Fair and Robust Estimation of Heterogeneous Treatment Effects for Optimal Policies in Multilevel Studies

Youmi Suk [*1], Chan Park [†2], Chenguang Pan [‡1], and Kwangho Kim [§3]

[1]Department of Human Development, Teachers College Columbia University
[2]Department of Statistics, University of Illinois Urbana-Champaign
[3]Department of Statistics, Korea University

Sep 26, 2024

Abstract

Recently, there have been growing efforts in developing fair algorithms for treatment effect estimation and optimal treatment recommendations to mitigate discriminatory biases against disadvantaged groups. While most of this work has primarily focused on addressing discrimination due to individual-level attributes (e.g., race/ethnicity), it overlooks the broader impact of societal structures and cultural norms (e.g., structural racism) beyond the individual level. In this paper, we formalize the concept of multilevel fairness for estimating heterogeneous treatment effects to improve fairness in optimal policies. Specifically, we propose a general framework for the estimation of conditional average treatment effects under multilevel fairness constraints that incorporate individual-level sensitive variables, cluster-level sensitive variables, and their combinations. Using this framework, we analyze the trade-off between fairness and the maximum achievable utility by the optimal policy. We evaluate the effectiveness of our framework through a simulation study and a real data study on advanced math courses using data from the High School Longitudinal Study of 2009.

Keywords: Heterogeneous treatment effects, Conditional average treatment effects, Fairness, Intersectionality, Fair algorithms, Optimal treatment regimes, Multilevel observational data, Math course-taking

1 Motivation: Multilevel Fairness

In recent decades, shifting from the average treatment effect (ATE), researchers have increasingly turned to heterogeneous treatment effects or *conditional average treatment effects* (CATEs), especially using machine learning methods (e.g., Hill, 2011; Kennedy, 2023; Künzel et al., 2019; Wager & Athey, 2018). The CATE is defined as the ATE among subgroups determined by pretreatment covariates, and allows for more individualized treatment effects. These estimates are often used to inform optimal treatment recommendations. For example, a data-driven recommendation model constructs an optimal rule or policy, where a treatment is recommended if the

[*]ysuk@tc.columbia.edu
[†]parkchan@illinois.edu
[‡]cp3280@tc.columbia.edu
[§]kwanghk@korea.ac.kr

CATE estimate exceeds a certain threshold, and otherwise, it is not (e.g., Ballarini et al., 2018; Lipkovich et al., 2017; Tsiatis et al., 2019). However, this type of model may be susceptible to potential discriminatory biases against individuals' characteristics, particularly those identified by race, gender, or other historically discriminated attributes. These characteristics are often called *sensitive* or *protected* variables. To mitigate discriminatory biases in CATEs or policies, a few technical solutions have been proposed, such as Kim and Zubizarreta (2023), Nabi et al. (2019), and Rateike et al. (2024). However, none have explicitly addressed the importance of considering discrimination arising from individual-level and cluster-level sensitive variables in fairness investigations. The overarching goal of this paper is to propose a framework for estimating CATEs with fairness considerations to develop fair and optimal polices in multilevel studies, where individuals (e.g., students) are nested within clusters (e.g., schools).

Ensuring algorithmic fairness has emerged as a critical issue across various research communities. The relevant literature offers formal, measurable metrics of fairness to detect discriminatory biases and provides both technical and non-technical solutions to resolve these biases (e.g., Barocas et al., 2023; Pessach & Shmueli, 2022). However, most work has primarily focused on individual-level sensitive variables without addressing the broader impact of societal structures and cultural norms. For example, recent studies have highlighted the issues with framing race analytically as an individual-level attribute, but this view overlooks systemic racism and other broader factors related to race (Boyd et al., 2020; Hanna et al., 2020; VanderWeele & Robinson, 2014). To more accurately represent discrimination and unfairness, therefore, it is essential to use a multilevel approach: one that includes individual experiences and characteristics, and another that includes cluster-level phenomena.

As a concrete example, consider estimating the CATE of taking an advanced math course on math scores in 9th grade to develop a math-course recommendation model. In historical educational data, participation in advanced math has been skewed towards white students rather than black students at the student level (given their prior achievement scores) (Byun et al., 2014; Dalton et al., 2007). Additionally, at the school level, participation in advanced math has varied based on school racial composition, such as the proportions of black or Hispanic groups that make up a school's student population. Such structural disparities in advanced math courses are undesirable, as they have contributed to an increase in the achievement gap and a lack of diversity in STEM fields (Byun et al., 2014; Sadler et al., 2014). Without fairness considerations, these disparities or biases in data are likely to be passed on to recommendation systems and potentially reinforce unfair dependencies between (multilevel) sensitive variables, decisions, and outcomes.

The main goal of this paper is to propose a framework for estimating the CATE under multilevel fairness constraints and to apply this framework to mitigate unfairness in optimal policies in multilevel studies. This paper accommodates sensitive variables from both the individual and cluster levels and seamlessly addresses intersectionality across multiple sensitive variables. Intersectionality is a theoretical framework that examines how overlapping categories like gender and race interact to create unique inequalities and challenges for disadvantaged or marginalized groups (Crenshaw, 1989; Foulds et al., 2020). Specifically, we extend the approach by Kim and Zubizarreta (2023) to multilevel contexts by characterizing multilevel fairness functions to define desirable fairness criteria and by accounting for clustering effects in propensity score and outcome models. We then evaluate the performance of our proposed approach by measuring the accuracy of the estimated CATE, the utility and fairness of the policy, and the trade-off efficiency between utility and fairness. Finally, we demonstrate the proposed approach by estimating a CATE-based, math-course recommendation model for 9th graders using data from the High School Longitudinal Study of 2009 (HSLS:09).

2 Prior Work and Our Contributions

Many prior studies on algorithmic fairness in machine learning provide formal, measurable metrics for group fairness. Group fairness aims to ensure that different groups determined by a sensitive variable should be treated similarly by an algorithm (Barocas et al., 2023).[1] Example metrics include statistical parity, separation, and sufficiency. Statistical parity requires that an algorithm's prediction (or decision) be marginally independent of the sensitive variable. Separation (or equalized odds) requires that the prediction be independent of the sensitive variable conditional on the outcome. Sufficiency requires that the outcome be independent of the sensitive variable conditional on the prediction; see Chapter 3 of Barocas et al. (2023) for a review on fairness notions.

Additionally, the group fairness metrics above can be modified by conditioning on additional variables (e.g., Corbett-Davies et al., 2017; Suk & Han, 2023) or relaxed by allowing for a certain amount of disparities between the subgroups (e.g., Feldman et al., 2015). Furthermore, these metrics can be converted to counterfactual versions by replacing the observed outcome with the potential control outcome (i.e., the outcome if individuals were untreated) (Coston et al., 2020; Mishler et al., 2021).[2] They can also be expanded to address intersectionality, for example by ensuring statistical parity among intersectional subgroups defined by gender and race (Foulds et al., 2020; Suk & Han, 2024). Except for highly unrealistic conditions, several of the fairness metrics cannot be simultaneously satisfied on the same data, and thus, researchers should choose which criterion to target based on their domain knowledge (Berk et al., 2021).

Beyond fairness metrics, recent discussions on causal fairness have sought to provide more intuitive explanations of unfairness in algorithms. Many of relevant studies advocate for fairness by blocking unfair pathways of the sensitive variable on the decision or prediction (Chiappa, 2019; Kusner et al., 2017; Mhasawade & Chunara, 2021; Nabi et al., 2019; Yang et al., 2021). An interesting point in these works is that counterfactuals are defined with respect to the sensitive variable (e.g., black versus white) rather than with respect to a decision variable (e.g., advanced math versus standard math). That is, they focus on questions like "What would the recommended math course be if a student had been of a different race their whole life?" rather than "What would the math score be if this student took an advanced math course?" This approach often faces controversy over whether it is meaningful to discuss counterfactuals of individuals' demographics since these are not typically manipulated (Glymour & Glymour, 2014; Holland, 1986). On the other hand, other causal fairness approaches, including Kim and Zubizarreta (2023) and Mishler and Kennedy (2022), define counterfactuals with respect to a decision variable and use them inside an algorithm's risk function under fairness constraints; our proposal also adopts this definition of counterfactuals. Regardless of the approaches, understanding fairness-utility trade-offs is crucial when designing fair algorithms because the most accurate models do not satisfy desired fairness criteria and the fairest models do not yield the maximum utility (e.g., Chan et al., 2024; Mishler et al., 2021). Therefore, it is essential to inspect the tension between fairness and utility, as a fair but ineffective algorithm is of little use in practice.

To the best of our knowledge, there is no work on the intersection of algorithmic fairness and the estimation of heterogeneous treatment effects in multilevel studies that incorporates the multidimensionality of sensitive variables from both the individual and cluster levels. Some studies have integrated aspects of algorithmic fairness and policy learning (Kim & Zubizarreta, 2023; Nabi et al., 2019; Viviano & Bradic, 2023), but all have focused on data where the study units are assumed to be independent and identically distributed (i.i.d.). One notable

[1] Another category of fairness metrics is individual fairness, which ensures that similar individuals should be treated similarly by an algorithm.

[2] Coston et al. (2020) and Mishler et al. (2021) consider potential outcomes regarding a decision variable.

exception is by Mhasawade and Chunara (2021) on multilevel fairness, but it focuses on the ATE where the counterfactuals depend on a sensitive variable. Therefore, it remains unclear how to extend existing methods in algorithmic fairness to design fair algorithms with different aspects of fairness issues in multilevel data.

Our contribution is to propose a general framework for estimating the CATE under multilevel fairness constraints and to apply the proposed framework to improve the fairness of optimal policies in multilevel studies. To achieve this, we integrate literature from algorithmic fairness, causal inference, multilevel modeling, and stochastic optimization. In this paper, we detail how to characterize multilevel fairness functions that accommodate individual-level sensitive variables, cluster-level sensitive variables, and their combinations. Following Kim and Zubizarreta (2023), we formulate our estimator with these multilevel fairness constraints as a convex optimization problem, which can be readily solved using off-the-shelf solvers. The resulting estimator is doubly robust under certain regularity conditions. We also demonstrate how the concept of intersectionality can be seamlessly integrated into our fairness functions to conduct more accurate fairness investigations in multilevel studies. Furthermore, we use this framework to analyze the trade-off between utility and fairness in the optimal policy, using visualization inspection and formal trade-off metrics recently developed by Chan et al. (2024). Lastly, we emphasize the importance of considering multilevel fairness in a real-world application using multilevel educational data. Our analysis reveals that an algorithm considered fair at the individual level may not be fair from a cluster-level perspective. These findings highlight the necessity and effectiveness of our approach for estimating fair heterogeneous treatment effects to develop fair optimal policies in multilevel settings.

