627. https://doi.org/10.1002/pam.10077

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*(1), 69–88. https://doi.org/10.1093/biostatistics/1.1.69

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hong, G. (2004). *Causal inference for multi-level observational data with application to kindergarten retention*. University of Michigan. https://search.proquest.com/openview/ddfd7b7994e62eb541ec3074f7430a2b/1?pq-origsite=gscholar&cbl=18750&diss=y

Hong, G., & Raudenbush, S. W. (2006). Evaluating Kindergarten Retention Policy. *Journal of the American Statistical Association*, *101*(475), 901–910. https://doi.org/10.1198/016214506000000447

Huber, M. (2023). *Causal Analysis: Impact Evaluation and Causal Machine Learning with Applications in R*. MIT Press.

Imbens, G. W. (2024). Causal Inference in the Social Sciences. *Annual Review of Statistics and Its Application*, *11*(1), annurev-statistics-033121-114601. https://doi.org/10.1146/annurev-statistics-033121-114601

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, *25*(1), 305–327. https://www.jstor.org/stable/2242722

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jiang, Z., & Ding, P. (2021). Identification of causal effects within principal strata using auxiliary variables. *Statistical Science*, *36*(4), 493–508. https://doi.org/10.1214/20-STS810

Jiang, Z., Yang, S., & Ding, P. (2022). Multiply robust estimation of causal effects under principal ignorability. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(4), 1423–1445. https://doi.org/10.1111/rssb.12538

Jin, H., & Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, *103*(481), 101–111. https://www.jstor.org/stable/27640023

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, *13*(4), 314. https://doi.org/10.1037/a0014207

Jo, B., Asparouhov, T., & Muthén, B. (2008). Intention-to-treat analysis in cluster randomized trials with noncompliance. *Statistics in Medicine*, *27*(27), 5565–5577. https://doi.org/10.1002/sim.3370

Jo, B., Asparouhov, T., Muthén, B., Ialongo, N. S., & Brown, C. H. (2008). Cluster randomized trials with treatment noncompliance. *Psychological Methods*, *13*(1), 1–18. https://doi.org/10.1037/1082-989X.13.1.1

Jo, B., & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, *28*(23), 2857–2875. https://doi.org/10.1002/sim.3669

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539. https://www.jstor.org/stable/27645858

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008–3049. https://doi.org/10.1214/23-EJS2157

Kim, J. S., & Quinn, D. M. (2013). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, *83*(3), 386–431. https://doi.org/10.3102/0034654313483906

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, *25*(3), 602–627. https://doi.org/10.1093/ectj/utac015

Lee, K. J., & Thompson, S. G. (2005). Clustering by health professional in individually randomised trials. *BMJ : British Medical Journal*, *330*(7483), 142–144.

Linero, A. R., & Zhang, Q. (2022). Mediation analysis using Bayesian tree ensembles. *Psychological Methods*, Advance online publication. https://doi.org/10.1037/met0000504

Liu, X., Liu, F., Miller-Graff, L., Howell, K. H., & Wang, L. (2023). Causal inference for treatment effects in partially nested designs. *Psychological Methods*, Advanced Online Publication. https://doi.org/10.1037/met0000565.supp

Lohr, S., Schochet, P., & Sanders, E. (2014). Partially nested randomized controlled trials in education research: A guide to design and analysis. *National Center for Education Research*. https://ies.ed.gov/ncer/pubs/20142000/index.asp

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141–165. https://doi.org/10.1037/met0000096

Luo, W., Cappaert, K. J., & Ning, L. (2015). Modelling partially cross-classified multilevel data. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 342–362. https://doi.org/10.1111/bmsp.12050

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. https://doi.org/10.3102/10769986221115446

McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, *24*(1), 20–35. https://doi.org/10.1037/met0000182

Nguyen, T. Q., Stuart, E. A., Scharfstein, D. O., & Ogburn, E. L. (2023). *Sensitivity analysis for principal ignorability violation in estimating complier and noncomplier average causal effects* (arXiv:2303.05032). arXiv. https://doi.org/10.48550/arXiv.2303.05032

Normand, S.-L. T., Ash, A. S., Fienberg, S. E., Stukel, T. A., Utts, J., & Louis, T. A. (2016). League tables for hospital comparisons. *Annual Review of Statistics and Its Application*, *3*(1), 21–50. https://doi.org/10.1146/annurev-statistics-022513-115617

Page, L. C., Feller, A., Grindal, T., Miratrix, L., & Somers, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation*, *36*(4), 514–531. https://doi.org/10.1177/1098214015594419

Polley, E. C., & van der Laan, M. S. (2017). *super learner prediction: R package, version 2.0-21*.

