Designing Optimal, Data-Driven Policies from Multisite Randomized Trials

Youmi Suk *[1] and Chan Park †[2]

[1]Department of Human Development, Teachers College Columbia University
[2]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania

September 18, 2023

Abstract

Optimal treatment regimes (OTRs) have been widely employed in computer science and personalized medicine to provide data-driven, optimal recommendations to individuals. However, previous research on OTRs has primarily focused on settings that are independent and identically distributed, with little attention given to the unique characteristics of educational settings, where students are nested within schools and there are hierarchical dependencies. The goal of this study is to propose a framework for designing OTRs from multisite randomized trials, a commonly used experimental design in education and psychology to evaluate educational programs. We investigate modifications to popular OTR methods, specifically Q-learning and weighting methods, in order to improve their performance in multisite randomized trials. A total of 12 modifications, 6 for Q-learning and 6 for weighting, are proposed by utilizing different multilevel models, moderators, and augmentations. Simulation studies reveal that all Q-learning modifications improve performance in multisite randomized trials and the modifications that incorporate random treatment effects show the most promise in handling cluster-level moderators. Among weighting methods, the modification that incorporates cluster dummies into moderator variables and augmentation terms performs best across simulation conditions. The proposed modifications are demonstrated through an application to estimate an OTR of conditional cash transfer programs using a multisite randomized trial in Colombia to maximize educational attainment.

Keywords: Optimal treatment regimes, Optimal treatment rules, Personalized learning, Q-learning, Weighting, Multilevel data, Clustered data, Conditional cash transfer

1 Introduction

Treatment effect heterogeneity has emerged as a critical concern in the social sciences when subpopulations exhibit differential gains from treatments. Treatment effects can differ from one subject to another and vary across subgroups. To address this heterogeneity from randomized control trials or observational studies, researchers in education and psychology have explored different approaches, including interactions, random effects, and finite mixture models (Feller &

*ysuk@tc.columbia.edu
†chanpk@wharton.upenn.edu

This article has been accepted for publication in *Psychometrika*, published by Springer.

Gelman, 2015; Raudenbush & Schwartz, 2020). More recently, machine learning methods have been increasingly employed to estimate heterogeneous treatment effects or *conditional average treatment effects* (CATEs) (e.g., Chernozhukov et al., 2018; Hill, 2011; Wager & Athey, 2018). While these various methods allow researchers to detect and estimate effect heterogeneity, there is limited research on how to utilize heterogeneous treatment effects to make informed decisions about *optimal* treatment allocation, especifically in the fields of education and psychology. The overarching goal of this paper is to propose a framework for designing optimal, data-driven educational policies by incorporating recent advances in personalized medicine and adapting them to multisite educational contexts.

Currently, data-driven, personalized recommendations have been popular in fields outside of education and psychology, notably in computer science, personalized medicine, public policy, and, most recently, criminal justice reform (e.g., Agniel et al., 2020; Murphy, Lynch, et al., 2007; Murphy, Oslin, et al., 2007; Nabi et al., 2019). In particular, in personalized medicine, there is now a widely accepted belief that every individual can benefit more from a data-driven, personalized treatment plan informed by vast amounts of historical patient data rather than a one-size-fits-all treatment plan for every patient (Tsiatis et al., 2019). A popular data-driven, recommendation model in personalized medicine is based on optimal treatment regimes (OTRs). OTRs use the observed effect heterogeneity (e.g., CATE estimates) under different treatment plans to find a decision rule that maximizes an outcome of interest (Chakraborty & Moodie, 2013; Murphy, 2003); see Section 3 for more details.

As a concrete example, suppose we are interested in developing an optimal regime for a conditional cash transfer (CCT) program to maximize students' attendance in school and we want to decide whether to recommend them for two students, say Emma and Peter, who come from economically disadvantaged backgrounds (Barrera-Osorio et al., 2011a). A decision rule can recommend both Emma and Peter to receive the CCT program, irrespective of their characteristics like age, grade, or family income, given the overall positive impact of the program, but the rule may not be *optimal*. In an optimal decision rule, Emma may be recommended to attend the CCT program because she is expected to show a strong positive effect given her characteristics, whereas Peter may be recommended not to receive the CCT program because of a potential negative effect. A critical feature of a personalized CCT program is that it considers each student's characteristics and finds an optimal decision rule that maximizes student attainment in school. But unfortunately, there is little research on translating OTRs to educational settings based on the potential effect heterogeneity and developing methodologies specifically tailored to the context of education and psychology.

In contrast, the field of personalized medicine has led to significant methodological advances in designing optimal recommendations. Common methods for OTRs include Q-learning (Q denoting "quality"; Murphy, 2005; Watkins & Dayan, 1992) and weighting methods (e.g., Chen et al., 2017; Qian & Murphy, 2011; Zhang et al., 2012; Zhao et al., 2012). Q-learning is based on outcome regression and involves two steps for constructing an optimal regime. In contrast, a weighting approach uses the treatment model and a part of outcome regression, while also directly searching for an OTR; see Section 3 for more information on Q-learning and weighting. These OTR methods have been designed for what we call single-level data, where study units are assumed to be independent and identically distributed (i.i.d.). When applied to clustered or multilevel data, there is no guarantee that they will produce consistent estimates for the OTR. Therefore, existing OTR methods may need to be modified to account for the underlying clustering or multilevel structures and to yield more precise and consistent estimates of the OTRs in multilevel studies. However, it remains an open question as to how to exactly make such modifications for different types of OTR methods.

The main goal of this paper is to investigate how to modify Q-learning and weighting methods to estimate an OTR robustly from multisite randomized trials that are frequently employed to evaluate programs or policies in education and psychology. Briefly, multisite randomized trials involve randomization of individuals into treatment and control groups within clusters. We define a total of 12 modifications—6 modifications for Q-learning and 6 modifications for weighting—by using different multilevel models, different sets of moderators (i.e., variables that interact with treatment), or different augmentations. Briefly, the proposed *multilevel OTR* methods based on Q-learning use different multilevel outcome models, such as fixed effects models, random effects models, or hybrid models. The proposed weighting methods for multilevel OTRs are fine-tuned by adding cluster dummies to moderator variables and/or augmentation terms; see Section 4 for details. We evaluate the performance of these modified OTR methods in comparison to their counterparts without modifications by measuring several performance criteria including accuracy and F1 score, which are used for binary classification. Finally, we demonstrate the modified OTR methods by designing an optimal regime for CCT programs using data from a multisite randomized trial in Colombia. We hope that multilevel OTR methods will aid in the design of more accurate, data-driven, optimal policies for individuals in education and the social sciences.

The remainder of the paper is organized as follows. Section 2 presents the setup, and Section 3 reviews the definition and estimation methods of OTR in i.i.d data settings. In Section 4, we discuss our proposed modifications for OTRs in multisite randomized studies. Section 5 outlines the design of our simulation study and presents its results. Section 6 demonstrates our proposals in empirical data about CCT programs. Finally, we provide discussion and conclusions in Section 7.

2 Setup

2.1 Notation, Decision Rule, and Potential Outcomes

Consider an i.i.d. sample of n study units, indexed by $i = 1, 2, \ldots, n$ from a randomized study. Let $T_i \in \mathcal{T} = \{0, 1\}$ denote the treatment where $T_i = 1$ indicates that individual i was treated and $T_i = 0$ indicates that individual i was untreated. Let Y_i denote the observed outcome for individual i, where larger values are assumed to be preferable without loss of generality. We denote $\mathbf{X}_i \in \mathcal{X}$ as finite-dimensional observed covariates available on individual i. To keep the notation simple, we define the first element of $\mathbf{X}_i$ to be 1 (denoted as $X_{0i} = 1$ for all i), so $\mathbf{X}_i$ can account for the intercept. A *decision rule* $d \in \mathcal{D}$ is a function that maps an individual's covariates $\mathbf{X}_i$ to a treatment option in $\mathcal{T}$ (Tsiatis et al., 2019), i.e., $d: \mathcal{X} \to \mathcal{T},\ \ \mathbf{X}_i \to d(\mathbf{X}_i)$. Here, $\mathcal{D}$ denotes the collection of all possible treatment rules/regimes d. As an example, consider a decision rule of the form $d(\mathbf{X}) = I(\mathsf{age} < 18 \text{ and } \mathsf{income} < 380000)$, where $I(\cdot)$ is the indicator function. Under this rule, an individual is given the treatment if their age is less than 18 years old and their income is less than 380,000 pesos, and otherwise, they are assigned to the control group. In practice, decision rules are often restricted by researchers' desire for them to be interpretable and easy to implement, while still being rich enough to encompass complex decision-making processes. In this paper, we denote a restricted subset of $\mathcal{D}$ as $\mathcal{D}_\beta$ where β is a finite-dimensional parameter that parameterizes the subset of decision rules considered by the investigator.

We use the potential outcomes framework to measure the quality of a decision rule. (Neyman, 1923; Rubin, 1974). Let $Y_i(t)$ be the potential outcome that would be achieved if individual i were to receive treatment option $t \in \mathcal{T}$. Specifically, $Y_i(1)$ denotes the potential treatment outcome if they were treated ($T_i = 1$) and $Y_i(0)$ denotes the potential control outcome if they were untreated ($T_i = 0$). In practice, only one of the potential outcomes is observed, $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$, where the relation assumes the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1986); see more explanation of SUTVA below. Also, under a given regime d, we define the potential outcome that an individual would achieve if the treatment were assigned according to regime d as: $Y_i(d) = d(\mathbf{X}_i)Y_i(1) + \{1 - d(\mathbf{X}_i)\}Y_i(0)$ (Tsiatis et al., 2019).

2.2 Treatment Effects and Benefit Scores

We define CATEs and *benefit scores* related to the decision rule. The CATE is the average linear contrast of potential outcomes between the treated and untreated groups among subgroups determined by $\mathbf{X}_i$, and is denoted as $\Delta(\mathbf{x})$ (Chen et al., 2017; Huling & Yu, 2021; Imbens & Rubin, 2015), i.e.,

$$\Delta(\mathbf{x}) = E\{Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}\}. \tag{1}$$

In the context of a CCT program, consider a scenario where $\mathbf{X}_i$ includes a student's sex X_{1i} (0=males, 1=females) and age X_{2i} measured in years as well as the intercept term $X_{0i} = 1$. Here, the CATE (i.e., $\Delta(\mathbf{x})$) represents the causal effect of the CCT program on school attendance for a specific subgroup of individuals, defined by the values of X_{1i} and X_{2i}. For example, $\Delta(1, 9)$ is the CATE of the CCT program on school attendance for females who are nine years old, while $\Delta(0, 10)$ is the CATE for males who are 10 years old. Using these CATEs, researchers can develop a decision rule, where the treatment is recommended if the CATE is positive, and otherwise, the control is recommended, i.e., $d(\mathbf{X}_i) = I\{\Delta(\mathbf{X}_i) > 0\}$.[1]

The observed effect heterogeneity can be represented by benefit scores, which are a monotonic transformation of $\Delta(\mathbf{X}_i)$ (Chen et al., 2017; Huling & Yu, 2021). A benefit score is defined to be any mapping $f(\mathbf{X}_i)$ that meets the following two properties: i) it is monotone in the treatment effect $\Delta(\mathbf{X}_i)$ and ii) it has a known cutpoint value c such that $f(\mathbf{X}_i) > c$ indicates that the treatment is more beneficial than the control. Examples of benefit scores include $f(\mathbf{X}_i) = \Delta(\mathbf{X}_i)/2$ and $f(\mathbf{X}_i) = \exp\{\Delta(\mathbf{X}_i)\}$. Continuing with the above example, one can design a decision rule based on benefit scores, e.g., $d(\mathbf{X}_i) = I\{f(\mathbf{X}_i) > 0\}$.

2.3 Causal Assumptions

The typical assumptions for identifying $\Delta(\mathbf{X}_i)$ or $f(\mathbf{X}_i)$ are the SUTVA, conditional ignorability, and positivity (Chakraborty & Moodie, 2013; Chen et al., 2017; Tsiatis et al., 2019). The SUTVA means that (i) individual i's potential outcomes are independent of others' treatment assignments and (ii) there is only one version of the treatment. In this paper, we assume SUTVA holds. Also, the conditional ignorability assumption i.e., $\{Y_i(1), Y_i(0)\} \perp T_i|\mathbf{X}_i$ is satisfied by design because this paper focuses on randomized trials and randomization ensures the treatment status T_i is independent of the potential outcomes $Y_i(1), Y_i(0)$ and covariates $\mathbf{X}_i$. Lastly, the positivity assumption states that the propensity score, which represents the probability of $T_i = 1$ given the covariates $\mathbf{X}_i$, falls between 0 and 1. Formally, it can be expressed as $0 < \pi(\mathbf{X}_i) := Pr(T_i = 1|\mathbf{X}_i) < 1$. In the context of randomized experiments, the positivity assumption is met by design because the propensity score is typically known and free of $\mathbf{X}_i$. For more details on the causal assumptions, see Chapter 2 of Tsiatis et al. (2019), Chapter 2 of Chakraborty and Moodie (2013), and Chen et al. (2017).

3 Review: Optimal Treatment Regimes

3.1 Definitions

An OTR aims to find the "best" decision rule $d^{opt} \in \mathcal{D}$ that maximizes the *value function* $\mathcal{V}(d)$, and formally, it is written as (Tsiatis et al., 2019):

$$d^{opt} = \underset{d \in \mathcal{D}}{\arg\max}\, \mathcal{V}(d). \tag{2}$$

[1] The CATE is similar to the *optimal blip to zero function* (Robins, 2004) comparing the expected outcomes between the control option and a particular treatment option of interest.

Typically, the standard value function is $\mathcal{V}(d) = E\{Y(d)\} = E[E\{Y_i(d(\mathbf{X}_i))|\mathbf{X}_i\}]$. In the observed data, an OTR d^{opt} is characterized formally as (Tsiatis et al., 2019):

$$d^{opt}(\mathbf{x}) = \arg\max_{t \in \mathcal{T}} E(Y_i \mid \mathbf{X}_i = \mathbf{x}, T_i = t), \tag{3}$$

when the aforementioned causal assumptions hold and the standard value function is used. That is, an optimal rule assigns treatments to individuals in a way that maximizes the average outcomes across the population (Chen et al., 2017; Tsiatis et al., 2019).

3.2 Q-learning

As mentioned above, Q-learning for estimating OTRs is an outcome regression approach. Let $Q(\mathbf{x}, t; \boldsymbol{\beta}) = E(Y_i \mid \mathbf{X}_i = \mathbf{x}, T_i = t; \boldsymbol{\beta})$ be an outcome regression model posited by a researcher with a finite-dimensional parameter $\boldsymbol{\beta}$. If the posited outcome regression value under treatment is larger than that under control, i.e., $Q(\mathbf{x}, 1; \boldsymbol{\beta}) > Q(\mathbf{x}, 0; \boldsymbol{\beta})$, it indicates that units with covariates $\mathbf{x}$ would have a larger outcome if treated. In other words, given $\boldsymbol{\beta}$, the optimal policy based on Q-learning is equivalent to the sign of the differences between two outcome regression values under treatment and control, i.e., $d(\mathbf{x}; \boldsymbol{\beta}) := I\{Q(\mathbf{x}, 1; \boldsymbol{\beta}) - Q(\mathbf{x}, 0; \boldsymbol{\beta}) > 0\}$. Therefore, to find the optimal treatment rule, it suffices to compare only decision rules $d(\mathbf{x}; \boldsymbol{\beta})$ over possible values of $\boldsymbol{\beta}$. We formally define this comparison as a decision space for Q-learning, i.e.,

$$\mathcal{D}_\beta = \big\{d(\mathbf{x}; \boldsymbol{\beta}) \big| d(\mathbf{x}; \boldsymbol{\beta}) = I\{Q(\mathbf{x}, 1; \boldsymbol{\beta}) - Q(\mathbf{x}, 0; \boldsymbol{\beta}) > 0\}\big\}. \tag{4}$$

Here, the parameter for Q function, $\boldsymbol{\beta}$, also parameterizes the decision rule, and this suggests that a simple substitution estimator for the decision rule can be constructed within the Q-learning framework.