The remainder of the paper is organized as follows. Section 3 presents the setup, and Section 4 discusses our framework and methods that account for multilevel fairness. Section 5 provides the design and results of our simulation study, and Section 6 demonstrates our framework in empirical data about advanced math courses in high school. Finally, discussion and conclusions are provided in Section 7.

3 Setting

3.1 Notation and Potential Outcomes

Let $j = 1, \ldots, J$ index J clusters (e.g., schools), and let $i = 1, \ldots, n_j$ index individuals (i.e., study units), where n_j is the number of individuals within cluster j and the total sample size is $n = \sum_{j=1}^{J} n_j$. Let $T_{ij} \in \mathcal{T} = \{0, 1\}$ denote treatment assignment (e.g., advanced math course assignment), where $T_{ij} = 1$ indicates the treated status of individual i in cluster j and $T_{ij} = 0$ indicates the untreated/control status. For example, $T_{ij} = 1$ means that individual i in cluster j (hereafter referred to as individual ij) takes the advanced math course, while $T_{ij} = 0$ means they take the standard math course. The observed outcome for individual ij is denoted as Y_{ij}, where larger values are assumed to be preferable without loss of generality (e.g., math achievement scores). We denote measured p-dimensional pre-treatment covariates for individual ij as $W_{ij} = (S_{ij}, X_{ij}) \in \mathcal{W}$, which consists of q-dimensional sensitive/protected variables $S_{ij} \in \{0, 1\}^q$ and $(p - q)$-dimensional insensitive/unprotected covariates $X_{ij} \in \mathcal{X}$. Here, $S_{ij} = (S_{ij}^{(1)}, S_j^{(2)})$ consists of q_1-dimensional individual-level sensitive variables $S_{ij}^{(1)} = \{S_{1ij}^{(1)}, S_{2ij}^{(1)}, \ldots, S_{q_1ij}^{(1)}\}$ and q_2-dimensional cluster-level sensitive variables $S_j^{(2)} = \{S_{1j}^{(2)}, S_{2j}^{(2)}, \ldots, S_{q_2j}^{(2)}\}$; likewise, $X_{ij} = (X_{ij}^{(1)}, X_j^{(2)})$ consists of individual-level (insensitive) covariates $X_{ij}^{(1)}$ and cluster-level (insensitive) covariates $X_j^{(2)}$. Thus, for each individual, we observe $O_{ij} = (Y_{ij}, T_{ij}, W_{ij})$.

We also use the potential outcomes notation (Neyman, 1923; Rubin, 1974) to define causal effects. We denote $Y_{ij}(1)$ as the potential treatment outcome if individual ij were treated (i.e.,

$T_{ij} = 1$), and $Y_{ij}(0)$ as the potential control outcome if individual ij were untreated (i.e., $T_{ij} = 0$). The observed outcome is linked to the potential outcomes as $Y_{ij} = T_{ij}Y_{ij}(1) + (1 - T_{ij})Y_{ij}(0)$ under the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1986), which implies that individuals' potential outcomes are independent of others' treatment assignments and that there are no hidden variations of the treatment.

3.2 Heterogeneous Treatment Effects and Decision Rule

Heterogeneous treatment effects are commonly characterized by the CATE. The CATE measures the ATE among a subgroup of individuals defined by pre-treatment covariates W_{ij}. In particular, we are interested in the CATE for the population *conditional on the current cluster membership*, that is, the CATE of individual treatment effects with equal weights over all the individuals in the population given their observed cluster assignments:

$$\tau(w) = E\{Y_{ij}(1) - Y_{ij}(0) \mid W_{ij} = w\}. \tag{1}$$

In our empirical example, if W_{ij} represents students' math identity, where $W_{ij} = 1$ means that they see themselves a math person and $W_{ij} = 0$ means that they do not, $\tau(1)$ is the ATE of taking an advanced math course on math scores among those who identify as a math person and $\tau(0)$ is the ATE among those who do not. Alternatively, if W_{ij} is defined by prior math scores, $\tau(60)$ is the ATE among students who scored 60, and $\tau(65)$ is the ATE among students who scored 65. However, unlike the binary case of math identity, when W_{ij} is defined by prior math scores, it requires estimating a function of a continuous variable. To estimate the ATE for students with a prior score of 63, i.e., $\tau(63)$, ideally, we need data from students who scored exactly 63. Otherwise, interpolation (or extrapolation) from students with nearby scores, like 60 or 65 is necessary, and the accuracy of the interpolation relies on assumptions about the functional form of $\tau(w)$. In this paper, our estimand of interest is the CATE under a series of multilevel fairness constraints, which will be further described in Section 4.1.

To reduce reliance on the functional assumptions, we instead target the best linear approximation of the CATE, defined by its projection onto the space spanned by linear combinations of covariates, and is formally written as:

$$\beta' = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; E\left[\left\{\tau(W_{ij}) - W_{ij}^\top \beta\right\}^2\right]. \tag{2}$$

Here, the linear function $W_{ij}^\top \beta'$ can be viewed as the best linear approximation of a potentially complex function $\tau(W_{ij})$. This kind of projection approach is often employed in causal inference (e.g., Kim & Zubizarreta, 2023; Semenova & Chernozhukov, 2020; Suk & Kang, 2022). There are several reasons why the above projection approach is preferred in our setting. First, even if the true function $\tau(W_{ij})$ is non-linear, the coefficients β' are still interpretable, especially, as the linear effect of changing W_{ij} on the treatment effect. Second, we can use doubly robust methods to estimate β', even when various constraints are present. Third, we can use off-the-shelf optimization algorithms to efficiently perform our estimation procedure. These points will be elaborated in the following section.

Researchers often use the CATE (or its transformation) to develop a *decision rule* or *policy* $d \in \mathcal{D}$, where $\mathcal{D}$ denotes the set of all possible treatment rules. The decision rule is a function that maps an individual's covariates W_{ij} to a treatment option in $\mathcal{T}$ (Tsiatis et al., 2019), i.e., $d: \; \mathcal{W} \to \mathcal{T}, \;\; W_{ij} \to d(W_{ij})$. In particular, an optimal treatment regime or optimal policy d^* is the "best" decision rule that maximizes the expected utility or *value* $\mathcal{V}(d)$ as:

$$d^*(W_{ij}) = \underset{d \in \mathcal{D}}{\operatorname{argmax}} \, \mathcal{V}(d), \tag{3}$$

where the standard value function is $\mathcal{V}(d) := E[d(W_{ij})Y_{ij}(1) + \{1 - d(W_{ij})\}Y_{ij}(0)]$. In this case, we can alternately express the optimal policy maximizing the value $\mathcal{V}(d)$ as:

$$d^*(W_{ij}) = \mathbb{1}\{\tau(W_{ij}) > 0\}. \tag{4}$$

Here, $\mathbb{1}$ represents an indicator function. This optimal rule recommends the treatment if the CATE is positive and otherwise, recommends the control.

In practice, researchers may need to restrict decision rules in order to make the rules more interpretable and easier to implement. We denote a restricted subset of $\mathcal{D}$ as $\mathcal{D}_\beta$ where β is a finite-dimensional parameter that defines the subset of decision rules considered by researchers. For example, we can restrict the form of CATE using the linear projection approach in Equation (2). With this restricted rule, we estimate the policy that maximizes the achievable value as:

$$d^*(W_{ij}; \beta') = \mathbb{1}(W_{ij}^\top \beta' > 0). \tag{5}$$

3.3 Causal Assumptions

The usual set of the working assumptions to identify the CATE in multilevel studies is (Imbens & Rubin, 2015):

(A1) Conditional Ignorability: $Y_{ij}(1), Y_{ij}(0) \perp T_{ij} \mid W_{ij}$

(A2) Positivity: $0 < P(T_{ij} = 1 \mid W_{ij}) < 1$.

Assumption (A1) states that within every value of the pre-treatment covariates W_{ij}, the treatment assignment T_{ij} is randomly assigned to individuals and thus, is independent of $Y_{ij}(1)$ and $Y_{ij}(0)$. Assumption (A2) states that for every value of the covariates, the probability of receiving treatment (i.e., the propensity score) is between zero and one. Therefore, (A2) ensures that the probability of having clusters containing only treated or only control units is zero. However, (A2) may not be empirically supported by the observed data, as there may be some clusters where all units receive either treatment or control. (A1) and (A2) are jointly referred to as *strong ignorability* (Rosenbaum & Rubin, 1983).

4 Our Proposal: Fair and Robust CATE in Multilevel Studies

4.1 Estimand

In this section, we provide a framework for estimating the CATE that maximizes the utility (i.e., value) while satisfying the desirable fairness criteria in multilevel studies. Following Kim and Zubizarreta (2023), we aim to find the best linear approximation of the CATE, defined by its projection onto a covariate space, but subject to K multilevel fairness constraints as:

$$\begin{aligned} \beta^* = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ & \mathbb{E}\left[\left\{Y_{ij}(1) - Y_{ij}(0) - W_{ij}^\top \beta\right\}^2\right] \\ \text{subject to} \quad & \left|\mathbb{E}\left\{\mathrm{uf}_k(O_{ij}) W_{ij}^\top \beta\right\}\right| \leq \delta_k, \quad k = 1, \ldots, K. \end{aligned} \tag{6}$$

Our estimand is now reformulated as the solution to the constrained stochastic optimization problem (6), where we do not assume anything about the true functional relationship between individual treatment effects (i.e., $Y_{ij}(1) - Y_{ij}(0)$) and covariates W_{ij}. The *(un)fairness function* $\mathrm{uf}_k : \mathcal{W} \to \mathbb{R}$ characterizes a desired fairness criterion specified by users to be incorporated into the CATE estimation (Mishler & Kennedy, 2022). Each uf_k is defined with sensitive variables from the individual and/or cluster levels, and a larger value indicates an increasing level of unfairness or larger disparity. In our setting, the CATE is fair not only if individuals from

different demographic groups have similar CATE but also if individuals from different cluster-level backgrounds have similar CATE estimates; see the next subsection for more details. The term δ_k is a pre-specified tolerance for the maximum acceptable level of unfairness in the k-th fairness criterion, where $\delta_k \geq 0$. Under this objective function, we aim to find the coefficients of the best-fitting function of individual treatment effects on the linear projection, subject to the K fairness constraints.[3]

4.2 Multilevel Fairness Functions

Multilevel fairness functions use sensitive variables from different structural levels, including individual-level variables, cluster-level variables, and their combinations. These functions are specified based on fairness metrics discussed in Section 2 by addressing different aspects of fairness concerns in multilevel data.

