Evaluating Treatment Effects in Educational Assessment Data

By

Youmi Suk

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Educational Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2021

Date of final oral examination: 5/6/2021

The dissertation is approved by the following members of the Final Oral Committee:

Jee-Seon Kim, Professor, Educational Psychology
Daniel M. Bolt, Professor, Educational Psychology
David Kaplan, Professor, Educational Psychology
Hyunseung Kang, Assistant Professor, Statistics
Peter M. Steiner, Associate Professor, Human Development and Quantitative Methodology, University of Maryland - College Park

Dedication

I dedicate this dissertation to my mother,

who has taught me the values of honesty, wisdom, and diligence.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Researchers often assess causal effects of educational programs or policies using educational assessment data. This dissertation explores novel methods of estimating causal effects in educational assessment data and is broken into three parts. The first part proposes a regression discontinuity design with an ordinal running variable to assess the effects of extended time accommodations for the National Assessment of Educational Progress. The second part investigates how to enhance the performance of machine learning methods to estimate causal effects in multilevel observational data. The third part discusses how to estimate effect heterogeneity that arises from unobservable, latent characteristics by using machine-learning - based methods for causal inference. Overall, the methods from each part provide investigators with modern tools to estimate causal effects in increasingly large and complex educational assessment data.

INTRODUCTION

Researchers who analyze educational assessment data often evaluate causal effects of educational programs or policies (Singer, Braun, & Chudowsky, 2018). In the social sciences, there are rich sets of educational assessment data that have been collected nationally and internationally. For example, national surveys in the United States include the Early Childhood Longitudinal Study (ECLS) and the National Assessment of Educational Progress (NAEP), and cross-national surveys include the Programme for International Student Assessment (PISA) from the Organization for Economic Cooperation and Development (OECD), and the Trends in International Mathematics and Science Study (TIMSS) from the International Association for the Evaluation of Educational Achievement (IEA). These educational assessment data can enable researchers to conduct evidenced-based education research for various purposes for estimating treatment effects, particularly in a setting where a randomized control trial is not feasible for ethical and/or practical reasons.

Despite a large and growing number of datasets publicly available, estimating treatment effects is more challenging in observational studies due to confounding bias, i.e., bias from covariates that affect the treatment assignment and the outcome. Therefore, it is necessary to use appropriate quasi-experimental or non-experimental designs in order to estimate unbiased or consistent treatment effects. In addition to confounding bias, the complexity of the sampling designs (e.g., clustered sampling designs) and the sources of treatment effect heterogeneity (e.g., latent or manifest) can be additional barriers to estimating treatment effects unbiasedly or consistently (Kaplan, 2016). Though there are numerous issues related to causal inference in educational assessment data, the appropriate use of causal inference methods and their

extensions has not been thoroughly discussed in causal inference literature. To fill this gap, this dissertation is a collection of three papers, each of which focuses on the specific context of causal inference in educational assessment data. Among quasi-experimental or non-experimental designs, particularly, this dissertation focuses on (i) regression discontinuity designs and (ii) matching and propensity score designs. Three studies in this dissertation address different methodological challenges and use distinct empirical examples from educational assessment data.

The first study proposes a regression discontinuity design with an ordinal running variable in order to assess the effects of extended time accommodations for the NAEP. This study discusses the causal identification and causal estimation of the average treatment effect and the local average treatment effect at the cutoff of the ordinal running variable and provides a series of sensitivity analyses associated with the scaling function, the cutoff point, and an unmeasured confounder.

The second study investigates how to enhance the performance of machine learning (ML) methods for causal inference to estimate the average treatment effect in multilevel observational data. This study provides different modifications to fine-tune Causal Forests, an ML method based on random forests (Wager & Athey, 2018), to consistently estimate treatment effects in different types of multilevel observational data: two-level data, three-level data, and cross-classified data. The proposed modifications were demonstrated for estimating the effects of private math lessons on students' math achievement scores in the TIMSS data.

The third study discusses how to reveal treatment effect heterogeneity that arises from unobservable, latent characteristics. This study defines the conditional average treatment effects given latent classes, and it proposes a two-step procedure which combines finite mixture models

and ML methods to estimate treatment effects within latent classes. This study's proposed method was applied to estimating the effects of private science lessons on students' science achievement scores in the TIMSS data. Overall, these three studies clarify and provide specific guidance on how researchers should use a regression discontinuity design (with an ordinal running variable) or ML-based causal inference method in educational assessment data. We hope that the methods from each study provide researchers with modern tools to estimate causal effects in increasingly large and complex educational assessment data.