To estimate an optimal rule based on Q-learning, researchers first need to estimate $\boldsymbol{\beta}$ by using appropriate parametric estimation techniques such as (quasi-)likelihood-based methods, M-estimation methods, or the generalized estimating equations (GEE) (e.g., Liang & Zeger, 1986; Stefanski & Boos, 2002; Tsiatis et al., 2019; van der Vaart, 2000). Then, they can construct a plug-in Q-learning estimator for the optimal decision rule using (4) as the basis. This Q-learning estimator, denoted as $\hat{d}_Q^{opt}$, is defined as:

$$\hat{d}_Q^{opt}(\mathbf{x}) = I\{Q(\mathbf{x}, 1; \hat{\boldsymbol{\beta}}) - Q(\mathbf{x}, 0; \hat{\boldsymbol{\beta}}) > 0\} = I\{\hat{\Delta}(\mathbf{x}) > 0\}. \tag{5}$$

In words, the rule $\hat{d}_Q^{opt}(\mathbf{x})$ is 1 if $Q(\mathbf{x}, 1; \hat{\boldsymbol{\beta}})$ is larger than $Q(\mathbf{x}, 0; \hat{\boldsymbol{\beta}})$ (i.e., $\hat{\Delta}(\mathbf{x}) > 0$); otherwise, $\hat{d}_Q^{opt}(\mathbf{x})$ is 0. Therefore, an alternative representation of $\hat{d}_Q^{opt}(\mathbf{x})$ is given by

$$\hat{d}_Q^{opt}(\mathbf{x}) = \arg\max_{t \in \mathcal{T}} Q(\mathbf{x}, t; \hat{\boldsymbol{\beta}}), \tag{6}$$

which means that using the rule (4) maximizes the expected value of the outcome regression.

To illustrate the procedure of Q-learning more concretely, consider a simple example where Y_i is continuous, T_i is binary, and there is only one measured covariate X_{1i}. A linear outcome model is $Q(\mathbf{X}_i, T_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 X_i + \beta_2 T_i + \beta_3 X_{1i} T_i$. We first estimate $\boldsymbol{\beta}$ using ordinary least squares (OLS) estimation, and we then construct an optimal decision rule based on the CATE estimates, $\hat{\beta}_2 + \hat{\beta}_3 X_{1i}$. A bit more formally, from (5), the optimal rule is $\hat{d}_Q^{opt}(\mathbf{X}_i) = I\{\hat{\Delta}(X_{1i}) > 0\}$ $= I(\hat{\beta}_2 + \hat{\beta}_3 X_{1i} > 0)$. This rule ensures that the highest expected outcome is achieved as discussed in (6). For more details on Q-learning methods, see Chapter 3 of Tsiatis et al. (2019) and Chapter 6 of Kosorok and Moodie (2015).

3.3 Weighting

Unlike Q-learning, a weighting method proposed by Chen et al. (2017) aims to estimate the treatment effect or its transformation only, using the propensity score, without having to fully specify an outcome regression. Let $\mathbf{Z}_i$ represent a set of moderators that interact with the treatment, where $\mathbf{Z}_i \subset \mathbf{X}_i$. Let $\widetilde{T}_i = 2T_i - 1$ and $\widetilde{\mathcal{T}} = \{-1, 1\}$, where 1 indicates treated units and -1 indicates untreated units. Note using $\widetilde{\mathcal{T}} = \{-1, 1\}$ provides some aesthetic advantages, such as working with only the sign of f and directly using $\widetilde{T}$ instead of $2T - 1$. (Kosorok & Moodie, 2015). An outcome model with $\widetilde{T}_i$ can be divided into two parts, $E(Y_i \mid \mathbf{X}_i, \widetilde{T}_i) = m(\mathbf{X}_i) + \widetilde{T}_i \Delta(\mathbf{Z}_i)$, where the main effect of covariates $m(\mathbf{X}_i) = 0.5\{E(Y_i|\widetilde{T}_i = 1, \mathbf{X}_i) + E(Y_i|\widetilde{T}_i = -1, \mathbf{X}_i)\}$ and the treatment effect $\Delta(\mathbf{Z}_i) = 0.5\{E(Y_i|\widetilde{T}_i = 1, \mathbf{Z}_i) - E(Y_i|\widetilde{T}_i = -1, \mathbf{Z}_i)\}$. A primary goal of the weighting method is to estimate $\Delta(\mathbf{Z}_i)$ (or its transformation $f(\mathbf{Z}_i)$) without estimating $m(\mathbf{X}_i)$. Formally, it minimizes the following objective function with respect to $f(\mathbf{Z}_i)$:

$$\hat{f}(\mathbf{z}) = \arg\min_{f} E\left[\frac{\mathbf{M}(Y_i,\ \widetilde{T}_i f(\mathbf{z}))}{\widetilde{T}_i \pi(\mathbf{x}) + (1 - \widetilde{T}_i)/2} \middle| \mathbf{X}_i = \mathbf{x} \right] \tag{7}$$

where $\pi(\mathbf{x})$ represents the propensity score. In observational studies, the propensity score needs to be estimated from the data, and in randomized trials, one can use the known propensity score or the estimated propensity score by computing the proportion of treated units. $\mathbf{M}(y, v)$ represents a convex function with respect to v for each y, e.g., the squared error loss, $\mathbf{M}(y, v) = (y - v)^2$. The choice of a loss function generally does not depend on the distributions of outcomes. For instance, when working with a continuous outcome, the squared error loss is a commonly used choice, but alternative loss functions such as the hinge loss can also be utilized. Furthermore, one must make modeling choices for the form of f. The form of f can be simple, such as linear forms, or it can be more flexible, such as regression trees and smoothing splines (Chen et al., 2017; Huling & Yu, 2021). Unlike Q-learning, estimator (7) directly maximizes a value function; see Chen et al. (2017) for more details on their internal value function and methods.

An optimal decision rule based on the estimated benefit score $\hat{f}(\mathbf{z})$ is formalized as:

$$\hat{d}_W^{opt}(\mathbf{z}) = \text{sign}\{\hat{f}(\mathbf{z})\} \quad \text{for} \quad \widetilde{\mathcal{T}} = \{-1, 1\}, \tag{8}$$

where the rule $\hat{d}_W^{opt}(\mathbf{z}) = 1$ if $\hat{f}(\mathbf{z})$ is larger than 0; otherwise, $\hat{d}_W^{opt}(\mathbf{z}) = 0$.

As a concrete example, suppose Y_i is continuous, T_i is binary, and there is one measured covariate X_{1i}. The true benefit score model $f(\mathbf{X}_i; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 X_{1i}$ and the known propensity score is 0.5 for every student. To design an optimal regime, we first choose a form of the benefit score model, and here, we assume a linear form: $f(\mathbf{X}_i; \boldsymbol{\alpha}) = \mathbf{X}_i^\intercal \boldsymbol{\alpha}$. We then estimate the benefit score f using (7) with $\pi(\mathbf{X}_i) = 0.5$ and a valid loss, say the squared error loss. After that, we use the estimated benefit score $\hat{f}$ in (8) to make optimal decisions. We remark that $\hat{\boldsymbol{\alpha}}$ obtained from weighting methods will not be exactly the same as $\hat{\boldsymbol{\beta}}$ used for the decision rule in Q-learning due to different underlying estimation procedures employed in each method.

3.4 Comparison Between Q-learning and Weighting

We end this section by comparing Q-learning and weighting methods. First, the vanilla Q-learning method can be understood as a two-step approach. In the first step, we obtain an estimate of the outcome regression Q that can be parametric, nonparametric, or semi-parametric. The aforementioned Q-learning is a special case where the working model is based on parametric models. In the second step, we determine an optimal decision rule $\hat{d}_Q^{opt}$ in a way that maximizes the estimated outcome regression obtained from the first step. This procedure does not directly search for an optimal rule over $\mathcal{D}_\beta$ (which is considered to be a collection of "good" decision rules), and thus, Q-learning is often called an *indirect method.* A problem of Q-learning is that it can be sensitive to misspecification of the outcome model. In the extreme case, Qian and Murphy (2011) revealed that although d_Q^{opt} belongs to $\mathcal{D}_\beta$, an optimal rule from Q-learning can be inconsistent due to a misspecified outcome model, e.g., fitting a linear model when the true outcome model is appropriate for logistic regression. It should also be noted that evaluating whether the outcome model is misspecified is not the primary goal for OTRs and it is often more challenging than finding an optimal rule.

The weighting method, on the other hand, is a type of *direct method*, and it searches for an OTR by directly maximizing the value function. Moreover, unlike Q-learning, it is not necessary to estimate the model for the main effects of covariates, i.e., $m(\mathbf{X}_i)$ in the weighting method because $m(\mathbf{X}_i)$ is a nuisance component (Zhao et al., 2012). Therefore, the weighting method is more robust to the potential misspecification of $m(\mathbf{X}_i)$. This is an important advantage because in practice there are often many covariates used in $m(\mathbf{X}_i)$, but far fewer treatment moderators for $\Delta(\mathbf{Z}_i)$. Also, while the weighting method does not require fitting a full outcome regression model, the function $m(\mathbf{X}_i)$ can be augmented into the weighting method to improve efficiency. Even if the augmented outcome model is misspecified, it does not affect the consistency of the estimated optimal rules (Chen et al., 2017); see details of the augmentation in Section 4.2. While both Q-learning and weighting methods are valid methods for estimating an OTR, how the current Q-learning and weighting methods work in data that have multilevel structures and what modifications are the most effective in such settings have not been well-explored so far.

4 Optimal Treatment Regimes with Multisite Randomized Studies

In this section, we consider a multisite randomized study with n individuals. We use the index $j = 1, \ldots, J$ to represent J clusters of units, and $i = 1, \ldots, n_j$ to represent individual units within cluster/site j. All variables now possess a subscript ij, indicating individual i in cluster/site j. For example, Y_{ij} represents the observed outcome for individual i in cluster j. We also assume the SUTVA, conditional ignorability, and positivity, discussed in Section 2.3, to identify CATEs or benefit scores. In multilevel settings, specifically, the first condition of the SUTVA implies the absence of interference both within clusters and between clusters.

A multisite or multilevel structure in observed data can have a significant impact on treatment effect estimation (Raudenbush & Schwartz, 2020) and more specifically, in this paper, the estimation of OTRs. One potential issue is the presence of clustering effects, where individuals within a cluster share similar characteristics, leading to a non-negligible intraclass correlation coefficient (ICC) in the outcome or propensity score model. This is commonly observed in education data, where students are nested within schools, and such school effects can affect OTR estimation through Q-learning or weighting methods. Another potential issue is that decision rules may be influenced by both individual-level and cluster-level covariates, and unmeasured cluster-level moderators may alter treatment decisions. In our data analysis of CCT programs, there were no school-level covariates available, and thus, if school-level covariates acted as moderators that made treatment effects heterogeneous, OTRs from default Q-learning and weighting methods would be inconsistent. Therefore, to address these issues and produce more robust estimates of OTRs using data from multisite randomized trials, we propose modifications using different multilevel models, different sets of moderators, or different specifications of augmentations. Specifically, we use a total of 12 modifications, including 6 for Q-learning and 6 for weighting.

4.1 Modifications for Q-learning

For Q-learning, we provide six modifications based on three different types of multilevel outcome models: the fixed effects model, random effects models, and (hybrid) models with de-meaned variables; see Algorithm 1 for a summary. Let $\mathbf{V}_{ij}$ denote a subset of $\mathbf{X}_{ij}$ that interacts with the treatment where the first element of $\mathbf{V}_{ij}$ is 1 to account for the main treatment effect. The fixed effects outcome model for Q-learning, denoted as Q-fe, is written as:

$$Q_{fe}(\mathbf{X}_{ij}, T_{ij}) = \mu_{0j} + \mathbf{X}_{ij}^{\intercal}\boldsymbol{\beta}_{\boldsymbol{x}} + T_{ij}\mathbf{V}_{ij}^{\intercal}\boldsymbol{\beta}_{\boldsymbol{v}}, \quad \widehat{\Delta}(\mathbf{V}_{ij}) = \mathbf{V}_{ij}^{\intercal}\widehat{\boldsymbol{\beta}}_{\boldsymbol{v}}, \tag{9}$$

where the term μ_{0j} is the cluster-level main effect term and absorbs the effects of both measured and unmeasured cluster-level covariates. Q-fe will remove the cluster-level impact on the outcome without requiring knowledge of cluster-level covariates and can be fitted by adding a $J-1$ cluster dummy matrix, $\mathbf{S}_j$, that indicates individual i's cluster membership in one of the $J-1$ clusters. This modification is motivated by the econometrics or causal inference literature (e.g., Suk & Kang, 2022a, 2022b; Wooldridge, 2010) where unobserved cluster-specific effects are often modeled as fixed effects.

Also, we propose two modifications for Q-learning based on random effects outcome models, and the models are written as:

$$Q_{ri}(\mathbf{X}_{ij}, T_{ij}) = \mu_{0j} + \mathbf{X}_{ij}^{\intercal}\boldsymbol{\beta}_{\boldsymbol{x}} + T_{ij}\mathbf{V}_{ij}^{\intercal}\boldsymbol{\beta}_{\boldsymbol{v}}, \quad \mu_{0j} \sim \mathcal{N}(0, \sigma_0^2), \quad \widehat{\Delta}(\mathbf{V}_{ij}) = \mathbf{V}_{ij}^{\intercal}\widehat{\boldsymbol{\beta}}_{\boldsymbol{v}}, \tag{10}$$

$$\begin{aligned} Q_{rs}(\mathbf{X}_{ij}, T_{ij}) &= \mu_{0j} + \mathbf{X}_{ij}^{\intercal}\boldsymbol{\beta}_{\boldsymbol{x}} + T_{ij}\mathbf{V}_{ij}^{\intercal}\boldsymbol{\beta}_{\boldsymbol{v}} + T_{ij}\mu_{1j}, \quad \boldsymbol{\mu}_j \sim \mathcal{N}(\mathbf{0}, \Sigma), \\ \widehat{\Delta}(\mathbf{V}_{ij}) &= \mathbf{V}_{ij}^{\intercal}\widehat{\boldsymbol{\beta}}_{\boldsymbol{v}} + \widehat{\mu}_{1j}. \end{aligned} \tag{11}$$

where the cluster-level main effect μ_{0j} requires two assumptions: (i) μ_{0j} is normally distributed with mean 0 and common variance σ_0^2 and (ii) μ_{0j} is independent of measured covariates. $\boldsymbol{\mu}_j = \{\mu_{0j}, \mu_{1j}\}^{\intercal}$ represents a 2×1 matrix that contains the cluster-level main effect μ_{0j} and the cluster-level treatment effect μ_{1j} where the respective means are zero and a 2×2 covariance matrix $\Sigma = \begin{bmatrix} \sigma_0^2 & \sigma_{0,1} \\ \sigma_{0,1} & \sigma_1^2 \end{bmatrix}$. Importantly, model (10) assumes the main treatment effect to be constant across clusters, while model (11) allows for random variation of the treatment effect across clusters. The motivation behind these modifications is that in the literature, cluster-specific effects are often modeled through cluster random effect terms (e.g., Lee et al., 2021; Raudenbush & Bryk, 2002; Suk et al., 2021). In the multilevel modeling literature, model (10), denoted as Q-ri, is often called the random intercept model, and model (11), denoted as Q-rs, is often called the random slope model (Raudenbush & Bryk, 2002). If the additional assumptions made in Q-ri and Q-rs hold, they can reduce the impact of (unmeasured) cluster-level covariates on the outcome, and in particular, Q-rs can capture a cluster-specific component in a decision rule.