Individual Level

Suppose we have one individual-level sensitive variable, $S^{(1)}_{1ij}$ (e.g., black versus white students), without loss of generality. The statistical parity function with $S^{(1)}_{1ij}$ is as follows:

$$\text{uf}_k(O_{ij}) = \frac{1 - S^{(1)}_{1ij}}{\mathbb{E}\{1 - S^{(1)}_{1ij}\}} - \frac{S^{(1)}_{1ij}}{\mathbb{E}\{S^{(1)}_{1ij}\}}. \tag{7}$$

This criterion allows our model to be marginally independent of the sensitive variable, and it leads to:

$$|\mathbb{E}\{W_{ij}^\top \beta | S^{(1)}_{1ij} = 0\} - \mathbb{E}\{W_{ij}^\top \beta | S^{(1)}_{1ij} = 1\}| \leq \delta_k.$$

For example, statistical parity aims to have similar means of CATE between white versus black groups, with a tolerance level δ_k. Additionally, we can use the conditional statistical parity as our fairness criterion:

$$\text{uf}_k(O_{ij}) = \frac{(1 - S^{(1)}_{1ij})\mathbb{1}(L_{ij} = l)}{\mathbb{E}\left\{(1 - S^{(1)}_{1ij})\mathbb{1}(L_{ij} = l)\right\}} - \frac{S^{(1)}_{1ij}\mathbb{1}(L_{ij} = l)}{\mathbb{E}\left\{S^{(1)}_{1ij}\mathbb{1}(L_{ij} = l)\right\}}, \tag{8}$$

where L_{ij} is a legitimate (or fair) factor used to specify the conditional statistical parity. L_{ij} is some function of X_{ij}, and it is categorical or categorized to be used within the fairness function, i.e., $L_{ij} \in \{1, 2, \ldots, h\}$. This criterion achieves our model's conditional independence of the sensitive variable given the legitimate/fair variable, and requires fitting h fairness constraints. Equation (8) leads to:

$$|\mathbb{E}\{W_{ij}^\top \beta | S^{(1)}_{1ij} = 0, L_{ij} = l\} - \mathbb{E}\{W_{ij}^\top \beta | S^{(1)}_{1ij} = 1, L_{ij} = l\}| \leq \delta_k.$$

Under this criterion, among students who have the same prior achievement levels, black and white groups have similar means of CATEs, with a tolerance level δ_k. Furthermore, researchers can use another fairness metric known as separation/equalized odds to allow our model to be independent of the sensitive variable conditional on the outcome. This can be done by replacing L_{ij} with a categorized version of Y_{ij} in Equation (8).

Importantly, all the fairness functions with $S^{(1)}_{1ij}$ above try to achieve fairness *across clusters* rather than *within clusters*. However, one may hope to protect these fairness criteria within

[3]To fit a more flexible approximation of the CATE, one may consider its projection onto a much larger finite-dimensional parametric model space spanned by the B distinct basis functions $b(W) = [b_1(W), \ldots, b_B(W)] \in \mathbb{R}^B$, including intersection terms or quadratic terms, instead of the linear projection; see Section 2.3 of Kim and Zubizarreta (2023) for details.

each cluster. More specifically, if we want to achieve within-cluster statistical parity, Equation (7) needs to be changed into a type of conditional statistical parity as:

$$\text{uf}_k(O_{ij}) = \frac{(1 - S_{1ij}^{(1)})\mathbb{1}(C_j = c)}{\mathbb{E}\left\{(1 - S_{1ij}^{(1)})\mathbb{1}(C_j = c)\right\}} - \frac{S_{1ij}^{(1)}\mathbb{1}(C_j = c)}{\mathbb{E}\left\{S_{1ij}^{(1)}\mathbb{1}(C_j = c)\right\}}, \tag{9}$$

where $C_j \in \{1, 2, ..., J\}$ represents the cluster membership or identifier. This criterion creates J fairness constraints and aims to achieve marginal independence within each cluster. Other fairness functions, such as Equation (8), can be easily adapted to within-cluster versions by additionally conditioning on the cluster membership.

We make two remarks about within-cluster fairness. First, while these notions of fairness may be of interest in applied research, they are often not feasible to use in finite samples when (i) some schools contain only one subgroup (e.g., only white students) or (ii) cluster sizes are small. Second, in such cases, researchers can group clusters based on their similarity (e.g., in terms of covariate distributions or treatment prevalence) and use the group membership that consists of similar clusters. This strategy helps to alleviate the two problems while achieving a more accurate representation of fairness in multilevel data.

Cluster Level

It is straightforward to characterize multilevel fairness functions with cluster-level sensitive variables (e.g., Hispanic-serving institution versus not). We apply fairness functions across clusters, like statistical parity in Equation (7), conditional statistical parity in Equation (8), or separation with a categorized outcome, to define fairness functions for cluster-level sensitive variables. Specifically, we replace $S_{1ij}^{(1)}$ with a cluster-level sensitive variable, say $S_{1j}^{(2)}$, in chosen across-cluster fairness functions to define desirable fairness criteria.

Intersectionality

Intersectionality is often addressed by forming intersectional groups based on multiple protected variables (Foulds et al., 2020; Russell & Kaplan, 2021; Suk & Han, 2024). Given this, we also form intersectional subgroups defined by multiple sensitive variables to achieve intersectional fairness. In multilevel data settings, these intersectional groups can be defined using individual-level sensitive variables (e.g., $S_{1ij}^{(1)}, S_{2ij}^{(1)}$), cluster-level sensitive variables (e.g., $S_{1j}^{(2)}, S_{2j}^{(2)}$), or a combination of both (e.g., $S_{1ij}^{(1)}, S_{1j}^{(2)}$).

To characterize intersectional fairness, suppose that intersectional subgroups are defined by one individual-level sensitive variable $S_{1ij}^{(1)}$ and one cluster-level sensitive variable $S_{1j}^{(2)}$, and the target fairness criterion is statistical parity. We denote intersectional subgroups as $S_{ij}^{*} = (S_{1ij}^{(1)}, S_{1j}^{(2)}) \in \{(0,0), (0,1), (1,0), (1,1)\}$, where the first group is a reference group and the last three groups are focal groups. The statistical parity function for intersectionality is as:

$$\text{uf}_k(O_{ij}) = \frac{1 - \mathbb{1}(S_{ij}^{*} = s)}{\mathbb{E}\left\{1 - \mathbb{1}(S_{ij}^{*} = s)\right\}} - \frac{\mathbb{1}(S_{ij}^{*} = s)}{\mathbb{E}\left\{\mathbb{1}(S_{ij}^{*} = s)\right\}}, \tag{10}$$

where $\mathbb{1}(S_{ij}^{*} = s)$ is the indicator for intersectional group membership representing the focal group s. This criterion aims to achieve marginal independence across all intersectional subgroups simultaneously. Similarly, we can apply other fairness functions to these intersectional subgroups, for example, by conditioning on a legitimate/fair variable, using a categorical version of Y, or considering cluster membership. It is important to note that Equation (10) for intersectionality differs from Equation (8), even if we replace L_{ij} with $S_{1j}^{(2)}$ in Equation (8),

because Equation (8) does not ensure similar means of CATEs with respect to $S_{1j}^{(2)}$. Therefore, researchers must carefully define the appropriate forms of fairness criteria in their real-world contexts.

4.3 Estimation

Once multilevel fairness functions are chosen by users, we can solve the above objective function in Equation (6) using an appropriate approximate program constructed from observed data. Let φ_t denote the augmented inverse probability weighted estimator for the potential outcomes as:

$$\varphi_t(O;\eta) = \frac{\mathbb{1}(T=t)\{Y-\mu_t(W)\}}{\pi_t(W)} + \mu_t(W) \tag{11}$$

where $\eta = \{\pi_t, \mu_t\}$ denotes a set of nuisance functions with $\pi_t(W_{ij}) = P(T_{ij} = t|W_{ij})$ and $\mu_t(W_{ij}) = \mathbb{E}(Y_{ij}|W_{ij}, T_{ij} = t)$. Estimator φ_t is doubly robust, which means that it is consistent if either the propensity score model π_t or outcome model μ_t is consistent, not necessarily both; see Park and Kang (2022) for technical details. To estimate the nuisance functions η, one may use appropriate parametric estimation techniques for multilevel or clustered data, such as generalized linear mixed effects models (GLMMs, McCulloch et al., 2008) or the generalized estimating equations (GEE, Liang & Zeger, 1986). We estimate our nuisance functions with the following GLMMs:

$$logit(\pi_1(W_{ij})) = \gamma_0 + \gamma_W^\top W_{ij} + V_j, \quad V_j \sim N(0, \sigma_V^2), \tag{12}$$

$$Y_{ij} = \beta_0 + \beta_T T_{ij} + \beta_W^\top W_{ij} + \beta_{TW}^\top T_{ij} W_{ij} + U_j + R_{ij}, \tag{13}$$

$$\text{with} \;\; U_j \sim N(0, \sigma_U^2), \quad R_{ij} \sim N(0, \sigma_R^2).$$

Here, $\gamma_{(\cdot)}$s and $\beta_{(\cdot)}$s represent the fixed-effects coefficients for the propensity score model (12) and outcome model (13), respectively. V_j and U_j denote the random effect terms in the propensity score model and outcome model, respectively.

After obtaining the estimates of η, we then estimate β^* via the following constrained quadratic program:

$$\begin{aligned} \hat{\beta}^* = &\underset{\beta}{\text{argmin}} \;\; \frac{1}{2}\beta^\top \mathbb{P}_n(WW^\top)\beta - \mathbb{P}_n\left[\{\varphi_1(O;\hat{\eta}) - \varphi_0(O;\hat{\eta})\}\, W^\top\right]\beta \\ &\text{subject to} \quad \left|\mathbb{P}_n\left\{\widehat{\text{uf}}_k(O)W^\top\right\}\beta\right| \le \delta_k, \quad k = 1, \ldots, K. \end{aligned} \tag{14}$$

Here, $\mathbb{P}_n$ denotes the sample average operator adapted to the cluster structure, i.e., $\mathbb{P}_n\{f(O)\} = J^{-1}\sum_{j=1}^{J} n_j^{-1}\sum_{i=1}^{n_j} f(O_{ij})$, and the estimated fairness functions $\widehat{\text{uf}}_k(O)$ are computed based on the sample data. The estimator $\hat{\beta}^*$ can be readily implemented using various efficient off-the-shelf solvers, say by using the alternating direction method of multipliers (e.g., Stellato et al., 2020), or by converting it into a conic or semidefinite programming problem (e.g., MOSEK ApS, 2019). Furthermore, our proposed estimator (14) is doubly robust, which means that it is consistent even if one of the two nuisance models is misspecified. This property is formally stated in the following proposition, which is a direct consequence of Theorem 3.2 of Kim and Zubizarreta (2023) and Theorem 2 of Hansen and Lee (2019).