R Core Team, Rf. & others. (2018). *R: A language and environment for statistical computing*. R foundation for statistical computing Vienna, Austria.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475–499. https://doi.org/10.1177/1098214015600515

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

Reed, D. K., Aloe, A. M., Reeger, A. J., & Folsom, J. S. (2019). Defining summer gain among elementary students with or at risk for reading disabilities. *Exceptional Children*, *85*(4), 413–431. https://journals.sagepub.com/doi/pdf/10.1177/0014402918819426

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338. https://doi.org/10.1037/met0000184

Ripley, B., & Venables, W. (2023). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models* (7.3-19) [Computer software]. https://cran.r-project.org/web/packages/nnet/index.html

Roberts, C. (1999). The implications of variation in outcome between health professionals for the design and analysis of randomized controlled trials. *Statistics in Medicine*, *18*(19), 2605–2615. https://doi.org/10.1002/(SICI)1097-0258(19991015)18:19

Roberts, C. (2021). The implications of noncompliance for randomized trials with partial nesting due to group treatment. *Statistics in Medicine*, *40*(2), 349–368. https://doi.org/10.1002/sim.8778

Roberts, C., Batistatou, E., & Roberts, S. A. (2016). Design and analysis of trials with a partially nested design and a binary outcome measure. *Statistics in Medicine*, *35*(10), 1616–1636. https://doi.org/10.1002/sim.6828

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, *2*(2), 152–162. https://doi.org/10.1191/1740774505cnO76oas

Roberts, C., & Walwyn, R. (2013). Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Statistics in Medicine*, *32*(1), 81–98. https://doi.org/10.1002/sim.5521

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591–593. https://doi.org/10.2307/2287653

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/10.1080/01621459.1986.10478355

Sanders, E. A. (2011). *Multilevel analysis methods for partially nested cluster randomized trials* [PhD Thesis]. University of Washington.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313. https://doi.org/10.1037/a0014268

Schochet, P. Z. (2015). Statistical theory for the RCT-YES software: Design-based causal inference for RCTs. *National Center for Education Evaluation and Regional Assistance*. https://ies.ed.gov/ncee/rel/regions/central/pdf/CE5.3.2-Statistical-Theory-for-the-RCT-YES-Software-DesignBased-Causal-Inference-for-RCTs.pdf

Schochet, P. Z., & Chiang, H. S. (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics*, *36*(3), 307–345. https://doi.org/10.3102/1076998610375837

Schweig, J. D., & Pane, J. F. (2016). Intention-to-treat analysis in partially nested randomized controlled trials with real-world complexity. *International Journal of Research & Method in Education*, *39*(3), 268–286. https://doi.org/10.1080/1743727X.2016.1170800

Serlin, R. C., Wampold, B. E., & Levin, J. R. (2003). Should providers of treatment be regarded as a random factor? If it ain't broke, don't" fix" it: A comment on Siemer and Joormann

(2003). *Psychological Methods*, *8*(4), 524–534. https://doi.org/10.1037/1082-989X.8.4.524

Sterba, S. K. (2017). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research*, *27*(4), 425–436. https://doi.org/10.1080/10503307.2015.1114688

Sterba, S. K., Preacher, K. J., Forehand, R., Hardcastle, E. J., Cole, D. A., & Compas, B. E. (2014). Structural equation modeling approaches for analyzing partially nested data. *Multivariate Behavioral Research*, *49*(2), 93–118. https://doi.org/10.1080/00273171.2014.882253

Stice, E., Shaw, H., Burton, E., & Wade, E. (2006). Dissonance and healthy weight eating disorder prevention programs: A randomized efficacy trial. *Journal of Consulting and Clinical Psychology*, *74*(2), 263–275. https://doi.org/10.1037/0022-006X.74.2.263

Stuart, E. A., & Jo, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Statistical Methods in Medical Research*, *24*(6), 657–674. https://doi.org/10.1177/0962280211421840

Tchetgen Tchetgen, E. J., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics*, *40*(3), 1816–1845. https://doi.org/10.1214/12-aos990

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. https://doi.org/10.1080/00273171.2011.569395

van Buuren, S. (2011). *Multiple imputation of multilevel data*. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203848852-14/multiple-imputation-multilevel-data-stef-van-buuren

van Der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1). https://doi.org/10.2202/1544-6115.1309

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, *20*(6), 880–883. https://doi.org/10.1097/ede.0b013e3181bd5638

VanderWeele, T. J., & Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, *1*(1), 1–20. https://doi.org/10.1515/jci-2012-0002

Vowels, M. J. (2023). Misspecification and unreliable interpretations in psychology and social science. *Psychological Methods*, *28*(3), 507–526. https://doi.org/10.1037/met0000429

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. https://doi.org/10.1080/01621459.2017.1319839

Walwyn, R., & Roberts, C. (2010). Therapist variation within randomised trials of psychotherapy: Implications for precision, internal and external validity. *Statistical Methods in Medical Research*, *19*(3), 291–315. https://doi.org/10.1177/0962280209105017

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites?

Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876. https://doi.org/10.1080/19345747.2017.1300719

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*, 1–17. https://doi.org/10.18637/jss.v077.i01

Yang, M., & Gaskin, D. J. (2023). Handling missing data in partially clustered randomized controlled trials. *Psychological Methods*, Advanced Online Publication. https://doi.org/10.1037/met0000612.supp

Zhang, J. L., Rubin, D. B., & Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, *104*(485), 166–176.

Zheng, W., & van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In M. J. van der Laan & S. Rose (Eds.), *Targeted Learning: Causal Inference for Observational and Experimental Data* (pp. 459–474). Springer. https://doi.org/10.1007/978-1-4419-9782-1_27