Moreover, we provide three modifications based on de-meaned variables where $Y_{ij}^* = Y_{ij} - n_j^{-1}\sum_{i=1}^{n_j} Y_{ij}$, $T_{ij}^* = T_{ij} - n_j^{-1}\sum_{i=1}^{n_j} T_{ij}$, $\mathbf{X}_{ij}^* = \mathbf{X}_{ij} - n_j^{-1}\sum_{i=1}^{n_j}\mathbf{X}_{ij}$, and $(T_{ij}\mathbf{V})_{ij}^* = T_{ij}\mathbf{V}_{ij} - n_j^{-1}\sum_{i=1}^{n_j} T_{ij}\mathbf{V}_{ij}$. The proposed outcome models based on de-meaned variables are written as follows:

$$Q_{dm}(\mathbf{X}_{ij}, T_{ij}) = \mathbf{X}_{ij}^{*\intercal}\boldsymbol{\beta}_{\boldsymbol{x}} + (T_{ij}\mathbf{V}_{ij})^{*\intercal}\boldsymbol{\beta}_{\boldsymbol{v}}, \quad \widehat{\Delta}(\mathbf{V}_{ij}) = \mathbf{V}_{ij}^{\intercal}\widehat{\boldsymbol{\beta}}_{\boldsymbol{v}}, \tag{12}$$

$$Q_{dm\text{-}ri}(\mathbf{X}_{ij}, T_{ij}) = \mu_{0j} + \mathbf{X}_{ij}^{*\intercal}\boldsymbol{\beta}_{\boldsymbol{x}} + (T_{ij}\mathbf{V}_{ij})^{*\intercal}\boldsymbol{\beta}_{\boldsymbol{v}}, \quad \mu_{0j} \sim \mathcal{N}(0, \sigma_0^2), \quad \widehat{\Delta}(\mathbf{V}_{ij}) = \mathbf{V}_{ij}^{\intercal}\widehat{\boldsymbol{\beta}}_{\boldsymbol{v}}, \tag{13}$$

$$\begin{aligned} Q_{dm\text{-}rs}(\mathbf{X}_{ij}, T_{ij}) &= \mu_{0j} + \mathbf{X}_{ij}^{*\intercal}\boldsymbol{\beta}_{\boldsymbol{x}} + (T_{ij}\mathbf{V}_{ij})^{*\intercal}\boldsymbol{\beta}_{\boldsymbol{v}} + T_{ij}^{*}\mu_{1j}, \quad \boldsymbol{\mu}_j \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\mu}}), \\ \widehat{\Delta}(\mathbf{V}_{ij}) &= \mathbf{V}_{ij}^{\intercal}\widehat{\boldsymbol{\beta}}_{\boldsymbol{v}} + T_{ij}\widehat{\mu}_{1j}. \end{aligned} \tag{14}$$

Model (12) uses the de-meaned outcome Y_{ij}^*, de-meaned covariates, and de-meaned treatment to estimate regression coefficients in outcome regression. The motivation for the three modifications is rooted in the idea that by subtracting cluster means from the original variables, we can create variables that are locally orthogonal to cluster-specific components in the original variables (as demonstrated in previous research by Athey et al. (2019) and Suk and Kang (2022b)). This is important because when unmeasured cluster-level covariates are present, they remain in the subspace that has cluster-specific variations only. By utilizing de-meaned variables that are locally orthogonal to this subspace, we can produce more robust estimates of OTRs from multilevel data faced with unmeasured cluster-level covariates. When we estimate $\Delta(\mathbf{V}_{ij})$, we use the original moderator variables $\mathbf{V}_{ij}$ so that the scale of $\Delta(\mathbf{V}_{ij})$ is not changed. Note that under the identity link, the estimates of $\boldsymbol{\beta_v}$ are identical between Q-fe and Q-dm, thus yielding the same decision rule and the same benefit scores. In contrast, Q-dm-ri and Q-dm-rs use the original outcome (i.e., Y_{ij}), de-meaned covariates, de-meaned treatment, and cluster-level random effect terms. The two models are often referred to as "hybrid models" (Firebaugh et al., 2013; Raudenbush, 2009) because they encompass the features of fixed effects models and random effects models. Hybrid models offer the fixed effects advantage of eliminating the effects of unmeasured cluster-level covariates. At the same time, like random effects models, they allow for estimating the impact of individual-level and cluster-level predictors, as well as estimating the random coefficients. Specifically, the estimates from Q-dm-ri are similar to those from Q-fe and Q-dm, but the hybrid models are robust to the violation of the underlying independence assumption made in random effects models.

Algorithm 1 summarizes the steps of Q-learning methods using data from multisite randomized trials. The cutoff value c is set to 0 by default. For implementation, we provide a function named *Qlearn* to run six outcome regression estimators in Algorithm 1 and the baseline outcome model with the OLS estimation. R codes for our modifications are available in the supplementary materials and can also be found at the first author's GitHub repository (https://github.com/youmisuk/multisiteOTR).

Algorithm 1 Q-learning methods with multisite randomized studies

```
Input: Outcome $Y_{ij}$, Treatment $T_{ij}$, covariates $\mathbf{X}_{ij}$, moderators $\mathbf{V}_{ij}$, cluster dummies $\mathbf{S}_j$
Input: Cutoff value $c = 0$
Input: Modification of Q-learning methods in {Q-fe, Q-ri, Q-rs, Q-dm, Q-dm-ri, Q-dm-rs}
  1: Estimate the chosen outcome regression $Q(\mathbf{x}, t)$ with a decision rule parametrized by
     regression coefficients $\boldsymbol{\beta}$.
  2: Predict the outcome for the control $Q(\mathbf{x}, 0; \hat{\boldsymbol{\beta}})$ and that for the treatment $Q(\mathbf{x}, 1; \hat{\boldsymbol{\beta}})$
  3: Compute $\hat{\Delta}(\mathbf{x}) = Q(\mathbf{x}, 1; \hat{\boldsymbol{\beta}}) - Q(\mathbf{x}, 0; \hat{\boldsymbol{\beta}})$
  4: Make treatment decisions based on $\hat{\Delta}(\mathbf{x})$ and the cutoff: $\hat{d}^{opt}(\mathbf{x}) = I\{\hat{\Delta}(\mathbf{x}) > 0\}$
Output: $\hat{\Delta}(\mathbf{x})$ and $\hat{d}^{opt}(\mathbf{x})$
```

4.2 Modifications for Weighting

We provide six modifications for the weighting approach to robustly estimate an OTR with multisite randomized studies. For interpretable models, we assume that f is linear with respect to a finite-dimensional parameter $\boldsymbol{\alpha}$. The first modification, denoted as W-noaug, is written as

$$\hat{f}_W(\mathbf{v}) = f(\mathbf{v}; \hat{\boldsymbol{\alpha}}) \ , \ \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\arg\min}\, E\left[\frac{\mathbf{M}(Y_{ij}, \widetilde{T}_{ij} f(\mathbf{v}; \boldsymbol{\alpha}))}{\widetilde{T}_{ij} p_j + (1 - \widetilde{T}_{ij})/2} \middle| \mathbf{V}_{ij} = \mathbf{v}\right], \quad \mathbf{V}_{ij} \subset \mathbf{X}_{ij}. \tag{15}$$

where $f(\mathbf{v}; \boldsymbol{\alpha})$ indicates that the benefit score is a function of moderators $\mathbf{V}_{ij} \subset \mathbf{X}_{ij}$. Treatment prevalence, $p_j = n_j^{-1} \sum_1^{n_j} I(T_{ij} = 1)$, is used as the propensity score. This modification allows us to account for potentially different propensity scores between clusters in multisite randomized trials. Note that researchers could additionally include measured covariates in the propensity score model to remove finite-sample differences between the treatment and control groups.

We also study two other variations with augmentation to improve the performance of W-noaug by employing the augmented loss function $\widetilde{\mathbf{M}}(y, v) = \mathbf{M}(y, v) + \mathbf{g}(\hat{m}(\mathbf{X}), v)$ instead of

$\mathbf{M}(y, v)$ in (15); the second term in the augmented loss function represents a loss for the main effects of covariates (Chen et al., 2017; Huling & Yu, 2021).[2] The basic idea of augmentation is to fit an outcome model for the main effects of covariates and shift the outcome based on the estimated main effects. The motivation behind modifications with augmentation is that the augmented loss function $\widetilde{\mathbf{M}}(y, v)$ does not change the optimality of the resulting decision rule and it potentially provides efficiency gains (Chen et al., 2017). This is because using the weighting estimator with the shifted outcome can be more efficient than using the original outcome (Huling & Yu, 2021).

Specifically, two augmentation models we use are: (i) augmentation with linear main effect terms of covariates $m(\mathbf{X}_{ij}) = \mathbf{X}_{ij}^{\intercal}\boldsymbol{\gamma}_{\boldsymbol{x}}$, denoted as W-aug, and (ii) augmentation with linear main effect terms of covariates and cluster dummies $m(\mathbf{X}_{ij}, \mathbf{S}_j) = \mathbf{X}_{ij}^{\intercal}\boldsymbol{\gamma}_{\boldsymbol{x}} + \mathbf{S}_j^{\intercal}\boldsymbol{\gamma}_{\boldsymbol{s}}$, denoted as W-augID. We use W-aug to check if a simple form of augmentation can improve the performance of W-noaug in our setting, and we study W-augID to examine if there is any additional gain from including cluster dummies in the augmentation. When an augmentation is added, we use the augmented loss function $\widetilde{\mathbf{M}}(y, v)$ in (15) instead of the original loss function to estimate benefit score f_W.

Moreover, among other three modifications, we extend the space of the moderator by adding cluster dummies $\mathbf{S}_j$, i.e., $\mathbf{V}_{ij} \subset \{\mathbf{X}_{ij}, \mathbf{S}_j\}$. denoted as W$_\mathsf{S}$-noaug, is written as:

$$\hat{f}_{W_S}(\mathbf{v}) = f(\mathbf{v}; \hat{\boldsymbol{\alpha}}) \ , \ \hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} E\left[\frac{\mathbf{M}(Y_{ij}, \widetilde{T}_{ij} f(\mathbf{v}; \boldsymbol{\alpha}))}{\widetilde{T}_{ij} p_j + (1 - \widetilde{T}_{ij})/2} \middle| \mathbf{V}_{ij} = \mathbf{v} \right], \quad \mathbf{V}_{ij} \subset \{\mathbf{X}_{ij}, \mathbf{S}_j\}. \qquad (16)$$

We add this modification because using an additional set of cluster dummies in moderator variables can help detect a decision rule driven by clusters, in particular, if there are no cluster-level covariates available. We also provide two more modifications with augmentation to potentially improve the performance of W$_\mathsf{S}$-noaug with respect to efficiency. Specifically, we use one modification augmented with linear main effect terms of covariates, denoted as W$_\mathsf{S}$-aug, and another modification augmented with linear main effect terms of covariates and cluster dummies, denoted as W$_\mathsf{S}$-augID.

Algorithm 2 summarizes the steps of W$_S$-augID among different weighting modifications. We set the cutoff value c to 0 by default. To implement weighting methods, we tweak functions named *fit.subgroup* and *propensity.func* from R package *personalized* (Huling & Yu, 2021). Regarding the loss function, we use the squared error loss with lasso penalty (Tibshirani, 1996), and we also allow the lasso penalty terms to be only applied to certain sets of covariates, say cluster dummies. Different loss functions can be used like the hinge loss or logistic loss functions, depending on outcome types.

5 Simulation Study

We conduct a large-scale simulation study to assess which of the proposed modifications improved the performance under which conditions. Throughout the simulations, we estimate the benefit scores and the optimal decisions using (1) our proposed modifications in Section 4, (2) the baseline outcome regression based on the OLS (denoted as Q-base), and (3) the weighting estimator that uses the constant propensity score (denoted as W-base). Specifically, our simulation study is divided into four designs. Design 1 uses linear main effects of $\mathbf{X}$ in the outcome model and assumes there is no unmeasured cluster-level moderator. Design 2 extends Design 1 by allowing an unmeasured cluster-level moderator and is intended to investigate which modifications are more robust in the presence of the unmeasured cluster-level moderator. Design 3

[2] In Section 3.3, we denote $E[Y|\mathbf{X}, \widetilde{T}] = m(\mathbf{X}) + \widetilde{T}_{ij}\Delta(\mathbf{Z})$ and $m(\mathbf{X}) = 0.5\{E(Y|\widetilde{T} = 1, \mathbf{X}) + E(Y|\widetilde{T} = -1, \mathbf{X})\}$. Note $\mathbf{g}(y, v)$ meets the same conditions required for $\mathbf{M}(y, v)$ (Chen et al., 2017; Huling & Yu, 2021).

```
Algorithm 2 Modification, W_S-augID
Input: Outcome $Y_{ij}$, treatment $T_{ij}$, covariates $\mathbf{X}_{ij}$, cluster dummies $\mathbf{S}_j$
Input: Treatment prevalence $p_j$
Input: Cutoff value $c = 0$
  1: Estimate $f$ by minimizing the following objective function with $\widetilde{T}_{ij} = 2T_{ij} - 1$.
       $\hat{f}_{W_S}(\mathbf{v}) = f(\mathbf{v}; \hat{\boldsymbol{\alpha}})\ ,\ \hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} E\left[\frac{\widetilde{\mathbf{M}}(Y_{ij}, \widetilde{T}_{ij} f(\mathbf{v}; \boldsymbol{\alpha}))}{\widetilde{T}_{ij} p_j + (1 - \widetilde{T}_{ij})/2} | \mathbf{V}_{ij} = \mathbf{v}\right], \quad \mathbf{V}_{ij} \subset \{\mathbf{X}_{ij}, \mathbf{S}_j\},$
     where the augmented loss function
       $\widetilde{\mathbf{M}}(Y_{ij}, \widetilde{T}_{ij} f(\mathbf{v}; \boldsymbol{\alpha})) = \mathbf{M}(Y_{ij}, \widetilde{T}_{ij} f(\mathbf{v}; \boldsymbol{\alpha})) + \mathbf{g}(\hat{m}(\mathbf{X}_{ij}, \mathbf{S}_j), \widetilde{T}_{ij} f(\mathbf{v}; \boldsymbol{\alpha})).$
  2: Make treatment decisions based on $\hat{f}_{W_S}(\mathbf{v})$ and the cutoff: $\hat{d}^{opt}(\mathbf{v}) = I\{\hat{f}_{W_S}(\mathbf{v}) > 0\}$
Output: $\hat{f}_{W_S}(\mathbf{v})$ and $\hat{d}^{opt}(\mathbf{v})$.
```

uses nonlinear main effects of $\mathbf{X}$ in the absence of an unmeasured cluster-level moderator, but if applicable, we still fit linear main effect terms in the estimators. We include this design to investigate the impact of model misspecification on the performance of Q-learning and weighting methods. In Appendix A, we conducted additional simulations where the treatment prevalence has more variation compared to Design 1. Specifically, the ICC estimate in the treatment model in Design 1 was 16.3%, whereas the ICC in Design 4 was 40.6%. The results of this additional design are similar to those from Design 1, and as expected, injecting cluster-level propensity scores inside the weighting estimator provides more accuracy gains as the ICC increases.

For each of the four designs, we vary the sample size where the number of clusters J and the cluster size n_j are either 25 or 150. That is, we use four sample size conditions (J, n_j): (25, 25), (150, 25), (25, 150), and (150, 150). Note that the sample size of (25, 150) is comparable to the size of our empirical data. Combining all of the variations across the three designs (plus the additional simulations in Appendix A), our simulation study considers 16 different data-generating models, and we examine the performance of the proposed methods in each data-generating model.