Proposition 1. *Let β^* and $\hat{\beta}^*$ be the optimal solutions to the programs* (6) *and* (14)*, respectively. Suppose that the nuisance regression functions π_t, μ_t are parametrically modeled through* (12) *and* (13)*, respectively. Further assume that each cluster size is uniformly bounded, i.e., there exists $M < \infty$ such that $n_j < M, \forall j$. Then the proposed estimator is $\sqrt{n}$-consistent, i.e., $\|\beta^* - \hat{\beta}^*\|_2 = O_{\mathbb{P}}(n^{-1/2})$, as long as one of the nuisance regression functions, but not necessarily both, is correctly specified.*

For practical implementation, in Algorithm 1, we summarize the steps of our proposed method using multilevel observational data. We provide two functions, named `GLMM_model` and `fairCATE_multilevel`. The `GLMM_model` function is based on `R` package `lme4` (Bates et al., 2015) and is used to estimates φ_t. The `fairCATE_multilevel` function is based on `R` package `Rmosek` (MOSEK ApS, 2019) for the MOSEK optimization and is used to estimate the fair and robust CATE (i.e., $W_{ij}^\top \hat{\beta}^*$) and the CATE-based treatment decision (i.e., $\hat{d}^*(W_{ij}) = \mathbb{1}(W_{ij}^\top \hat{\beta}^* > 0)$). `R` codes for our proposal are available in the supplementary materials and can also be found at the first author's GitHub repository (https://github.com/youmisuk/FairCATE-Multilevel).

Algorithm 1 Fair and Robust CATE Estimation in Multilevel Studies

Input: Outcome Y_{ij}, treatment T_{ij}, pre-treatment covariates $W_{ij} = (S_{ij}, X_{ij})$

1: Choose multilevel fairness functions, uf_k and their tolerance levels δ_k.

2: Estimate nuisance models $\eta = \{\pi_t, \mu_t\}$ and use $\hat{\eta}$ to compute φ_t for $t = 0, 1$.

3: Estimate β^* by minimizing the following objective function with fairness constraints.

$$\hat{\beta}^* = \underset{\beta}{\text{argmin}} \ \frac{1}{2}\beta^\top \mathbb{P}_n(WW^\top)\beta - \mathbb{P}_n\left[\{\varphi_1(O;\hat{\eta}) - \varphi_0(O;\hat{\eta})\}\, W^\top\right]\beta$$
$$\text{subject to} \quad \left|\mathbb{P}_n\left\{\widehat{\text{uf}}_k(O)W^\top\right\}\beta\right| \leq \delta_k, \quad k = 1, \ldots, K.$$

4: Estimate the fair and robust CATE, $W_{ij}^\top \hat{\beta}^*$ for all individuals ij.

5: Make the optimal decision based on the CATE: $\hat{d}^*(W_{ij}) = \mathbb{1}\{W_{ij}^\top \hat{\beta}^* > 0\}$.

Output: $W_{ij}^\top \hat{\beta}^*$ and $\hat{d}^*(W_{ij})$ for all individuals ij.

5 Simulation Study

5.1 Designs and Evaluation

We conduct a simulation study to assess the performance of our proposed method. Throughout the simulations, we estimate the CATE and the optimal decisions using our proposal discussed in Section 4 with varying tolerance levels of δ_k. For comparison, we use two popular existing estimators for CATE based on machine learning: Bayesian additive regression trees (BART, Hill, 2011) and causal forests (Wager & Athey, 2018). However, these estimators do not allow for fairness constraints in their CATE estimation.

Specifically, our simulation study is divided into two designs. Design 1 introduces unfairness in both individual-level and cluster-level sensitive variables without an intersectionality issue. Design 2 is based on Design 1, but adds unfairness among subgroups defined by individual-level and cluster-level sensitive covariates. For simplicity, our simulation uses one individual-level sensitive variable $S_{1ij}^{(1)}$, one cluster-level sensitive variable $S_{1j}^{(2)}$, and fairness metrics based on statistical parity, such as Equation (7). We provide details of the data-generating processes for both designs in Appendix S1. We also demonstrate the application of another fairness metric in Section 6 on real data analysis. For each design, we use a sample size of 300 clusters and 25 individuals per cluster.

In each replicate, we evaluate the performance of the methods with respect to (i) the mean squared error (MSE) of the CATE estimates (i.e., $\hat{\tau}(W) = W^\top \hat{\beta}^*$), (ii) the relative utility of a target optimal policy ($\hat{d}^* = \mathbb{1}(W^\top \hat{\beta}^* > 0)$), and (iii) the mean unfairness of the target policy

across fairness functions ($k = 1, 2, \ldots, K$) as:

$$\text{Mean Squared Error (MSE)} = \mathbb{P}_n \left\{ \hat{\tau}(W) - \tau(W) \right\}^2, \tag{15}$$

$$\text{Relative Utility} = \frac{\mathcal{V}(\hat{d}^*) - \mathcal{V}(\hat{d}_{\text{random}})}{\mathcal{V}(\hat{d}_{\text{random}})}, \tag{16}$$

$$\text{Mean Unfairness} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{U}_k(\hat{d}^*), \;\; \mathcal{U}_k(\hat{d}^*) = \left| \mathbb{P}_n \{ \widehat{\text{uf}}_k(O) \hat{d}^* \} \right|. \tag{17}$$

Here, the MSE measures the mean squared deviation between the estimated CATE and the true CATE across all study units. The relative utility measures the increase in value $\mathcal{V}$ of the target policy, compared to a random policy $\hat{d}_{\text{random}}$ (i.e., randomly assigned recommendations). The mean unfairness measures the average differences in the proportions of recommended treatments across the K fairness functions investigated, where $K = 2$ for Design 1 and $K = 3$ for Design 2. Additionally, we compute the mean unfairness in CATE estimates, and provide the results with their accuracy and fairness in Appendix S2.

For our proposed method, we further consider two trade-off metrics: Fairness-Utility Relative Gain (FURG) and Fairness-Utility Trade-off Ratio (FUTR). These metrics are recently proposed in Chan et al. (2024), and are written as:

$$\text{Fairness-Utility Relative Gain (FURG)} = \text{UG} + \text{UD}, \tag{18}$$

$$\text{Fairness-Utility Trade-off Ratio (FUTR)} = -\frac{\text{UD}}{\text{UG}}, \tag{19}$$

$$\text{where} \;\; \text{UG} = \frac{\mathcal{V}(\hat{d}^*_{\text{fair}}) - \mathcal{V}(\hat{d}^*_{\text{unfair}})}{\mathcal{V}(\hat{d}^*_{\text{unfair}}) - \mathcal{V}(\hat{d}_{\text{random}})}, \quad \text{UD} = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathcal{U}_k(\hat{d}^*_{\text{unfair}}) - \mathcal{U}_k(\hat{d}^*_{\text{fair}})}{\mathcal{U}_k(\hat{d}^*_{\text{unfair}})}.$$

Here, $\hat{d}^*_{\text{unfair}}$ represents the target unfair policy estimated using our proposed method with a large value of δ, and $\hat{d}^*_{\text{fair}}$ represents the target fair policy with a small value of δ. The FURG measures the total combined gain that sums utility gain (UG) and unfairness drop (UD), giving equal importance to both. The FUTR metric is the negative ratio between UD and UG, and computes the reduction in unfairness per unit of utility loss.[4] For both FURG and FUTR, a larger metric value indicates a better fairness-utility trade-off. We repeat our simulation 500 times, and compute the average score for each performance metric across these repetitions.

5.2 Results

Table 1 presents the MSE (of CATE estimates), relative utility, and mean unfairness for each method under Design 1 with no intersectionality. We use the true CATE $\tau(W)$ as our reference and compare our results with two popular CATE methods, BART and causal forests. For our proposed method, we consider two scenarios: the first only with an individual-level fairness function (i.e., an insufficient representation of fairness) and the second with both an individual-level and a cluster-level fairness function (i.e., accurate representation of fairness). In both scenarios, we use two tolerance levels: $\delta = \infty$ and $\delta = 0$. Specifically, $\delta = \infty$ indicates no fairness constraints, which we set $\delta = 4$ in our simulations, and results in an unfair policy; $\delta = 0$ indicates zero tolerance for unfairness, which we set $\delta = 0.0001$ for computational efficiency, and results in a fair policy.

From Table 1, our proposed method without fairness constraints (i.e., both scenarios with $\delta = \infty$) demonstrates performance comparable to BART in terms of relative utility and mean

[4]Following Chan et al. (2024), we replace $\min(\text{UG}, -0.01)$ to maintain the minimum utility loss of 0.01.

Table 1: Results of the methods under Design 1: No intersectionality

	MSE	Utility	Unfairness	FURG	FUTR
True CATE	0.000	0.121	0.305	-	-
BART	0.043	0.120	0.305	-	-
causal forests	0.188	0.113	0.225	-	-
Ours					
[1] uf_1 with $S_{1ij}^{(1)}$					
$\delta = \infty$	0.086	0.118	0.294	-	-
$\delta = 0$	0.340	0.107	0.130	0.319	4.506
[2] uf_1 with $S_{1ij}^{(1)}$ and uf_2 with $S_{1j}^{(2)}$					
$\delta = \infty$	0.086	0.118	0.294	-	-
$\delta = 0$	0.424	0.103	0.010	0.814	7.592

NOTE: uf_k represents the k-th (un)fairness function. $S_{1ij}^{(1)}$ and $S_{1j}^{(2)}$ represent individual-level and cluster-level sensitive variables, respectively. MSE represents the mean squared error of the conditional average treatment effect estimates. "Utility" and "Unfairness" represent the relative utility and mean unfairness of the optimal policy, respectively. FURG represents fairness-utility relative gain, and FUTR represents fairness-utility trade-off ratio.

unfairness, but with a slightly larger MSE of the CATE estimates. While neither BART nor causal forests can accommodate fairness constraints, our proposed method does so by reducing δ to zero. With a zero value of δ, our proposed method effectively reduces unfairness. Specifically, in the second scenario, which incorporates both individual-level and cluster-level fairness constraints, the remaining unfairness decreases to 0.010. In contrast, the first scenario, which includes only the individual-level fairness constraint, still exhibits residual unfairness due to the absence of cluster-level fairness consideration. Additionally, we observe a trade-off between utility and fairness; reducing unfairness comes at a slight cost of sacrificing the maximum achievable value (approximately 1% loss). The second scenario shows a better fairness-utility trade-off compared to the first scenario as indicated by higher FURG and FUTR values. These findings highlight the importance of formalizing a multilevel approach to algorithmic fairness in multilevel studies.