In each replicate, we evaluate the performance of the methods by investigating (i) accuracy i.e., the proportion of correctly assigned optimal decisions, (ii) the F1 score (=2 (Precision × Recall)/(Precision + Recall))[3], (iii) the rank correlation between the true benefit score $f(\mathbf{v})$ with the estimate $\hat{f}(\mathbf{v})$, and (iv) their area under the receiver operating characteristic curves (AUC) with respect to the true optimal decisions. Note that accuracy, F1-score and AUC are commonly used performance criteria for binary classification, where the original labels (also known as ground truth labels) are compared with the predicted labels produced by a machine learning model. A higher value of these measures indicates better performance. As a comparison criterion, we also measure the congruence among the estimators, i.e., the proportion of being in agreement with respect to the estimated optimal decisions. We repeat our simulation $r = 1, \dots, 500$ times on independent test datasets.

5.1 Design 1: Linear Main Effects and No Cluster-level Moderator

The data-generating model is based on those from Huling and Yu (2021) and our empirical data, and it is stated below.

[3] Precision $= \frac{\text{True Positive}}{\text{True Positive + False Positive}}$; $Recall = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$

1. For multisite/multilevel structures, create $j = 1, \ldots, J$ clusters, each with n_j individuals per cluster.
2. For each individual $i = 1, \ldots, n_j$ in cluster j, generate five individual-level covariates, $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij}, X_{3ij}, X_{4ij}, X_{5ij})$, and one cluster-level covariate U_j that is assumed to be unmeasured. All the covariates follow a normal distribution with a mean of 0 and variance of 1, i.e., $\mathcal{N}(0, 1)$.
3. Generate individual treatment status T_{ij} with cluster-specific propensity scores in the following logistic regression.

$$logit(e_{ij}) = 0.8 \cdot C_j, \quad C_j \sim \mathcal{N}(0,1), \qquad T_{ij} \sim Bernoulli(e_{ij})$$

   Here, e_{ij} is the propensity score for individual i in cluster j.
4. Generate the true treatment effects $\Delta(\mathbf{V}_{ij})$.

$$\Delta(\mathbf{V}_{ij}) = 0.2 + 0.5X_{3ij} - 0.2X_{4ij} - 0.2X_{5ij} + \beta_1 U_j$$

   Here, β_1 represents the coefficient of the cluster-level moderator and is set to 0 in Design 1.
5. Generate the potential outcomes $Y_{ij}(1), Y_{ij}(0)$ and the observed outcome Y_{ij} from the following linear regression model:

$$\begin{aligned} Y_{ij}(t) &= 0.7 + X_{1ij} + X_{2ij} - 2X_{3ij} + X_{4ij} + 0.5X_{5ij} + t \cdot \Delta(\mathbf{V}_{ij}) + 2.5U_j + \epsilon_{ij}, \\ Y_{ij} &= T_{ij}Y_{ij}(1) + (1 - T_{ij})Y_{ij}(0), \qquad \epsilon_{ij} \sim N(0,1) \end{aligned}$$

   Here, ϵ_{ij} represents the random error for individual i in cluster j.
6. Generate the true optimal decisions: $d^{opt} = I\{\Delta(\mathbf{V}_{ij}) > 0\}$.

Figure 1 summarizes the accuracy performance of the estimators with different sample size conditions in Design 1, placing the mean accuracy at the top of each boxplot. Each row in the figure represents the sample size, denoted by (J, n_j), with the number of clusters J and the size of each cluster n_j. For example, the first row category is a condition with $J = 25$ clusters and each cluster has 25 individuals. Note that we use different ranges of the y-axis to permit a clearer comparison of the estimators within each sample size condition in Figure 1.

Among Q-learning methods, the baseline estimator based on OLS regression (Q-base) shows the worst performance compared to the proposed modifications across different sample sizes as expected. Six modifications outperform the baseline estimator and increase the accuracy rates by about 6-15%. The pattern of results suggests that if hierarchical dependencies are present, it is necessary to account for clustering or hierarchical structures in the outcome model by using multilevel models (e.g., random effects, fixed effects, or hybrid models) to improve the accuracy in finite samples. When we compare Q-learning modifications, all the modifications show similar rates of accuracy in general, but those that allow for random treatment effects (Q-rs and Q-dm-rs) exhibit a slight loss of accuracy (92.4% to 90.8% or 90.7%) under the small sample size condition of (25, 25). This may be due to the fact that random slopes introduce unnecessary complexities in the outcome model under Design 1 and using the small sample size poses additional difficulties in estimating them reliably and accurately. However, when either the number of clusters or the cluster size increases, Q-rs and Q-dm-rs show almost the same accuracy rates as other modifications.

Among weighting methods, not surprisingly, the baseline weighting estimator that uses the constant propensity score without augmentation (i.e., W-base) shows the lowest accuracy rates. Injecting cluster-specific propensity scores inside the weighting estimator (W-noaug) improves

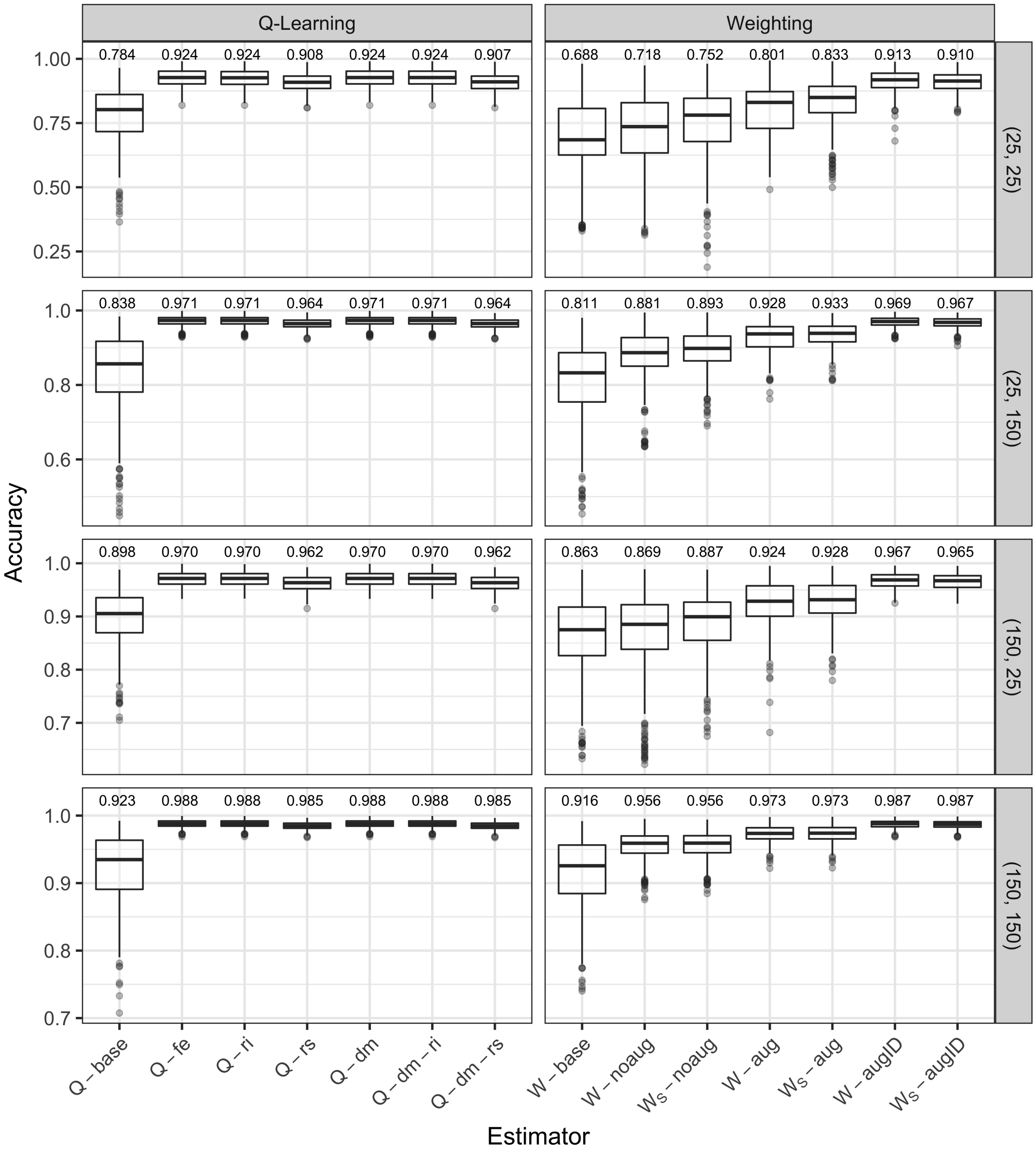

Figure 1: Accuracy of the estimators in Design 1.

the mean accuracy across sample size conditions. It also reduces the variability of the accuracy estimates, in particular, under a large cluster size of 150. In addition, including cluster dummies in the moderator variables (W$_{\text{S}}$-noaug) generally does not provide additional accuracy gains. This is expected given that there is no cluster-level moderator that alters treatment decisions in Design 1. Moreover, adding augmentation terms into W-noaug and W$_{\text{S}}$-noaug increases the accuracy rates and reduces the variability of accuracy estimates. Between the two types of augmentation terms, those with both covariates and cluster dummies (W-augID and W$_{\text{S}}$-augID) provide more accuracy gains than those with covariates only (W-aug and W$_{\text{S}}$-aug) and show similar accuracy rates to those from the modifications for Q-learning. This implies that when there are non-negligible clustering effects in the outcome model, accounting for them via augmentation is essential in raising the accuracy of the weighting estimator. For other performance criteria (i.e., F1 score, correlation, and AUC), we summarize results in Appendix A, and we find that the result patterns are generally similar to those of accuracy.

Figure 2 summarizes the congruence among the estimators in Design 1 when the sample size condition is equal to 25 clusters with a size of 150. Here, we focus on the sample size condition that is the most comparable to our empirical data because the congruence results are relatively similar across different sample size conditions. In Figure 2, the upper triangular part of the congruence matrix displays the heatmap and the lower triangular part reports the numerical results. To better visualize the congruence results, we list the weighting methods in reverse order. From Figure 2, we observe that there are high congruence rates of 98% or larger (highlighted in red) among Q-learning methods, except for Q-base; Q-base shows congruence rates that range from 83% to 92%.

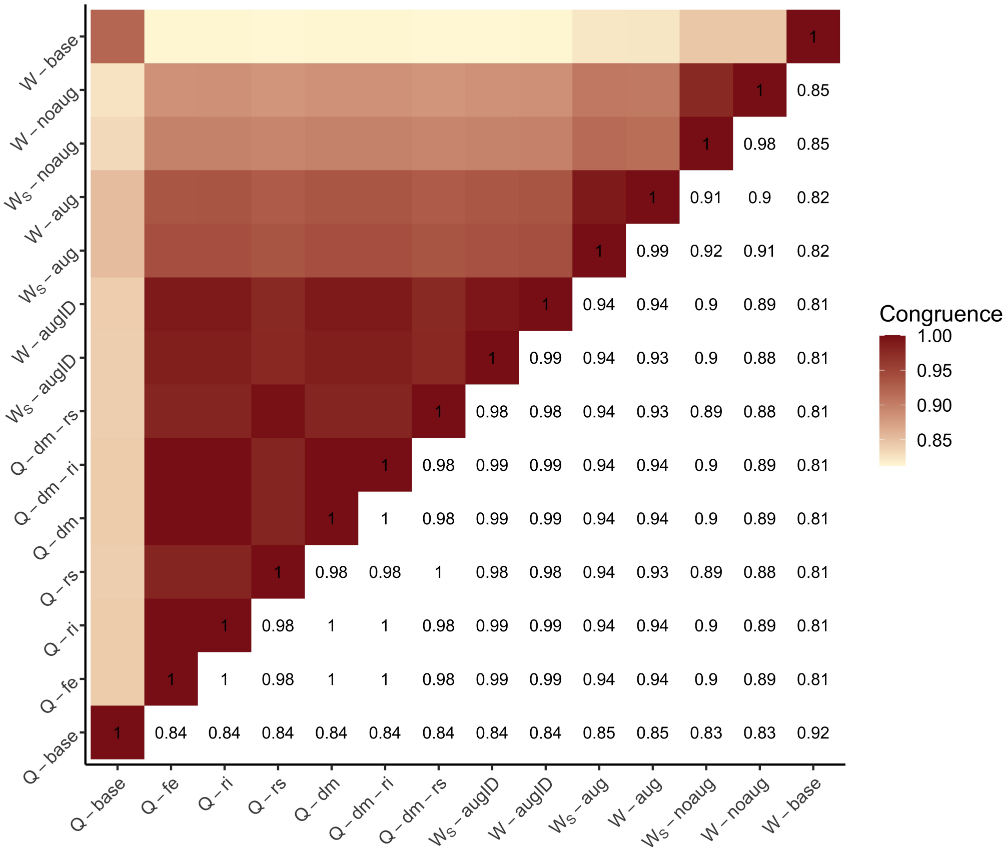

Figure 2: Congruence among the estimators in Design 1 when the number of clusters is 25 and the cluster size is 150.

As for weighting methods, six modifications show higher congruence rates than W-base that uses the constant propensity score, and in particular, augmentation types matter much. Those without augmentation (W-noaug, $\mathsf{W_S}$-noaug) exhibit at most 92% rates with other methods, i.e., those with augmentation and Q-learning methods. In contrast, using augmentation raises the congruence rates. Specifically, weighting methods augmented with covariates only (W-aug and $\mathsf{W_S}$-aug) show congruence rates of about 93-94% with Q-learning modifications, whereas weighting methods augmented with both covariates and cluster dummies (W-augID and $\mathsf{W_S}$-augID) increase congruence rates up to about 99%.