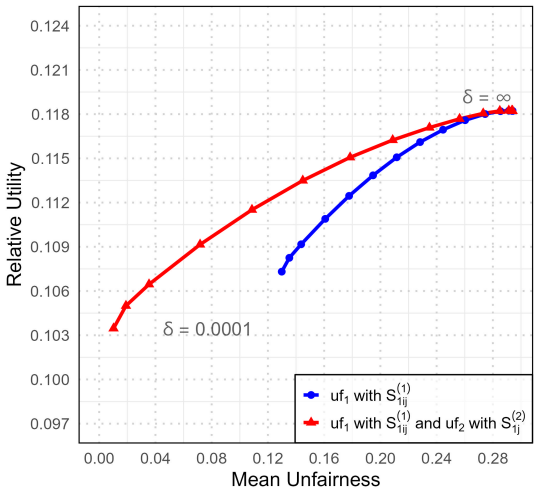


Figure 1: Trade-off between utility and unfairness under Design 1 with varying threshold values, from $\delta = \infty$ (i.e., an unfair model with no constraints) to $\delta = 0.0001$ (i.e., a fair model). The blue curve represents the scenario with only an individual-level fairness function. The red curve represents the scenario with multilevel fairness, where both individual-level and cluster-level fairness functions are specified.

We further examine the trade-off between utility and fairness using our proposed method with varying values of δ. Figure 1 visualizes this trade-off. In both scenarios, there are no noticeable changes in utility and fairness when δ is decreased to a certain threshold (from $\delta = 4$ to $\delta = 0.9$ in our case). However, when δ is reduced further from 0.9 to 0.0001, the mean unfairness effectively decreases, albeit with a slight reduction in relative utility. Between the first and second scenarios, the second scenario with multilevel fairness constraints demonstrates a more favorable trade-off, where a greater reduction in unfairness comes with only a small loss in utility. The remaining unfairness in the first scenario emphasizes that an algorithm considered fair at the individual level may still exhibit unfairness from the cluster level.

Table 2 summarizes the results for Design 2 with intersectionality concerns, where we generate unfairness among intersectional groups defined by an individual-level sensitive variable and a cluster-level sensitive variable; see Appendix S1 for more details of the data-generating process. In this design, we examine the performance of our proposed method with three different scenarios: the first with only an individual-level fairness function, the second with individual-level and cluster-level fairness functions, and the third with intersectional fairness functions. The first two scenarios neglect intersectionality, while the third scenario ensures intersectional fairness in our proposed method. From Table 2, we observe a trade-off between utility and unfairness, similar to that in Design 1. As the value of δ is reduced from a large value to zero (specifically, 0.0001), the mean unfairness decreases across all three scenarios, but with a 1% loss in relative utility. We achieve a more favorable trade-off when we specify our fairness functions more accurately, moving from an individual-level perspective (the first scenario) to an intersectional one (the third scenario). The first scenario shows the largest residual unfairness and the lowest FURG and FUTR values. The second scenario still exhibits some residual unfairness due to intersectional disparities. In contrast, the third scenario, which accounts for intersectional fairness, effectively reduces unfairness to 0.012, and shows the highest FURG and FUTR values.

Table 2: Results of the methods under Design 2: Intersectionality

	MSE	Utility	Unfairness	FURG	FUTR
True	0.000	0.135	0.269	-	-
BART	0.046	0.133	0.275	-	-
causal forests	0.217	0.121	0.169	-	-
Ours					
[1] uf_1 with $S_{1ij}^{(1)}$					
$\delta = \infty$	0.095	0.131	0.262	-	-
$\delta = 0$	0.293	0.119	0.065	0.338	4.911
[2] uf_1 with $S_{1ij}^{(1)}$ and uf_2 with $S_{1j}^{(2)}$					
$\delta = \infty$	0.095	0.131	0.262	-	-
$\delta = 0$	0.283	0.119	0.036	0.729	9.356
[3] uf_k with $(S_{1ij}^{(1)}, S_{1j}^{(2)})$ for $k = 1, 2, 3$					
$\delta = \infty$	0.097	0.131	0.262	-	-
$\delta = 0$	0.286	0.119	0.012	0.850	10.781

NOTE: uf_k represents the k-th (un)fairness function. $S_{1ij}^{(1)}$ and $S_{1j}^{(2)}$ represent individual-level and cluster-level sensitive variables, respectively. MSE represents the mean squared error of the conditional average treatment effect estimates. "Utility" and "Unfairness" represent the relative utility and mean unfairness of the optimal policy, respectively. FURG represents fairness-utility relative gain, and FUTR represents fairness-utility trade-off ratio.

Figure 2 visualizes the trade-off between utility and fairness using our proposed method with different values of δ. For large values of δ, all three scenarios behave similarly in terms of relative utility and mean fairness. However, when the value of δ approaches zero, the mean unfairness decreases, with a slight reduction in relative utility. However, in the first and second

scenarios, non-negligible unfairness remains due to disparities arising from either cluster-level sensitive variables or intersectional inequalities. In contrast, in the third scenario, we minimize unfairness as much as possible.

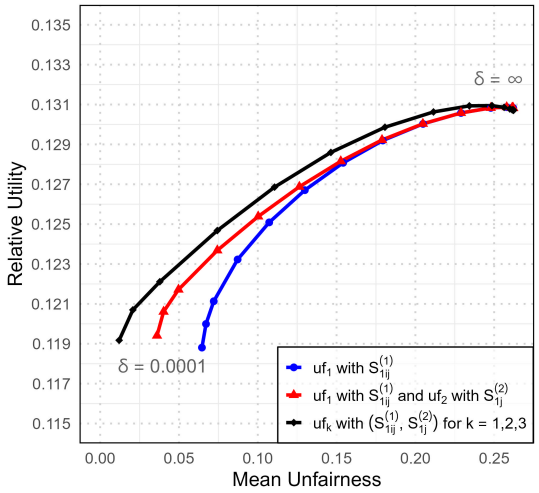


Figure 2: Trade-off between utility and unfairness under Design 2 with varying threshold values, from $\delta = \infty$ (i.e., an unfair model with no constraints) to $\delta = 0.0001$ (i.e., a fair model). The blue, red, and black curves represent the scenarios with fairness functions on the individual level only, on both the individual and cluster levels, and on each intersectional group, respectively.

Overall, the findings from Designs 1 and 2 highlight the importance of incorporating multilevel fairness considerations beyond the individual level when estimating CATEs and developing optimal policies in multilevel studies. By correctly specifying fairness functions, our method provides a flexible approach to minimizing unfairness while maintaining utility.

6 Empirical Example: Advanced Math Course in High School

6.1 Data and Variables

We use the restricted-use version of the HSLS:09 data, a nationally representative longitudinal study of 9th graders in 2009, to estimate the CATEs of taking an advanced math course and to develop fair and optimal math course-taking plans. We also use transcript data collected from the schools in 2013 after most students had completed high school. The transcript data includes detailed information on the math courses taken by each student, credits earned in each math course, GPA in mathematics, and more. In developing our recommendation models, we combine student data, transcript data, and school data from the HSLS:09 study. The analytic sample consists of 9,910 students from 800 schools. Note that numbers are rounded to the nearest tens to comply with IES rules.

The treatment variable is whether students took an advanced math course in 9-th grade during the academic year of 2009-10; $T_{ij} = 1$ indicates that they took an advanced math course (i.e., geometry or a higher-level course) and $T_{ij} = 0$ otherwise. The outcome of interest Y_{ij} is students' weighted GPA in mathematics during 9-th grade, which is adjusted for the difficulty levels of math courses. For pre-treatment (insensitive) covariates, we use 11 student-level covariates, such as their 8-th math proficiency level, participation in math club activities, perceptions of math utility, and math identity, as well as 12 school-level covariates, such as student-to-math-teacher ratio, math courses requirements, and school type. We also consider two sensitive covariates: racial groups (white versus black) at the student level and proportions of black students (high versus low proportions) at the school level. For the school-level sensitive variable, we

use a threshold of 20% to determine schools with high versus low proportions of black students. A proportion above 20% indicates a notable shift in student composition (Bohrnstedt et al., 2015), and this threshold is the median in our empirical data. For desirable fairness functions, we apply the statistical disparity criterion to the school-level sensitive variable, whereas we use the conditional statistical disparity criterion for the individual-level, race variable where the legitimate/fair variable is whether a student's pre-high school math achievement is above the average level or not. We impute missing values in the covariates with predictive mean matching. For a complete list of variables used in this study, see Appendix S3.

For the HSLS:09 data analysis, we apply our proposed method with the chosen fairness criteria using two different values of δ: a large value of 20 for an unfair model and a (near-)zero value of 0.0001 for a fair model. For comparison, we also implement our proposed method with only an individual-level fairness constraint to examine the impact of ignoring cluster-level fairness constraints in this real-data application.