5.2 Design 2: Linear Main Effects and a Cluster-level Moderator

Design 2 used the same data-generating model as Design 1 except that the coefficient β_1 is now set to 0.2 to investigate the potential impact of an unmeasured cluster-level moderator on the performance of the methods. Figure 3 summarizes the simulation results under Design 2

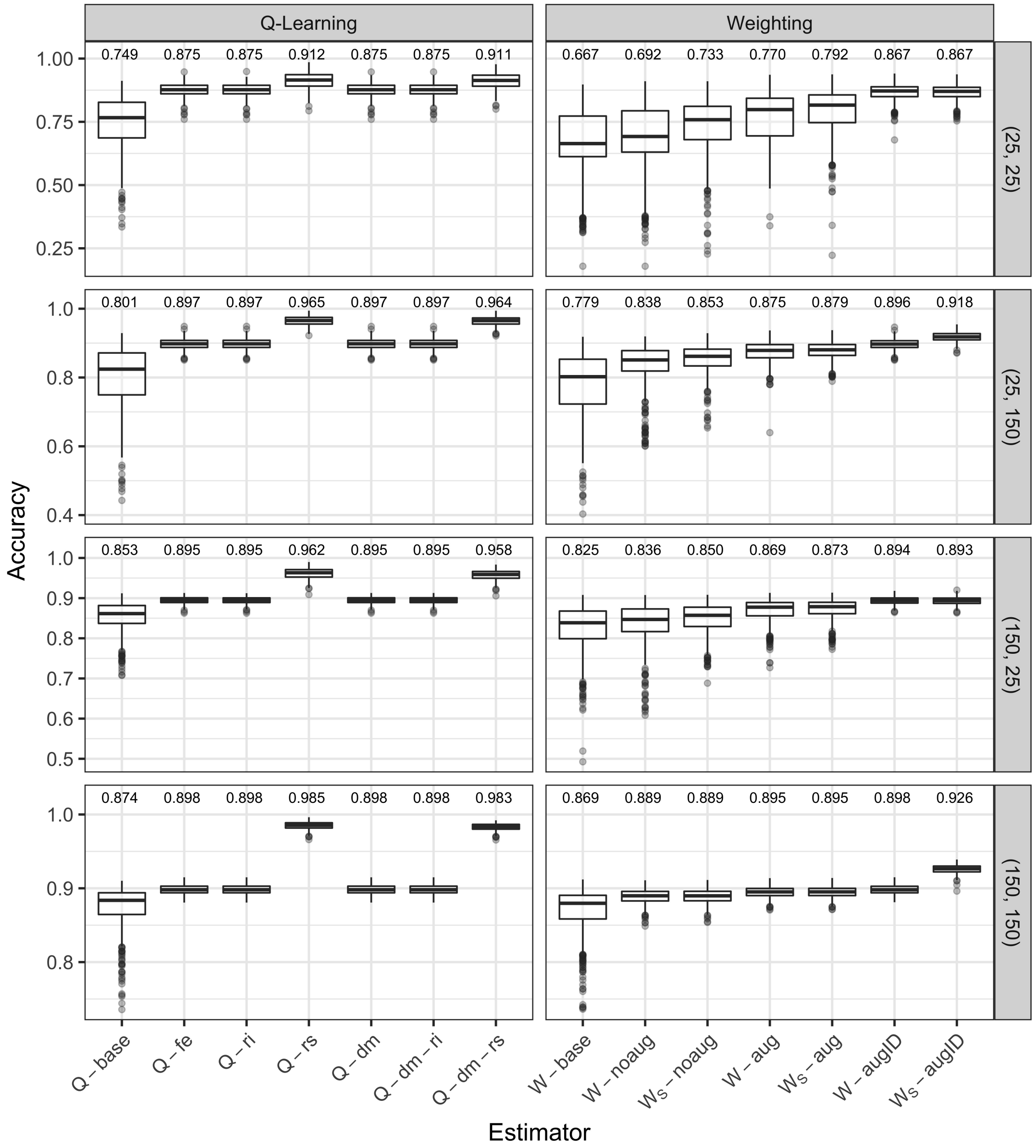

Figure 3: Accuracy of the estimators in Design 2.

in terms of accuracy. In this design, six modifications for Q-learning no longer show similar rates of accuracy. Q-rs and Q-dm-rs outperform other Q-learning modifications, unlike Design 1, and the accuracy gains from Q-rs and Q-dm-rs increase as either the number of clusters or the cluster size increases. Their superior performance is because they use random treatment effects that potentially capture the impact of the unmeasured cluster-level moderator in the design of OTRs. These findings indicate that incorporating the cluster-level impact through random effects will likely lead to higher accuracy rates than using the methods that ignore it. Among weighting methods, W_S-augID that incorporates cluster dummies into the moderator variables and augmentation model performs best across different simulation conditions. This highlights the importance of incorporating multilevel considerations in weighting methods to enhance accuracy. For other performance criteria (i.e., F1 score, correlation, and AUC), see the summary results in Appendix A.

Figure 4 summarize the congruence results under Design 2 when the number of clusters is 25 and the cluster size is 150. Similar to Figure 2, we focus on one sample size condition due

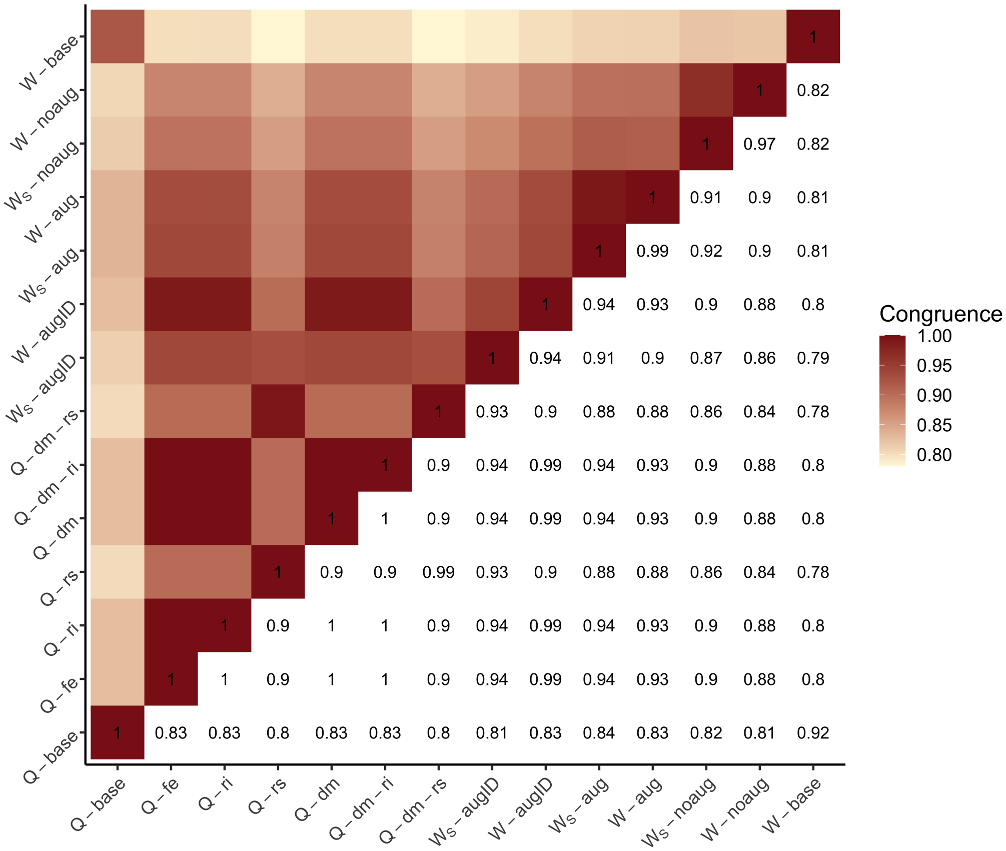

Figure 4: Congruence among the estimators in Design 2 when the number of clusters is 25 and the cluster size is 150.

to similarities in the result patterns across different sample size conditions. Unlike Design 1, Q-learning modifications no longer highly agree with each other and the modifications that share common features (e.g., random treatment effects) show high congruence measures. Specifically, Q-rs and Q-dm-rs show a congruence rate of 99%, but they have about 90% agreement with other Q-learning modifications. Among weighting methods, the congruence rate between W-augID and W$_S$-augID decreases by about 5% from the rate in Design 1. Also, W$_S$-augID exhibits 93% agreement with Q-rs and Q-dm-rs, the two top performers in Q-learning under Design 2, but W-augID shows 3% lower agreement, i.e., 90%. Given these, W$_S$-augID may be a safe way to estimate an OTR using data from multisite randomized trials potentially faced with unmeasured cluster-level moderators.

5.3 Design 3: Nonlinear Main Effects and No Cluster-level Moderator

In Design 3, we generated multisite data similar to Design 1, except we use non-linear main effects in the outcome model as follows:

$$\begin{aligned} Y_{ij}(t) = 0.7 + X_{1ij} + X_{2ij} - 2X_{3ij} + X_{4ij} + 0.5X_{5ij} + \exp(-0.6X_{1ij}X_{2ij} + 0.3X_{1ij}) + \\ t \cdot \Delta(\mathbf{V}_{ij}) + 2.5U_j + \epsilon_{ij} \end{aligned}$$

Although nonlinear main effects of covariates are present, we fit only linear main effect terms to compare the performance of Q-learning versus weighting methods under model misspecification. That is, under Design 3, the outcome regression models for Q-learning are misspecified, and the augmentation models in weighting methods are misspecified.

Figure 5 summarizes the accuracy performance of the estimators in Design 3. As expected, the accuracy rates decrease in this design because of model misspecification. Among Q-learning methods, we observe that the accuracy loss is large (about 9%) under the small sample sizes (here, 25 clusters with the cluster size of 25), and the loss is reduced to about 3-5% as either the number of clusters or the cluster size increases. We also observe that the accuracy rates of Q-rs and Q-dm-rs are slightly smaller than those from other Q-learning methods. Such a performance may be because model misspecification poses more difficulties in estimating random effects that are assumed to follow normal distributions.

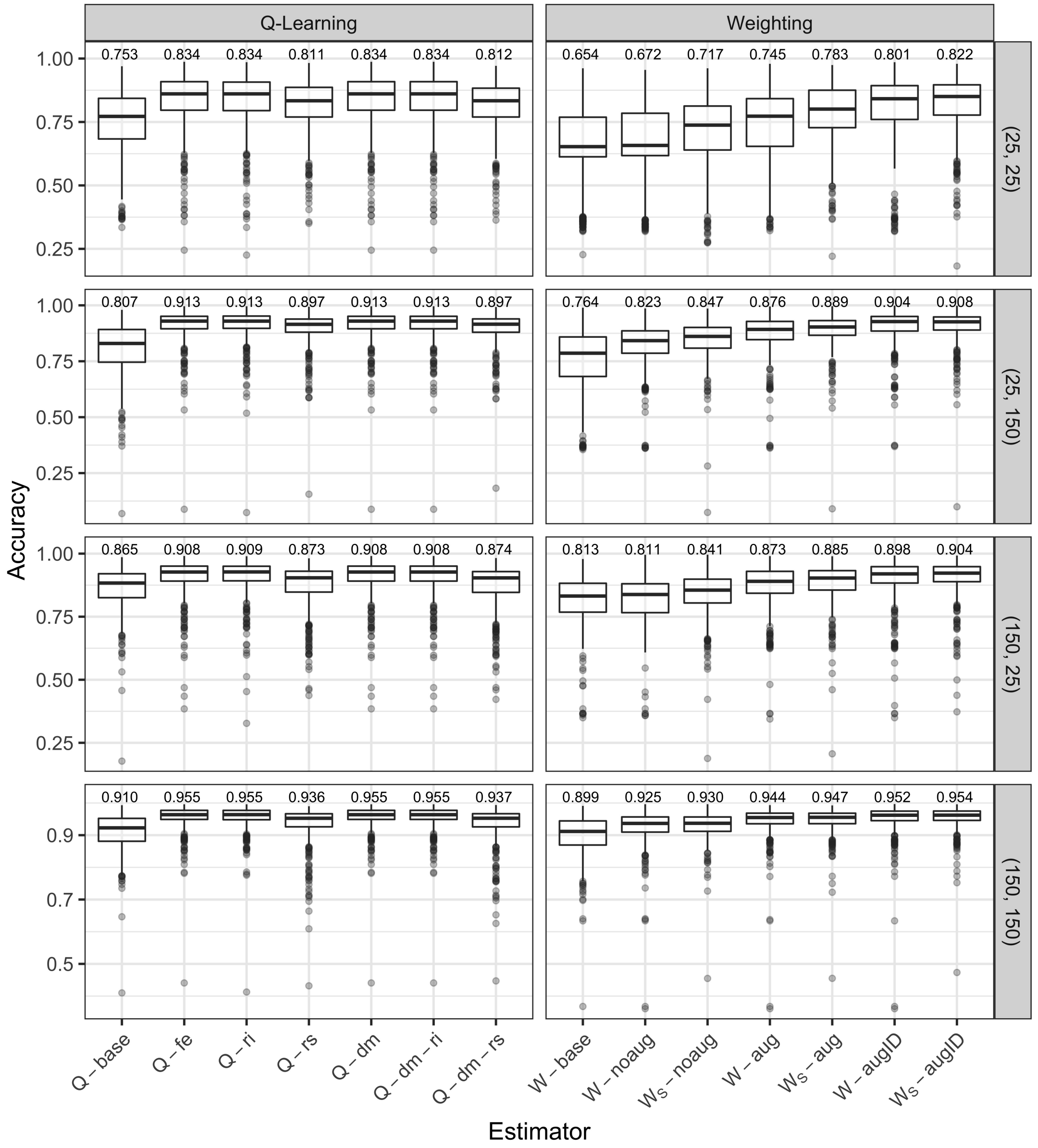

Figure 5: Accuracy of the estimators in Design 3.

Likewise, all the weighting methods show some accuracy losses. Including non-linear terms in the data-generating model deteriorates the performance of modifications without augmentation (i.e., W-noaug and W_S-noaug). Specifically, they exhibit 3-6% loss rates. But adding augmentation, while partially misspecified, into W-noaug and W_S-noaug increases accuracy rates. Similar to Design 1, weighting methods augmented with both covariates and cluster dummies (W-augID and W_S-augID) perform better or no worse than those with covariates only (W-aug and W_S-

aug). Between the modifications without and with augmentation, W-augID and $\mathsf{W_S}$-augID tend to show higher loss rates of accuracy than their counterparts without augmentation, but when the sample sizes increase to our largest sample size, i.e., (150, 150), all weighting modifications show similar loss rates. For other criteria (i.e., F1 score, correlation, AUC, and congruence), see Appendix A.

5.4 Takeaways from Simulations

Our simulation study provides guidelines on how to use Q-learning and weighting methods with multisite randomized trials. These guidelines aim to serve as useful reference points for future empirical or theoretical analyses of OTR methods for multisite randomized trials. Our findings suggest the following:

1. Incorporating multilevel, hierarchical structures through fixed effects models, random effects models, and hybrid models improves the accuracy of Q-learning methods. Specifically, the modifications based on random treatment effects (Q-rs and Q-dm-rs) perform particularly well when a cluster-level moderator is present.
2. Among weighting methods, the modification that includes cluster dummies in moderator variables and augmentation terms (i.e., $\mathsf{W_S}$-augID) shows the most promise in improving accuracy, regardless of the presence of a cluster-level moderator.
3. When the main effect terms of covariates are partially misspecified, the accuracy of Q-learning modifications decreases. But using modifications that account for clustering still improves accuracy compared to baselines that ignore a clustering structure.
4. When augmentation terms are misspecified, the accuracy of weighting modifications decreases in finite samples, but $\mathsf{W_S}$-augID still performs best.
5. Overall, Q-learning modifications perform well when the outcome model is correctly specified. But when the outcome model is misspecified, there is no guarantee that Q-rs and Q-dm-rs will perform best and may even perform worse than the top performer in weighting, $\mathsf{W_S}$-augID.

6 Empirical Example: Conditional Cash Transfer

6.1 Data and Variables

CCT programs have become a type of social assistance program aimed at assisting families living in poverty in developing countries (Barrera-Osorio, Linden, et al., 2019). Barrera-Osorio et al. (2011a) conducted two experiments in Bogotá, Colombia in 2005 to evaluate whether CCT programs could have a greater impact on educational attainment for students from economically disadvantaged backgrounds. In this paper, we focus on the data from their first experiment in San Cristobal, one locality of Bogotá. Barrera-Osorio, Bertrand, et al. (2019) used an oversubscription model, and the eligibility criteria required that students had completed grade 5 (i.e., they must have been in grades 6-11) and that their families were classified into the bottom two categories on Colombia's poverty index, known as the SISBEN. Eligible registrants were randomly assigned between the control group, the basic treatment, and the savings treatment[4], and there were some variations in treatment proportions among schools; see Barrera-Osorio et al. (2011a) for more details of the experiment. We categorized the treatment status into the CCT

[4] The basic treatment gave students US \$15 per month as long as the child attended at least 80 percent of the days that month. The savings treatment gave students two thirds of the amount (i.e., US \$10) and held the remaining third in account for the purpose of preparation for the next school year. The control group didn't receive any of the treatments

treatment versus control (a binary treatment) in the data analysis because basic and savings treatments increase attendance rates to about the same extent in the experiment (Barrera-Osorio et al., 2011a). Our goal in the analysis is to demonstrate our proposed modifications for Q-learning and weighting methods by designing an optimal regime for the CCT program from a multisite randomized trial where students are nested within schools.

For the data analysis, we focused on schools that had attendance data. Note that due to budget constraints, Barrera-Osorio et al. (2011a) collected subsequent attendance data in some schools with a large number of registered students. Further, we excluded (i) households with more than one sibling and (ii) schools with only one student. We deleted the former cases to eliminate spillover effects within households, and we used the household ID variable to check the number of registered siblings. For cases with missing values on the household ID, we borrowed the SISBEN database to identify whether any of the students had the same values for all the household data in the SISBEN database.[5] When they had identical values, we assumed that they came from the same household and thus, we excluded those students. After sample exclusion, our final analysis sample consisted of 3,872 students from 26 schools where the mean school size was 148.9, ranging from 48 to 388.