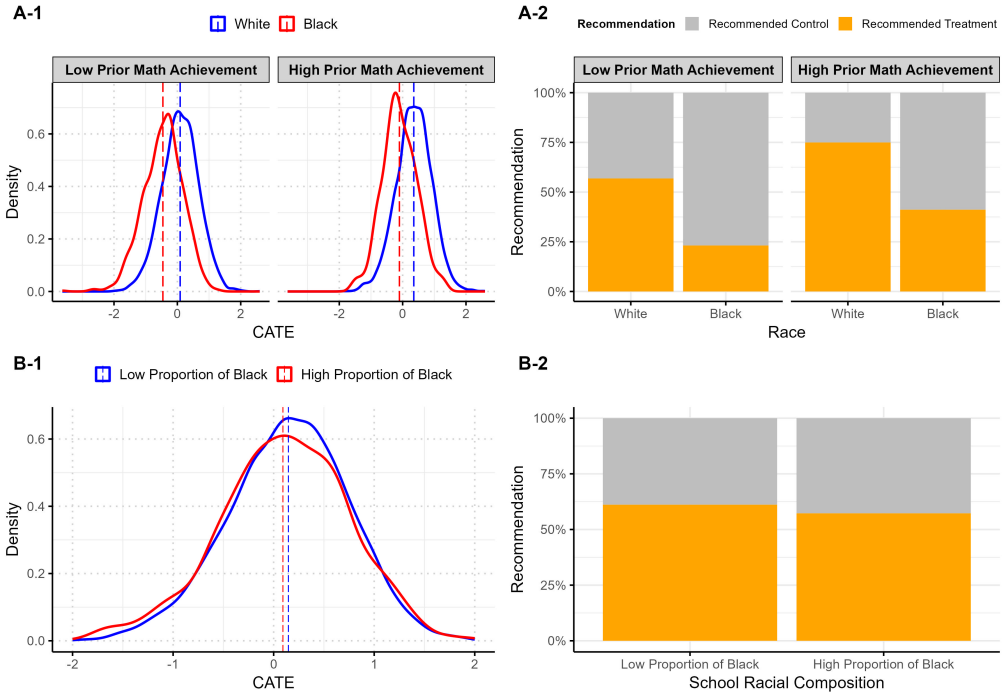
6.2 Results

Figure 3 provides the results about distributions of the CATE estimates and proportions of recommended treatments using our proposed method with multilevel fairness functions. We use two different values of δ and summarize each case in Figure 3-(1) and Figure 3-(2), respectively. The row categories in this figure represent sensitive variables, one at the student level and the other at the school level. Plot A-1 of Figure 3-(1) shows notable differences in CATE estimates between black and white students given their prior math achievement using the unfair model. These differences lead to disparities in course recommendations, as shown in Plot A-2 of Figure 3-(1); Black students are less likely to be recommended for advanced math courses compared to white students. For the school-level sensitive variable, we observe that students from schools with a higher proportion of black students have slightly lower CATE estimates compared to those from schools with a lower proportion, thereby leading to a slight difference in the recommendation rates for advanced math courses; see Plots B-1 and B-2 of Figure 3-(1). In contrast, when a zero value of δ is used in our proposed method with multilevel fairness functions, Figure 3-(2) shows no differences in the mean CATE estimates and proportions of recommended treatments for both student-level and school-level sensitive variables. This demonstrates that our proposed method ensures algorithmic fairness from both individual-level and cluster-level perspectives.

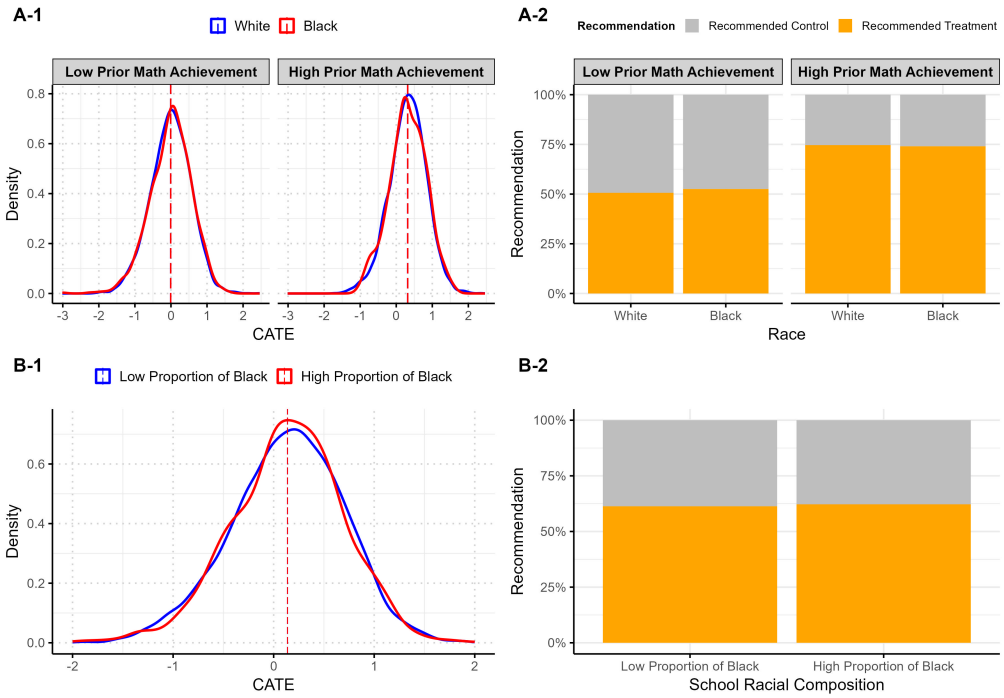
Additionally, we consider the case that includes only a student-level fairness function using a zero value of δ. Figure 4 summarizes the results. We maintain fair distributions in CATE estimates and recommended rates associated with the student-level sensitive variable, but unfairness with the school-level sensitive variable inadvertently increases when cluster-level fairness functions are ignored. Therefore, when developing fair algorithms in multilevel studies, it is essential to characterize multilevel fairness beyond the individual level as an algorithm considered fair at the individual level may not be fair from a cluster-level perspective.

7 Discussion and Conclusions

This paper provides a framework for estimating heterogeneous treatment effects under multilevel fairness considerations to develop fair and optimal policies in multilevel studies. Specifically, we propose a constrained optimization approach for CATE estimation with multiple fairness functions based on individual-level and cluster-level sensitive variables, as well as their intersectional groups. Our simulation study reveals that imposing only individual-level fairness constraints is insufficient to eliminate potential structural unfairness observed in multilevel studies. It is important to accurately characterize multilevel fairness functions by incorporating both individual-level and cluster-level sensitive variables and using appropriate fairness metrics. In



(1) Unfair model with a large value of δ



(2) Fair model with a zero value of δ

Figure 3: Distributions of the CATE estimates of taking the advanced math course and personalized recommendations in a student-level sensitive variable (A-1 and A-2) and a cluster-level sensitive variable (B-1 and B-2) using our proposed method with multilevel fairness functions: An unfair model (1) and a fair model (2).
SOURCE: Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLS:09).

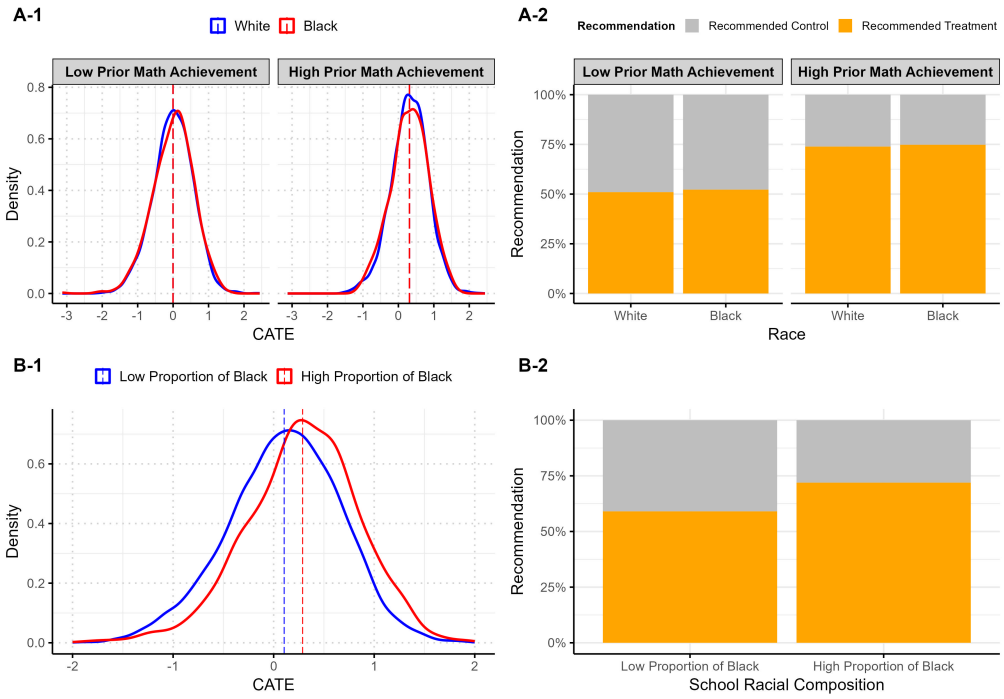


Figure 4: Distributions of the CATE estimates of taking the advanced math course and personalized recommendations in a student-level sensitive variable (A-1 and A-2) and a cluster-level sensitive variable (B-1 and B-2) using our proposed method with only an individual-level fairness function and a zero value of δ.
SOURCE: Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLS:09).

the simulations, we also find inevitable trade-offs between utility and fairness, where the best trade-off efficiency is achieved when multilevel fairness functions are correctly specified. Moreover, we demonstrate our proposed method using the HSLS:09 data, and our empirical results suggest that incorporating multilevel fairness considerations is necessary to ensure the fair distributions of CATE estimates and personalized recommendations in multilevel observational data.

Importantly, our proposed method requires users to choose a tolerance level, δ_k. As demonstrated in our simulation and real data studies, researchers may choose a tolerance level of zero if their goal is to achieve equality in fairness criteria. However, we empathize that the value of δ does not necessarily have to be zero; it can be set higher than zero to allow for some acceptable differences in certain or all fairness functions. Researchers should leverage domain knowledge about fairness in specific real-world applications (e.g., teaching, hiring, promotion, training) to determine appropriate tolerance levels. For example, they could use the 80% rule or statistical significance tests, which are commonly used for evaluating adverse impact in employment discrimination contexts (Biddle, 2006). Under the 80% rule, also known as the "four-fifths rule," researchers first investigate if the recommendation rate for a disadvantaged group is less than 80% of the rate for the advantaged group under no fairness constraints. If this is the case, they then gradually reduce the value of δ until the 80% rule is satisfied. With a statistical significance test, researchers can use non-zero values of δ such that differences in the mean CATE estimates or proportions of the recommended treatments that are statistically insignificant.