In the CCT data, we used the average attendance rate as the outcome Y_{ij}. As mentioned above, our treatment of interest is binary with $T_{ij} = 1$ denoting that a student received a CCT treatment and $T_{ij} = 0$ denoting that a student did not receive it. We used 20 pre-treatment covariates, such as gender, age, household income, and grade, as well as school dummies. We determined covariates that potentially interact with the treatment, i.e., moderators, based on prior works from Barrera-Osorio et al. (2011a) and Barrera-Osorio, Linden, et al. (2019). These included grade, the number of years older the child is for their grade, house possession, estrato classification[6], income categories ($\leq$ 380,000 vs. $>$ 380,000 pesos), and school dummies. For more detail about data and variables, see Barrera-Osorio et al. (2011a) and the codebook from Barrera-Osorio et al. (2011b).

We used our proposed modifications, which account for multilevel structures in Section 4, to design optimal recommendations about CCT treatment for students in grades 6-11. We compared them with the baseline estimators that ignore multilevel structures in an outcome model and/or propensity score model (i.e., Q-base and W-base). In our empirical data, the estimate of the average propensity score was 0.63, and the estimates of cluster-specific propensity scores ranged from 0.57 to 0.72. Also, the ICC estimate in the outcome model was 0.05. As for software, we used the R package *personalized* (Huling & Yu, 2021) for weighting methods and our own function named *Qlearn*, based on the R package *lme4* (Bates et al., 2015), for Q-learning methods. R codes for the data analysis are available in the supplemental materials and the first author's GitHub repository.

6.2 Results

Figure 6 summarizes the percentages of students who are recommended to receive the CCT program within each estimator, and the numbers at the bottom of each bar indicate the percentage values of the recommended group. More than 65% of the students are recommended to receive the CCT program across all the estimators, and in particular, all the Q-learning methods recommend the CCT program to more than 72%. For weighting methods, the modifications with augmentation recommend the CCT treatment to a larger number of students than their counterparts without augmentation and behave similarly to Q-learning modifications with respect

[5] We use 13 variables named "s_durables", "s_edadhead", "s_estcivil", "s_estrato", "s_infraest_hh", "s_ingtotal", "s_num18", "s_puntaje", "s_single", "s_tpersona", "s_teneviv", "s_utilities", and "s_yrshead" in the codebook.

[6] A geographic poverty index in Colombia. The estrato numbers range from 1 (the worst streets) to 6 (the best) (Arboleda & Valverde, 2021).

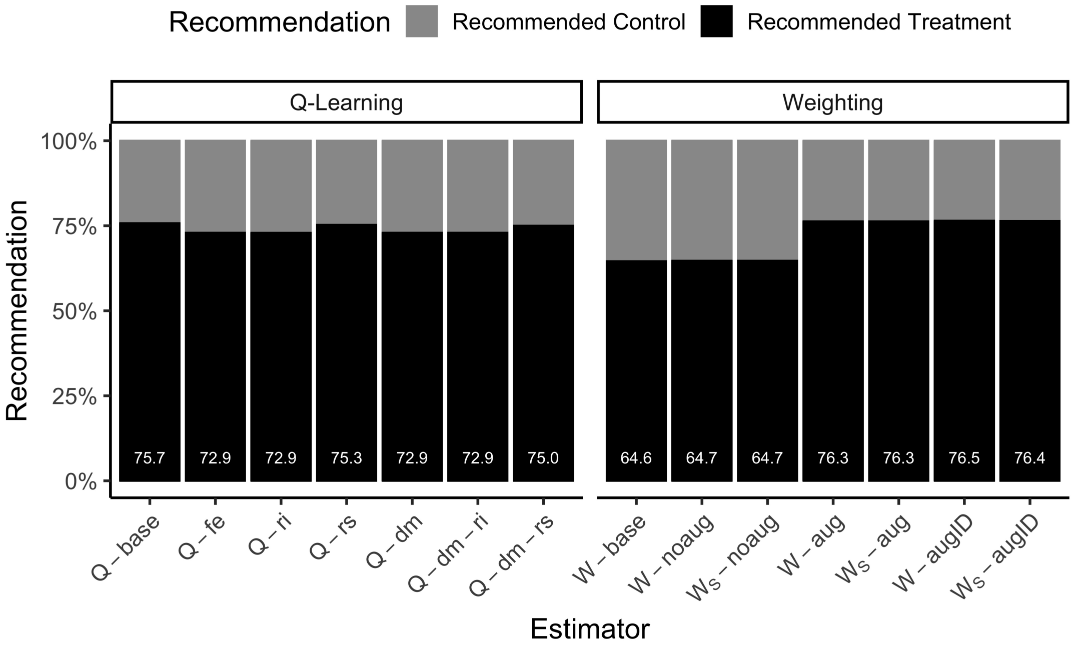

Figure 6: Percentages of recommended and not recommended groups for conditional cash transfer programs

to the percentage of the recommended group.

We also investigated congruence rates across different estimators as in our simulation study; see Figure 7. Again, the congruence rates represent the proportions of individuals who received the same optimal decisions between the estimators. As seen from Figure 7, there are high congruence rates among Q-learning methods (highlighted in red). Among weighting methods, congruence rates are high within the modifications without augmentation or within those with augmentation, but between those without and with augmentation, there are relatively low congruence rates. When comparing Q-learning and weighting methods in our empirical data, weighting methods with augmentation behave more similarly to Q-learning methods, as we observe in our simulations.

We further investigated which covariates work as important moderators that affect CCT recommendations. Figure 8 summarizes the frequencies of the recommended group by income categories: larger than 380,000 pesos and less than or equal to 380,000 pesos (about the bottom two income terciles). For convenience, I call them not-low income and low income, respectively. The numbers at the bottom of each bar represent the percentage values of students from low-income backgrounds among the recommended group. As seen from Figure 8, all the estimators recommend the CCT treatment to students from low-income families more highly than those from not-low-income families, and about 60% of the students among the recommended group come from low-income backgrounds across different estimators. This allocation pattern aligns with the intended goal of the CCT program, which is to encourage the academic participation of students who are economically disadvantaged.

Furthermore, we summarize the distributions of benefit scores by grade among four estimators of interest in Figure 9. The four estimators are (i) the Q-learning estimator with de-meaned variables (Q-dm), (ii) the Q-learning estimator that is based on Q-dm, but adds random school effects and random treatment effects (Q-dm-rs), (iii) the weighting estimator with school-specific propensity scores only (W-noaug), and (iv) the weighting estimator that is based on W-noaug, but adds school dummies into the moderator variables and augmentation terms (W$_S$-augID). The numbers at the top of each boxplot represent the percentages of the recommended group in

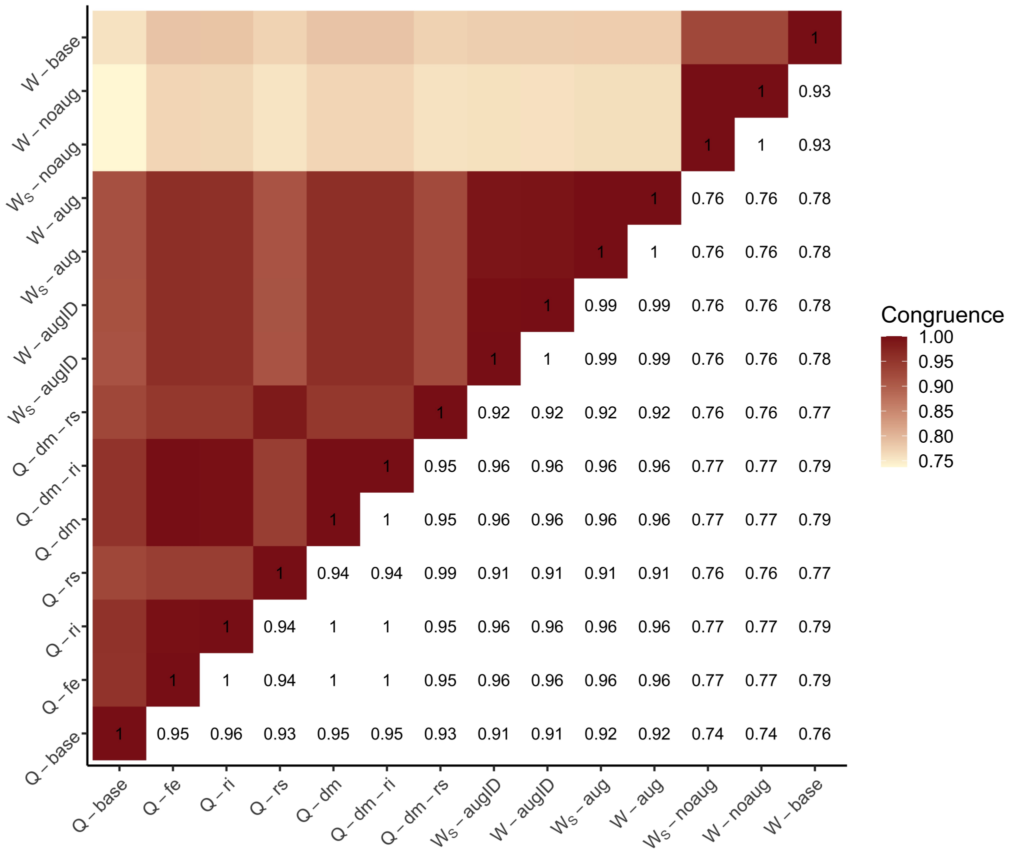

Figure 7: Congruence rates among the estimators in conditional cash transfer programs

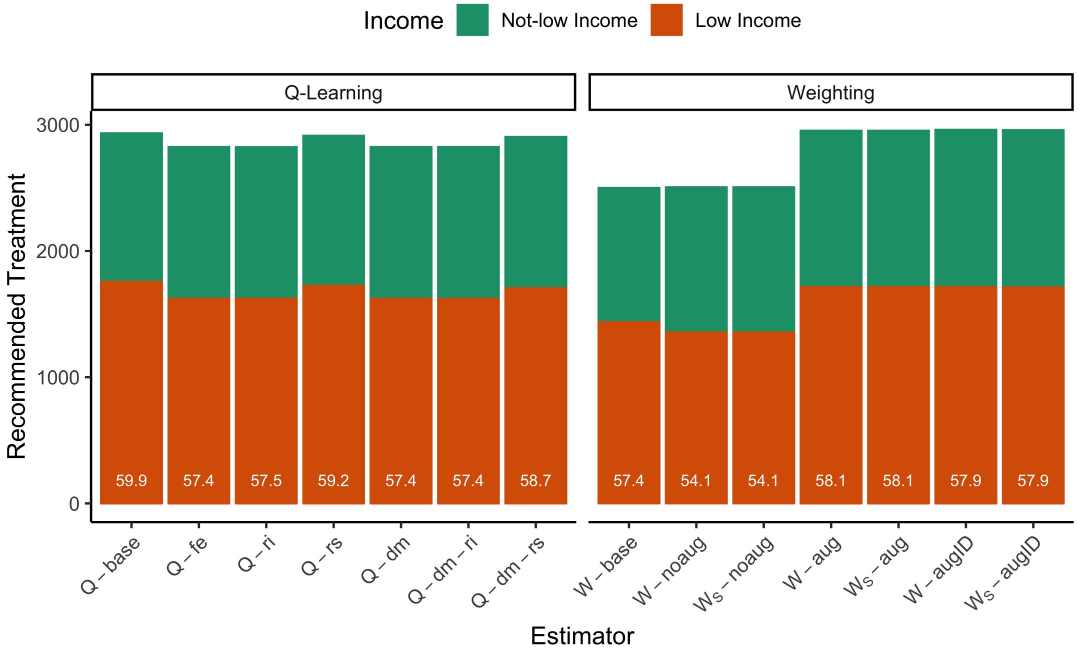

Figure 8: Frequencies of the recommended group by income for conditional cash transfer programs

each grade. Across different estimators, we observe that the distributions of benefit scores vary depending on grades and that grade-specific recommendation rates generally agree with one another, except for W-noaug. There are higher benefit scores and higher rates of the recommended group in grades 6, 9, and 11 than in other grades among the methods except for W-noaug (i.e., Q-dm, Q-dm-rs, and W$_S$-augID), though variations in the benefit scores are relatively small in W$_S$-augID. However, W-noaug's recommendation rates for 8-th and 11-th graders are less than 50% and significantly different from those using other modifications, which show more than 86%. This highlights the importance of including school dummies in the moderator space and augmentation term for weighting methods to accurately reflect potential school/cluster effects.

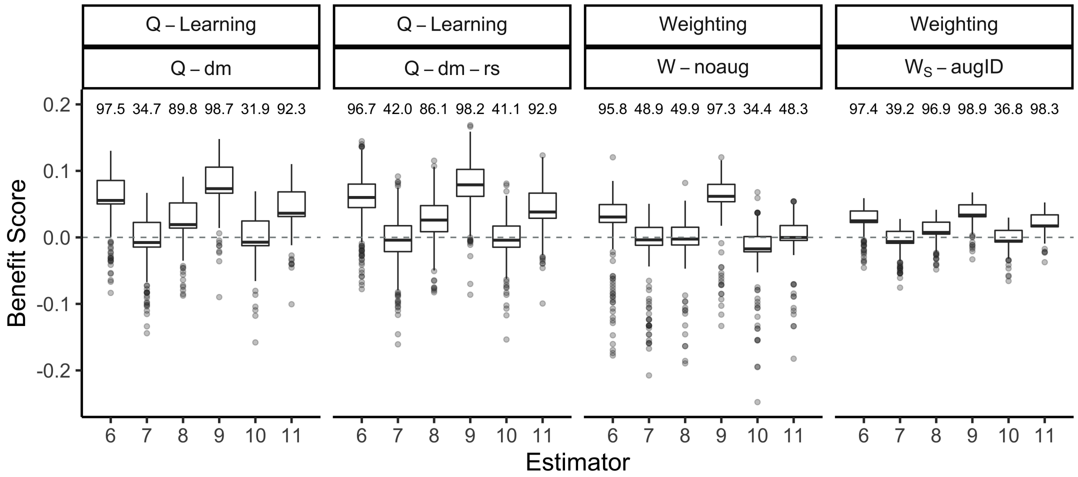

Figure 9: Distributions of benefit scores by grade for conditional cash transfer programs

7 Discussion and Conclusions

In this paper, we proposed different sets of modifications for Q-learning and weighting methods to design optimal, data-driven policies from multisite randomized trials. Our simulation studies show that incorporating multilevel structures in Q-learning methods improves performance. Specifically, the Q-learning estimator with random treatment effects (Q-rs) and the Q-learning estimator with de-meaned variables and random treatment effects (Q-dm-rs) perform well in capturing cluster-specific rules if cluster-level moderators are not measured. Among weighting methods, incorporating cluster dummies into moderator variables and augmentation terms (W$_S$-augID) is found to be the most effective in estimating an OTR from multisite randomized trials. When the outcome model is misspecified, Q-rs and Q-dm-rs may not perform best, but under a misspecified augmentation model, W$_S$-augID is still the top performer and can be even better than Q-rs and Q-dm-rs.

Additionally, our data analysis demonstrates the efficacy of the proposed OTR methods in studying the optimal regime of CCT programs in Colombia. We find that all Q-learning modifications recommend the CCT program to approximately three-quarters of the students, and weighting modifications with augmentation recommend the CCT treatment to a larger number of students than those without augmentation. We also observe a high degree of congruence between the Q-learning modifications and weighting modifications with augmentation, and CCT recommendations are largely influenced by income categories and grades.