Based on the findings of this paper, we provide some suggestions for future research. First, the performance of the proposed estimator (14) can be significantly influenced by the choice of nuisance estimators for η. For simplicity, we used linear specification of GLMMs for nuisance

estimation, and did not incorporate any basis expansion in approximating τ more accurately. Future research would explore more flexible models for the nuisance functions and consider a much larger model space for projecting τ through basis expansions. Second, while our proposed method requires binary or categorical sensitive variables, it is limited in handling continuous sensitive variables unless they are categorized. Future research would investigate how to accommodate continuous sensitive variables in our constrained optimization method, potentially using Kolmogorov-Smirnov or Wasserstein distances. Third, when considering intersectionality across many sensitive variables, the number of constraints can become excessive. In such cases, making a prior distributional assumption on the differences among these intersectional groups could help reduce the number of parameters, similar to multilevel random-effects models. Future research could further explore intersectional fairness in our proposed framework. Despite these limitations, our proposed method complements the existing literature by introducing multilevel fairness considerations in estimating CATEs and developing optimal policies. We hope that our framework serves as a useful tool for more accurately defining fairness dimensions and promoting more equitable algorithmic decision-making in multilevel observational data.

References

Ballarini, N. M., Rosenkranz, G. K., Jaki, T., König, F., & Posch, M. (2018). Subgroup identification in clinical trials via the predicted individual treatment effect. *PloS one*, *13*(10), e0205971.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. http://www.fairmlbook.org

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Biddle, D. (2006). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge. https://doi.org/10.4324/9781315263298

Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., & Chan, D. (2015). *School composition and the black-white achievement gap. nces 2015-018* (tech. rep.). National Center for Education Statistics. U.S. Department of Education.

Boyd, R. W., Lindo, E. G., Weeks, L. D., & McLemore, M. R. (2020). On racism: A new standard for publishing on racial health inequities. *Health Affairs Forefront*.

Byun, S.-y., Irvin, M. J., & Bell, B. A. (2014). Advanced math course taking: Effects on math achievement and college enrollment. *The Journal of Experimental Education*, *83*(4), 439–468. https://doi.org/10.1080/00220973.2014.919570

Chan, E., Liu, Z., Qiu, R., Zhang, Y., Maciejewski, R., & Tong, H. (2024). Group fairness via group consensus. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3630106.3659006

Chiappa, S. (2019). Path-specific counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 7801–7808. https://doi.org/10.1609/aaai.v33i01.33017801

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 797–806. https://doi.org/10.1145/3097983.3098095

Coston, A., Mishler, A., Kennedy, E. H., & Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372851

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 139–167.

Dalton, B., Ingels, S. J., Downing, J., & Bozick, R. (2007). *Advanced mathematics and science coursetaking in the spring high school senior classes of 1982, 1992, and 2004* (tech. rep. NCES Publication No. 2007-312). National Center for Education Statistics. http://nces.ed.gov/pubs2007/2007312.pdf

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2783258.2783311

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1918–1921. https://doi.org/10.1109/icde48307.2020.00203

Glymour, C., & Glymour, M. R. (2014). Commentary: Race and sex are causes. *Epidemiology*, *25*(4), 488–490. https://doi.org/10.1097/ede.0000000000000122

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372826

Hansen, B. E., & Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of econometrics*, *210*(2), 268–290.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945. https://doi.org/10.2307/2289064

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. https://doi.org/10.1017/cbo9781139025751

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, K., & Zubizarreta, J. R. (2023). Fair and robust estimation of heterogeneous treatment effects for policy learning. *International Conference on Machine Learning*, 16997–17014.

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

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *2017 Conference on Neural Information Processing Systems (NIPS)*. https://doi.org/30

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Lipkovich, I., Dmitrienko, A., & B D'Agostino Sr, R. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine*, *36*(1), 136–196.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models* (2nd). Wiley.

Mhasawade, V., & Chunara, R. (2021). Causal multi-level fairness. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3461702.3462587

Mishler, A., & Kennedy, E. H. (2022). Fade: Fair double ensemble learning for observable and counterfactual outcomes. *2022 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3531146.3533167

Mishler, A., Kennedy, E. H., & Chouldechova, A. (2021). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445902

MOSEK ApS. (2019). *Rmosek: The r to mosek optimization interface* [R package version 1.3.5]. https://CRAN.R-project.org/package=Rmosek

Nabi, R., Malinsky, D., & Shpitser, I. (2019). Learning optimal fair policies. *Proceedings of the 36th International Conference on Machine Learning*, *32*(1), 4674–4682. https://doi.org/10.1609/aaai.v32i1.11553

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Section 9 (with discussion). *Statistical Science*, *4*, 465–480.

Park, C., & Kang, H. (2022). A more efficient, doubly robust, nonparametric estimator of treatment effects in multilevel studies.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1–44. https://doi.org/10.1145/3494672

Rateike, M., Valera, I., & Forré, P. (2024). Designing long-term group fair policies in dynamical systems. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3630106.3658538

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 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. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*(396), 961–962. https://doi.org/10.2307/2289065

Russell, M., & Kaplan, L. (2021). An intersectional approach to differential item functioning: Reflecting configurations of inequality. *Practical Assessment, Research, and Evaluation*, *26*(1), 21. https://doi.org/https://doi.org/10.7275/20614854

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2014). The role of advanced high school coursework in increasing stem career interest. *Science Educator*, *23*(1), 1–13.

Semenova, V., & Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. https://doi.org/10.48550/arXiv.1702.06240

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., & Boyd, S. (2020). Osqp: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, *12*(4), 637–672.

Suk, Y., & Han, K. T. (2023). A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning. *Journal of Educational and Behavioral Statistics*, *49*(2), 151–172. https://doi.org/10.3102/10769986231171711

Suk, Y., & Han, K. T. (2024). Evaluating intersectional fairness in algorithmic decision making using intersectional differential algorithmic functioning. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986241269820

Suk, Y., & Kang, H. (2022). Robust machine learning for treatment effects in multilevel observational studies under cluster-level unmeasured confounding. *Psychometrika*, *87*(1), 310–343. https://doi.org/10.1007/s11336-021-09805-x

Tsiatis, A. A., Davidian, M., Holloway, S. T., & Laber, E. B. (2019). *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429192692

VanderWeele, T. J., & Robinson, W. R. (2014). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, *25*(4), 473–484. https://doi.org/10.1097/ede.0000000000000105

Viviano, D., & Bradic, J. (2023). Fair policy targeting. *Journal of the American Statistical Association*, *119*(545), 730–743. https://doi.org/10.1080/01621459.2022.2142591

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

Yang, K., Loftus, J. R., & Stoyanovich, J. (2021). Causal intersectionality and fair ranking. *2021 Symposium on Foundations of Responsible Computing*. https://doi.org/10.4230/LIPIcs.FORC.2021.7

Supplementary Materials

S1 Data-Generating Processes

Design 1 introduces the unfairness associated with both individual-level and cluster-level sensitive variables with no intersectionality issue. Design 2 is based on Design 1 but adds intersectional unfairness among the subgroups defined by individual-level and cluster-level sensitive variables. The details of the data simulation processes for Design 1 and Design 2 are provided below.

S1.1 Design 1

1. For multilevel data, create $j = 1, \ldots, J$ clusters and $i = 1, \ldots, n_j$ individuals within cluster j.
2. For each individual $i = 1, \ldots, n_j$ in cluster j, generate four individual-level insensitive variables $X_{1ij}^{(1)}$, $X_{2ij}^{(1)}$, $X_{3ij}^{(1)}$, $X_{4ij}^{(1)}$ and an individual-level sensitive variable $S_{1ij}^{(1)}$:

$$S_{1ij}^{(1)} \sim Bernoulli(0.4), \quad X_{1ij}^{(1)} \sim \mathcal{N}(0,1), \quad X_{2ij}^{(1)} \sim \mathcal{N}(-2S_{1ij}^{(1)} + 1, 1),$$
$$X_{3ij}^{(1)} \sim \mathcal{N}(-0.5S_{1ij}^{(1)}, 1), \quad X_{4ij}^{(1)} = \mathbb{1}(X_{ij}^{*} > 0.1), \quad X_{ij}^{*} \sim \mathcal{N}(-S_{1ij}^{(1)} + 0.25, 1),$$

   and generate three cluster-level insensitive variables $X_{1j}^{(2)}$, $X_{2j}^{(2)}$, $X_{3j}^{(2)}$ and a cluster-level sensitive variable $S_{1j}^{(2)}$:

$$S_{1j}^{(2)} \sim Bernoulli(0.3), \quad X_{1j}^{(2)} \sim \mathcal{N}(0,1),$$
$$X_{2j}^{(2)} \sim \mathcal{N}(-S_{1j}^{(2)} + 0.5, 1), \quad X_{3j}^{(2)} \sim \mathcal{N}(-0.3S_{1j}^{(2)}, 1)$$

3. Generate individual treatment status T_{ij} from the following logistic regression with a random effect V_j:

$$\begin{aligned} logit(\pi(W_{ij})) = &-0.7 + 0.3 \cdot (X_{1ij}^{(1)} + X_{2ij}^{(1)} + X_{3ij}^{(1)} + X_{4ij}^{(1)} + S_{1ij}^{(1)}) \\ &+ 0.3 \cdot (X_{1j}^{(2)} + X_{2j}^{(2)} + X_{3j}^{(2)} - S_{1j}^{(2)}) + V_j, \quad \text{with } V_j \sim \mathcal{N}(0, 1.950). \end{aligned}$$

   Here, $\pi(W_{ij})$ represents an individual's propensity score, i.e., the probability of receiving the treatment given measured covariates.

4. Generate the true CATE, $\tau(W_{ij})$:

$$\begin{aligned} \tau(W_{ij}) = &\, 0.8 + 0.3X_{1ij}^{(1)} + 0.2X_{2ij}^{(1)} + 0.2X_{3ij}^{(1)} + 0.1X_{4ij}^{(1)} - 0.5S_{1ij}^{(1)} \\ &+ 0.5X_{1j}^{(2)} + 0.4X_{2j}^{(2)} + 0.2X_{3j}^{(2)} - 0.2S_{1j}^{(2)}. \end{aligned}$$

5. Generate the potential outcomes $Y_{ij}(1), Y_{ij}(0)$ and the observed outcome Y_{ij} from the following linear regression model with a random effect U_j:

$$\begin{aligned} Y_{ij}(t) = &\, 4 + 0.4 \cdot (X_{1ij}^{(1)} + X_{2ij}^{(1)} + X_{3ij}^{(1)} + X_{4ij}^{(1)} + S_{1ij}^{(1)} + X_{1j}^{(2)} - X_{2j}^{(2)} + X_{3j}^{(2)} - S_{1j}^{(2)}) \\ &+ t \cdot \tau(W_{ij}) + U_j + R_{ij}, \\ Y_{ij} = &\, T_{ij}Y_{ij}(1) + (1 - T_{ij})Y_{ij}(0), \qquad U_j \sim \mathcal{N}(0, 0.078),\ R_{ij} \sim \mathcal{N}(0, 0.665). \end{aligned}$$

   Here, R_{ij} represents the random error for an individual. In this data-generating process, we achieve the conditional intra-class correlation (ICC) of 0.105 for the outcome model and 0.372 for the treatment model, which aligns with those observed in the HSLS:09 data.