When selecting between Q-learning and weighting methods, it is crucial for practitioners to understand the strengths and limitations of each approach (as discussed in Section 3.4) and carefully consider the characteristics of the available data. Specifically, Q-learning methods employed in this study assume the correct specification of the outcome model. Weighting methods, on the other hand, require larger datasets compared to Q-learning methods to achieve the same level of accuracy, as indicated by our simulations. Therefore, if researchers are confident about the correct model specification, Q-learning methods may be recommended. However, when there is insufficient knowledge about the outcome model, weighting methods can be utilized. Regardless of the chosen approach, it is important to incorporate the proposed modifications in Q-learning and weighting methods when working with multisite randomized datasets.

In the fields of education and psychology, the proposed methods hold great potential for various data applications that use multisite randomized designs and seek to tailer interventions based on individuals' differential gains from them. For example, our methods can be particularly valuable in designing course policies within schools. Given that the effect of an advanced math course varies among individual students, our methods can assist in determining whether to offer the course to each student, based on the observed effect heterogeneity. Furthermore, our methods can be effectively employed in designing optimal allocations of extracurricular activities, such as after-school programs, by considering individuals' heterogeneous treatment effects. By utilizing the proposed methods, we believe that individualized education programs can be promoted, offering a more tailored approach instead of a one-size-fits-all approach.

While our proposed methods show promise, we have some suggestions for future research. First, the design rule $\mathcal{D}_\beta$ used in this study was limited to a parametric form and may not perform well if the true OTR has a complex, non-linear form. Future research would explore the use of nonparametric or semiparametric OTRs in multilevel data settings. Second, our additional modifications for weighting were based on the squared loss function with lasso penalty, and using other loss functions may affect the performance in clustered settings. Third, this paper focused on randomized trials and further research is needed to examine the use of these methods in multilevel observational studies. Fourth, researchers may want to use a *within-cluster* approach that estimates an OTR within each cluster by viewing each cluster as a population. However, the within-cluster approach requires large numbers of individuals per cluster, and the cluster-specific OTRs from the approach cannot be generalized to new clusters in future experiments. Fifth, the effectiveness of the proposed modifications may be diminished in small data samples, such as the (25, 25) condition in our study. In such cases, one can consider employing a grouping strategy that combines clusters based on treatment prevalence (Lee et al., 2021; Suk, 2023), or Bayesian OTR estimation methods (Murray et al., 2018).

Sixth, the estimated OTR is subject to uncertainty due to the random nature of the observed data. However, the inference of the estimated OTR is challenging due to the presence of multiple sources of uncertainty (e.g., the propensity score, the OTR itself), making it another active area of research (e.g., Chakraborty et al., 2013; Logan et al., 2019). Future research would investigate the inference of OTRs in multilevel data settings. Seventh, OTR transportability depends on the characteristics of covariates and cluster dummies, as well as our fitting models. When new samples' covariate distributions are not observed in training data, treatment effects by covariates can be extrapolated or interpolated via parametric form assumptions. For new clusters, only random effects or hybrid models allow for treatment effect extrapolation due to their normality assumptions. The quality of extrapolation relies on these assumptions. Eighth, we assumed SUTVA in CCT data analysis due to its plausibility, where there are no spillover effects through interference within each school. To address interference in the same CCT data, we refer readers to the work of Park and Kang (2022). Lastly, we did not incorporate fairness considerations into the design of OTRs although many researchers have raised concerns about fairness-related bias from automatic recommendations and there are some tools available that can detect and resolve such fairness-related biases (Kim & Zubizarreta, 2023; Mitchell et al., 2021; Nabi et al., 2019; Suk & Han, 2023). Future research would examine ways to design OTRs with multilevel data, considering both the performance/utility and fairness aspects, where the definition of fairness is often provided by policymakers and administrators.

Despite these limitations, our proposed modifications for Q-learning and weighting methods have the potential to improve the performance of their baseline counterparts in multisite randomized trials. Additionally, the main ideas presented in this paper can be applied to other methods for OTRs, such as A-learning. We hope that the proposed modifications will serve as useful guidelines for researchers looking to revise OTR methods or apply robust techniques to multilevel studies, with the ultimate goal of providing personalized services in education and the social sciences.

Acknowledgements

The authors thank Jim Corter, Bryan Keller, Larry DeCarlo, Yi Feng, and Weicong Lyu for their useful comments on this work. This research was based upon work supported by the National Science Foundation under Grant No. 2225321. Any opinions, findings, and conclusions or recommendations expressed in the research are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Also, the original collector of the data, ICPSR, and the relevant funding agency bear no responsibility for use of the data or for interpretations or inferences based upon such uses.

References

Agniel, D., Almirall, D., Burkhart, Q., Grant, S., Hunter, S. B., Pedersen, E. R., Ramchand, R., & Griffin, B. A. (2020). Identifying optimal level-of-care placement decisions for adolescent substance use treatment. *Drug and Alcohol Dependence*, *212*, 107991. https://doi.org/10.1016/j.drugalcdep.2020.107991

Arboleda, F. L. T., & Valverde, M. (2021). The travels of a set of numbers: The multiple networks enabled by the Colombian 'estrato' system. *Social & Legal Studies*, *30*(5), 685–703. https://doi.org/10.1177/0964663920960536

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2019). *Replication data for: Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia* (tech. rep.). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E113783V1

Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2011a). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia. *American Economic Journal: Applied Economics*, *3*(2), 167–195. https://doi.org/10.1257/app.3.2.167

Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2011b). Replication data for: Improving the design of conditional transfer programs: Evidence from a randomized education experiment in colombia [Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12]. https://doi.org/10.3886/E113783V1

Barrera-Osorio, F., Linden, L. L., & Saavedra, J. E. (2019). Medium-and long-term educational consequences of alternative conditional cash transfer designs: Experimental evidence from Colombia. *American Economic Journal: Applied Economics*, *11*(3), 54–91. https://doi.org/10.1257/app.20170008

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

Chakraborty, B., Laber, E. B., & Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, *69*(3), 714–723.

Chakraborty, B., & Moodie, E. (2013). *Statistical Methods for Dynamic Treatment Regimes.* Springer. https://doi.org/10.1007/978-1-4614-7428-9

Chen, S., Tian, L., Cai, T., & Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, *73*, 1199–1209. https://doi.org/10.1111/biom.12676

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

Feller, A., & Gelman, A. (2015). Hierarchical models for causal effects. https://doi.org/10.1002/9781118900772.etrds0160

Firebaugh, G., Warner, C., & Massoglia, M. (2013). Fixed effects, random effects, and hybrid models for causal analysis. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 113–132). Springer. https://doi.org/10.1007/978-94-007-6094-3_7

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

Huling, J. D., & Yu, M. (2021). Subgroup identification using the personalized package. *Journal of Statistical Software*, *98*(5), 1–60. https://doi.org/10.18637/jss.v098.i05

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

Kim, K., & Zubizarreta, J. R. (2023). Fair and robust estimation of heterogeneous treatment effects for policy learning. *Proceedings of the 40-th International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2306.03625

Kosorok, M. R., & Moodie, E. E. M. (2015). *Adaptive treatment strategies in practice* (E. E. M. Moodie & M. R. Kosorok, Eds.). Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9781611974188

Lee, Y., Nguyen, T. Q., & Stuart, E. A. (2021). Partially pooled propensity score models for average treatment effect estimation with multilevel data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(4), 1578–1598. https://doi.org/10.1111/rssa.12741

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

Logan, B. R., Sparapani, R., McCulloch, R. E., & Laud, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using bayesian additive regression trees. *Statistical Methods in Medical Research*, *28*(4), 1079–1093.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 331–355. https://doi.org/10.1111/1467-9868.00389

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, *24*(10), 1455–1481. https://doi.org/10.1002/sim.2022

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R., & TenHave, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence*, *88*, S24–S30. https://doi.org/10.1016/j.drugalcdep.2006.09.008

Murphy, S. A., Oslin, D. W., Rush, A. J., & Zhu, J. (2007). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology*, *32*(2), 257–262. https://doi.org/10.1038/sj.npp.1301241

Murray, T. A., Yuan, Y., & Thall, P. F. (2018). A bayesian machine learning approach for optimizing dynamic treatment regimes. *Journal of the American Statistical Association*, *113*(523), 1255–1267. https://doi.org/10.1080/01621459.2017.1340887

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). Efficient Semiparametric Estimation of Network Treatment Effects Under Partial Interference [asac009]. *Biometrika*.

Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics*, *39*(2), 1180–1210. https://doi.org/10.1214/10-AOS864

Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Education Finance and Policy*, *4*(4), 468–491. https://doi.org/10.1162/edfp.2009.4.4.468

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

Raudenbush, S. W., & Schwartz, D. (2020). Randomized experiments in education, with implications for multilevel causal inference. *Annual Review of Statistics and Its Application*, *7*(1), 177–208. https://doi.org/10.1146/annurev-statistics-031219-041205

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. *Proceedings of the Second Seattle Symposium in Biostatistics*, 189–326. https://doi.org/10.1007/978-1-4419-9076-1_11

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

Stefanski, L. A., & Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, *56*(1), 29–38. https://doi.org/10.1198/000313002753631330

Suk, Y. (2023). A within-group approach to ensemble machine learning methods for causal inference in multilevel studies. *Journal of Educational and Behavioral Statistics*, 107699862311620. https://doi.org/10.3102/10769986231162096

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*, 107699862311717. https://doi.org/10.3102/10769986231171711

Suk, Y., & Kang, H. (2022a). 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

Suk, Y., & Kang, H. (2022b). Tuning random forests for causal inference under cluster-level unmeasured confounding. *Multivariate Behavioral Research*, *0*(0), 1–33. https://doi.org/10.1080/00273171.2021.1994364

Suk, Y., Kang, H., & Kim, J.-S. (2021). Random forests approach for causal inference with clustered observational data. *Multivariate Behavioral Research*, *56*(6), 829–852. https://doi.org/10.1080/00273171.2020.1808437

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.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

van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press. https://doi.org/10.1017/cbo9780511802256

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

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3), 279–292. https://doi.org/10.1023/a:1022676722315

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT press.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., & Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat*, *1*(1), 103–114. https://doi.org/10.1002/sta.411

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, *107*(499), 1106–1118. https://doi.org/10.1080/01621459.2012.695674

A Simulation Results

A.1 Design 1

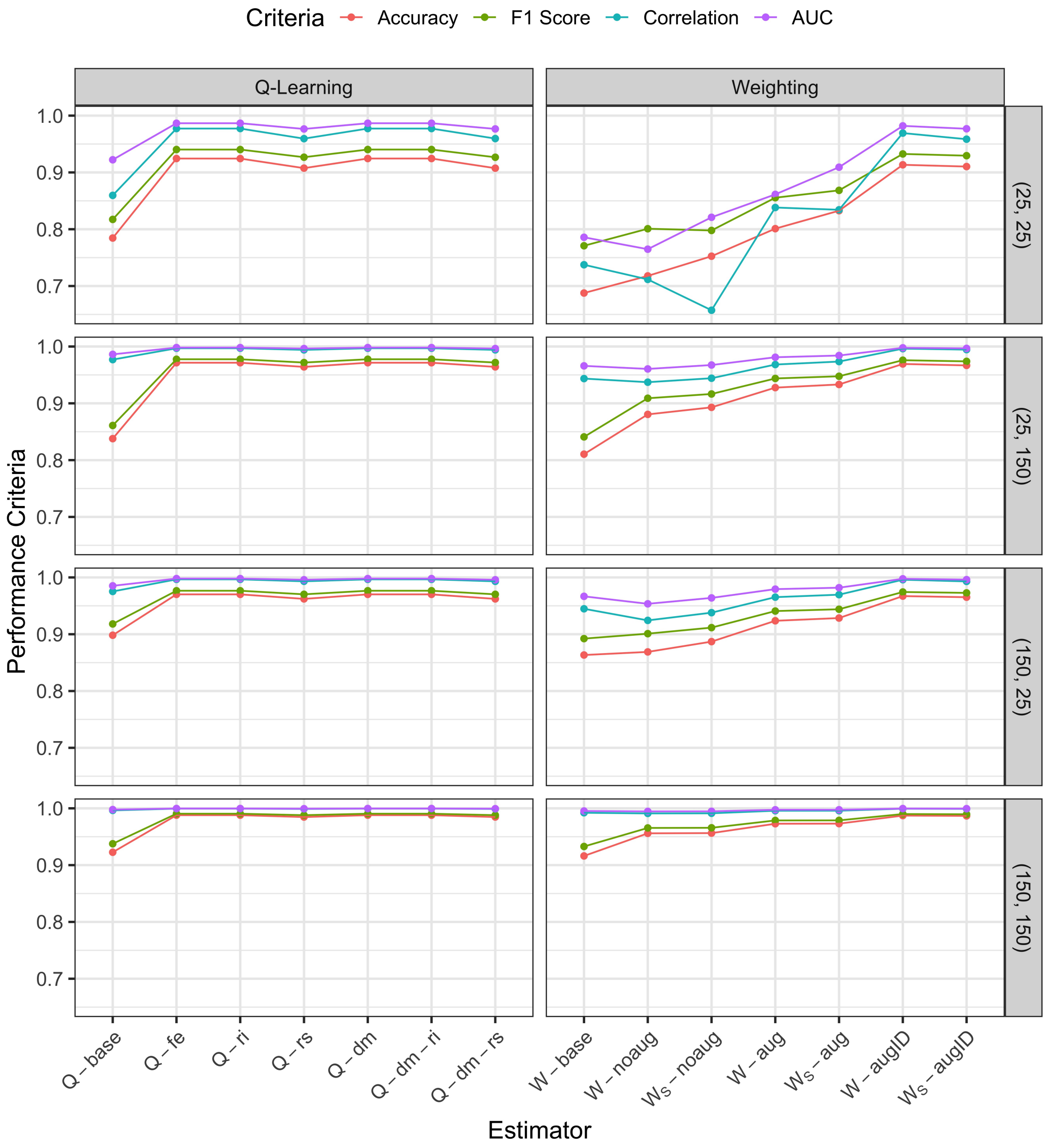

Figure 10: Performance of the estimators in Design 1: accuracy, F1 score, correlation, and the receiver operating characteristic curve (AUC)