S1.2 Design 2

1. Design 2 builds on Design 1, and introduces unfairness by intersectional groups determined by individual-level sensitive variable $S_{1ij}^{(1)} \sim Bernoulli(0.4)$ and cluster-level sensitive variable $S_{1j}^{(2)} \sim Bernoulli(0.3)$. The intersectional subgroups are defined as $S_{ij}^{*} = (S_{1ij}^{(1)}, S_{1j}^{(2)}) \in \{(0,0),(1,0),(0,1),(1,1)\}$, where the first group is a reference group and $S_{ij,s}^{*}$ represents the group indicator, i.e., $S_{ij,s}^{*} = \mathbb{1}(S_{ij}^{*} = s)$.

2. Generate four individual-level insensitive variables $X_{1ij}^{(1)}$, $X_{2ij}^{(1)}$, $X_{3ij}^{(1)}$, and $X_{4ij}^{(1)}$ (that are correlated with sensitive variables):

$$X_{1ij}^{(1)} \sim \mathcal{N}(-S_{ij,(0,1)}^{*}, 1), \quad X_{1ij}^{(2)} \sim \mathcal{N}(-2S_{ij,(1,1)}^{*}, 1), \quad X_{1ij}^{(3)} \sim \mathcal{N}(S_{ij,(1,0)}^{*}, 1),$$
$$X_{4ij}^{(1)} = \mathbb{1}(X_{ij}^{*} > 0.1), \quad X_{ij}^{*} \sim \mathcal{N}(-S_{ij,(1,0)}^{*} + 0.25, 1),$$

   and generate three cluster-level insensitive variables $X_{1j}^{(2)}$, $X_{2j}^{(2)}$, and $X_{3j}^{(2)}$:

$$X_{1j}^{(2)} \sim \mathcal{N}(0,1), \quad X_{2j}^{(2)} \sim \mathcal{N}(0.5,1), \quad X_{3j}^{(2)} \sim \mathcal{N}(-1,1).$$

3. Generate individual treatment status T_{ij}. This model is the same as in Design 1.

4. Generate the true CATE, $\tau(W_{ij})$:

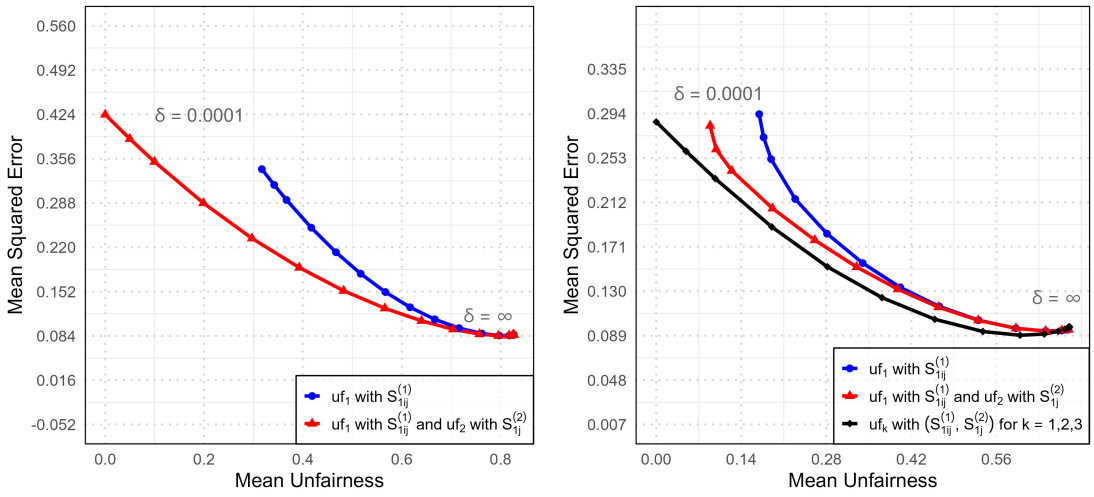
$$\begin{aligned}\tau(W_{ij}) = 0.8 &+ 0.3X_{1ij}^{(1)} + 0.2X_{2ij}^{(1)} + 0.2X_{3ij}^{(1)} + 0.1X_{4ij}^{(1)} - S_{ij,(1,0)}^{*} \\ &+ 0.5S_{ij,(0,1)}^{*} - 0.5S_{ij,(1,1)}^{*} + 0.5X_{1j}^{(2)} + 0.4X_{2j}^{(2)} + 0.2X_{3j}^{(2)}.\end{aligned}$$

5. Generate the potential outcomes $Y_{ij}(1), Y_{ij}(0)$ and the observed outcome Y_{ij} from the following linear regression model with a random effect U_j:

$$\begin{aligned}Y_{ij}(t) &= 4 + 0.4 \cdot (X_{1ij}^{(1)} + X_{2ij}^{(1)} + X_{3ij}^{(1)} + X_{4ij}^{(1)} + X_{1j}^{(2)} - X_{2j}^{(2)} + X_{3j}^{(2)}) - S_{ij,(1,0)}^{*} \\ &\quad + 0.5S_{ij,(0,1)}^{*} - 0.5S_{ij,(1,1)}^{*} + t \cdot \tau(W_{ij}) + U_j + R_{ij}, \\ Y_{ij} &= T_{ij}Y_{ij}(1) + (1 - T_{ij})Y_{ij}(0), \qquad U_j \sim \mathcal{N}(0, 0.078),\ R_{ij} \sim \mathcal{N}(0, 0.665).\end{aligned}$$

S2 Simulation Results: CATE

We summarize simulation results by investigating the trade-off between the mean squared error (MSE) and mean unfairness of the CATE estimates. Here, the mean unfairness is defined as the average difference in CATE estimates between subgroups of interest, unlike the unfairness of policy used in Section 5. From Figure S1, we find that as the value of δ moves from a large value ($\delta = \infty$) to $\delta = 0.0001$, the MSE increases but the unfairness effectively decreases. Reducing fairness comes at a slight cost of estimation accuracy. Additionally, in cases where the fairness functions are correctly specified, we observe a better trade-off between accuracy and fairness compared to using inaccurate fairness functions (e.g., only an individual-level function or no intersectional fairness functions). Overall, we need a multilevel approach to fairness considerations in order to achieve the high trade-off efficiency between utility and unfairness in multilevel studies.



(1) Design 1 (2) Design 2

Figure S1: Trade-off between accuracy and unfairness of the conditional average treatment effect estimates in Designs 1 and 2 with varying threshold values from $\delta = \infty$ (i.e., unfair model) to $\delta = 0.0001$ (i.e., fair model).

S3 Description of Variables

Table S1 provides a list of the variables used for this study from the HSLS:09 data.

Table S1: Variables used in this study from HSLS:09 data

Variable	Description	Level
X0TGPAMAT09	**Outcome**: weighted math GPA for 9th graders, constructed from transcript dataset.	Student
X0THIMATH09	**Treatment**: whether a 9th grader took an advanced math course (i.e., Geometry or higher level) or not; similar to X3THIMATH9, revised and reconstructed from transcript dataset.	Student
X1RACE	**Sensitive**: student's race, either Black or White.	Student
X1SCHBLACK	**Sensitive**: school's proportion of black students.	School
X0PREHSMAT	**Legitimate/fair**: student's pre-high school math achievement, based on S1M8 and S1M8GRADE.	Student
X1PAREDU	Parents'/guardians' highest level of education.	Student
X1SEX	Student's sex.	Student
X1SES	Student's socio-economic status.	Student
X1MTHUTI	Scale of student's mathematics utility.	Student
X1MTHID	Scale of student's mathematics identity.	Student
X0STAGE	Student's age in the year of 2009, based on X1STDOB.	Student
X0MATHACT	Participation in math activities since the 2008-09 school year, based on S1MCAMP, S1MCLUB, S1MCOMPETE, and S1MTUTOR.	Student
X1PAREDEXPCT	How far the parent thinks the student will go in school.	Student
X1STUEDEXPCT	How far the 9th grader thinks they will go in school.	Student
X0PARSTEM	Whether parent works in a STEM field, based on X1PAR1OCC_STEM1 and X1PAR2OCC_STEM1.	Student
X0MSTRATIO	Student-to-math-teacher ratio, based on A2HSSIZE and A1TOTMTCHRS.	School
A1MTHREQS	School requires the completion of specific math course(s) for graduation.	School
X1REGION	School's geographic region.	School
A1G9SUMMER	Offers pre-high school summer reading/math instruction for struggling 9th graders.	School
X0CSLHELP	High school counselors' assistance in transition from middle school to high school, based on C1TRANSCRS, C1TRANSCNSL, C1TRANPLCY, C1TRANPRNT, and C1TRANPRES.	School
C1G9MSAME	All 9th graders are placed in the same math course.	School
X0SCHHELP	School's assistance in transition from middle school to high school, based on C1TRANSTUDPR, C1TRANSTFFPR, C1TRANVISIT, C1TRANCLASS, C1TRANADMIN, C1TRANTCHRS, C1TRANBUDDY, C1TRANLRNCOM, C1TRANSUMMER, C1TRANFALL, and C1TRANSOTH.	School
C1G9MPLACTST	Importance of placement tests for 9th-grade math placement.	School

Continued on next page

Variable	Description	Level
X1LOCALE	School locale (urbanicity).	School
X1CONTROL	School control: public, private, or Catholic.	School
X0MATHRM	Whether the school provides a remedial math course, based on A1ONRMTH and A1OFFRMTH.	School
A1MTHSTREQ	Describe how required math course(s) for graduation compare with state's requirements.	School

NOTE: The transcript dataset is named stu_course.sas7bdat in the restricted version of HSLS:09 datasets. Variables beginning with X0 are composite variables constructed for this study, while all others were constructed by HSLS:09. We revised and reconstructed the math pipeline courses since the original coding strategy in the raw transcript dataset was inconsistent across the states.
SOURCE: Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLS:09).