A.2 Design 2

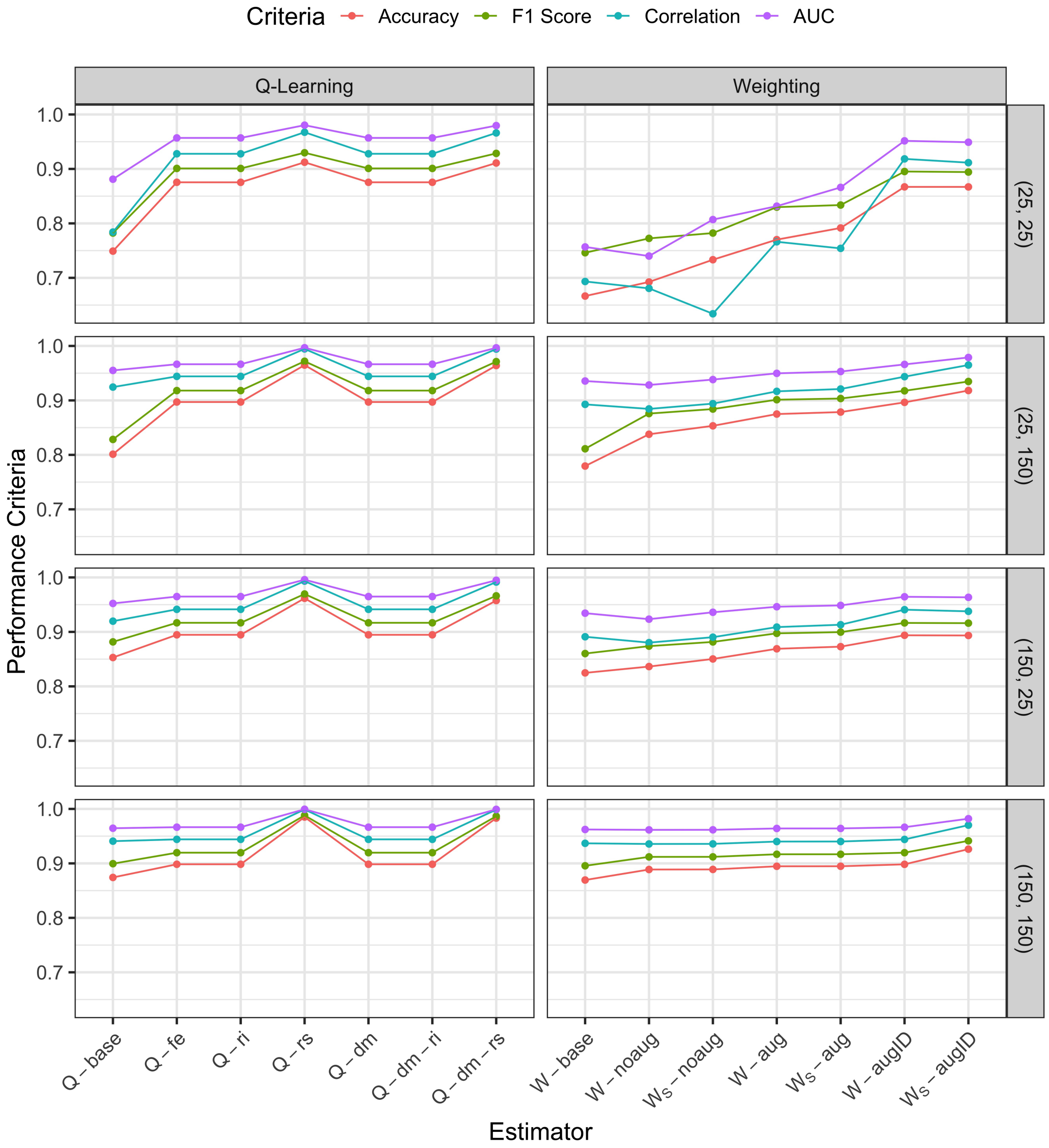

Figure 11: Performance of the estimators in Design 2: accuracy, F1 score, correlation, and the receiver operating characteristic curve (AUC)

A.3 Design 3

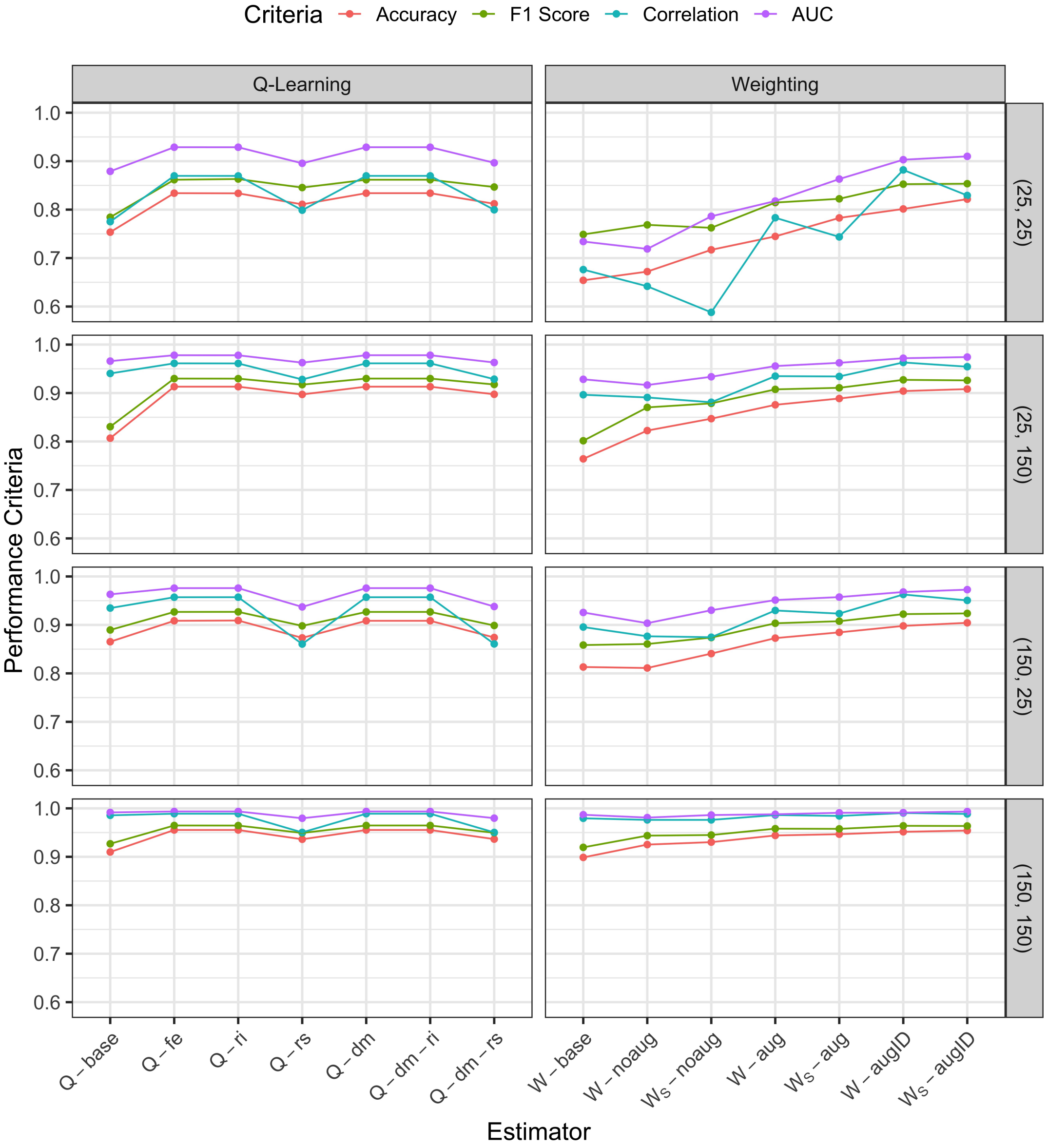

Figure 12: Performance of the estimators in Design 3: accuracy, F1 score, correlation, and the receiver operating characteristic curve (AUC)

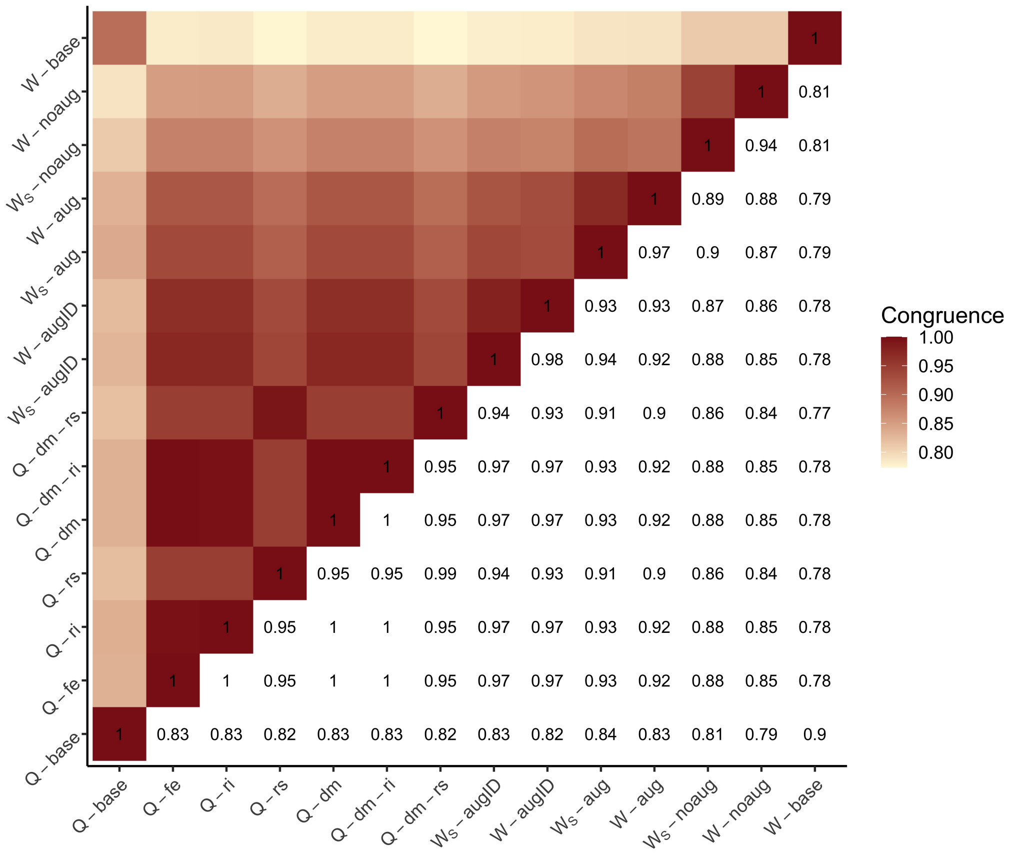

Figure 13: Congruence among the estimators in Design 3 when the number of clusters is 25 and the cluster size is 150.

A.4 Design 4

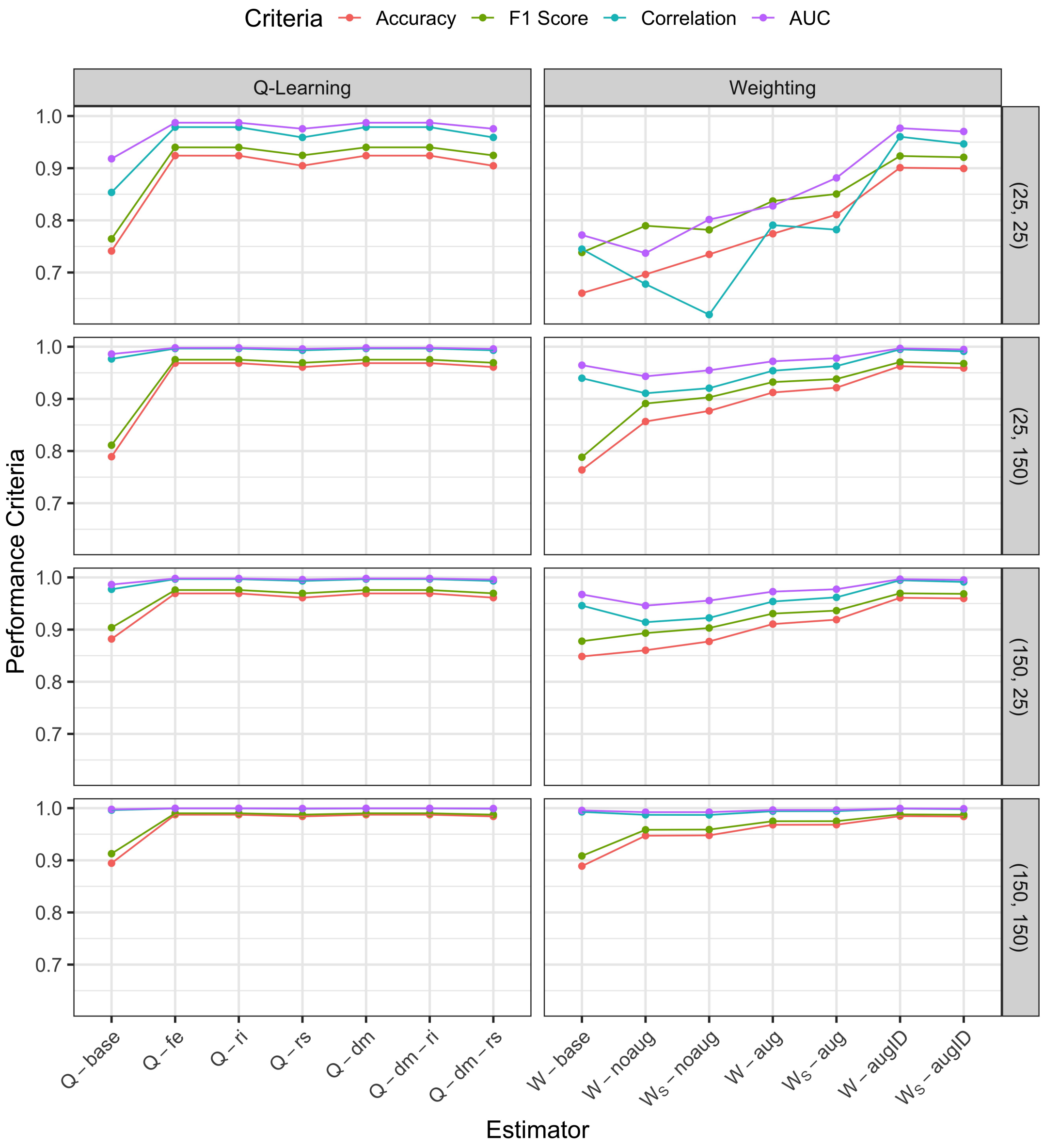

Figure 14: Performance of the estimators in Design 4: accuracy, F1 score, correlation, and the receiver operating characteristic curve (AUC)

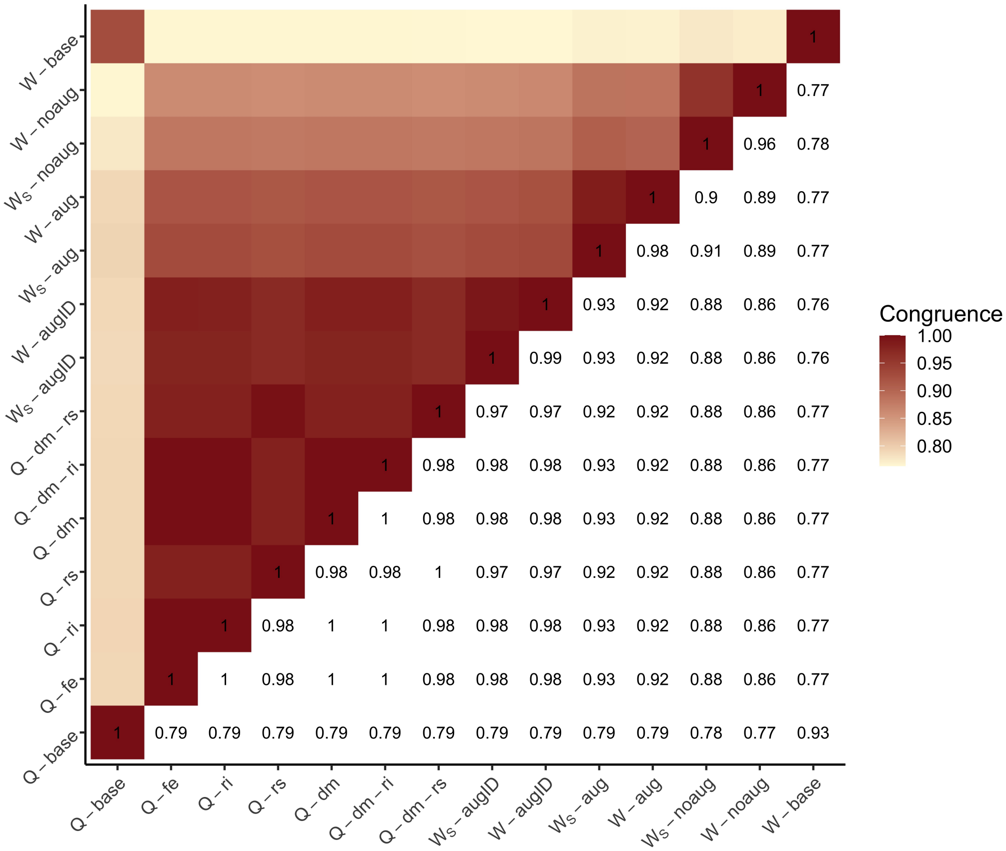

Figure 15: Congruence among the estimators in Design 4 when the number of clusters is 25 and the cluster size is 150.