STUDY 1 : REGRESSION DISCONTINUITY DESIGNS WITH AN ORDINAL RUNNING VARIABLE: EVALUATING THE EFFECTS OF EXTENDED TIME ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNERS

Abstract

Regression discontinuity designs are commonly used for program evaluation with continuous treatment assignment variables. But in practice, treatment assignment is frequently based on discrete or ordinal variables. In this study, we propose a regression discontinuity design with an ordinal running variable to assess the effects of extended time accommodations (ETA) for English language learners (ELL). ETA eligibility is determined by ordinal ELL English proficiency categories of National Assessment of Educational Progress data. We discuss the identification and estimation of the average treatment effect, intent-to-treat effect, and the local average treatment effect at the cutoff. We also propose a series of sensitivity analyses to probe the effect estimates' robustness to the choices of scaling functions and cutoff scores, and unmeasured confounding.

This research was supported by a grant from the American Educational Research Association Division D.

Suk, Y., Steiner, P. M., Kim, J.-S., & Kang, H. (2021) Regression discontinuity designs with an ordinal running variable: evaluating the effects of extended time accommodations for English language learners. *PsyArXiv.* Retrieved from psyarxiv.com/sgqjv doi:10.31234/osf.io/sgqjv

Introduction

In educational assessment, there have been ongoing efforts to include English language learners (ELLs) and students with disabilities (SD) in the National Assessment of Educational Progress (NAEP) assessment by providing appropriate testing accommodations (National Research Council, 2002). Among various testing accommodations, the extended time accommodation (ETA) is the most frequently offered accommodation in the NAEP assessment and other testing programs (Gregg & Nelson, 2012); the recent 2017 NAEP assessment included about 90% of ELLs and SD students, and about 10% of these students received ETA (National Center for Education Statistics, 2017a, 2017b). Despite the common usage of ETA, there is little guidance on how to evaluate ETA and no systematic research studies assessing ETA's effectiveness (Jonson, Trantham, & Usher-Tate, 2019). Given that it is unethical and impractical to conduct a randomized experiment in this setting, the goal of this paper is to propose a regression discontinuity (RD) design with an ordinal running variable for evaluating program effectiveness.

RD designs have been used for policy and program evaluation where subjects' treatment status is determined by whether their treatment assignment variable (also called running or forcing variable) exceeds a pre-defined cutoff. If the running variable is continuous, as required by standard RD designs, the average treatment effect (ATE) at the cutoff is non-parametrically identified and can be estimated by comparing the average outcomes of subjects "just below" and "just above" the cutoff (Hahn, Todd, & van der Klaauw, 2001; Imbens & Lemieux, 2008; Lee & Lemieux, 2010). However, in some settings, the running variable is discrete, reported in coarse intervals, or an ordinal variable with a few categories only. For instance, in NAEP assessments, student eligibility for ETA is determined by ELL English proficiency scores, an ordinal variable

with six categories: No Proficiency, ELL Beginning, ELL Intermediate, ELL Advanced, Formerly ELL, and Never ELL. Students with an ELL Advanced (here, the cutoff) or lower proficiency level are offered ETA. Due to the running variable's discrete and ordinal scale, the ATE at the cutoff is no longer non-parametrically identified because in the close vicinity of the cutoff score only ETA eligible students are observed but no ineligible students (i.e., there is no overlap of eligible and ineligible students). Thus, the identification of the ATE at the cutoff requires an appropriate scaling of the ordinal categories together with a correctly specified parametric outcome model to correctly extrapolate the average control outcome of ineligible ETA students to the cutoff category of ELL Advanced students.

In this paper, we extend the identification and estimation strategy proposed by Lee and Card (2008) for discrete running variables to ordinal variables. With ordinal running variables, RD designs face several challenges: First, the categories of the ordinal running variable need to be mapped onto a numeric scale by choosing an appropriate scaling function. For instance, using the ranks {1, ..., 6} is a possible but not necessarily optimal choice for the six ELL English proficiency categories above. Second, a meaningful cutoff value between ELL Advanced and Formerly ELL must be determined with respect to the chosen numeric scale. The choice of the cutoff score determines the target population (at the cutoff) but also the magnitude of the ATE (in the case of effect heterogeneity). Third, the parametric functional form to the left and the right of the cutoff score must be correctly specified with respect to the chosen numeric scale such that valid extrapolations to the cutoff score are guaranteed. Fourth, the statistical model uncertainty due to the discreteness of the scaled running variable should be accounted for when estimating standard errors (Lee & Card, 2008). Fifth, given the increased number of assumptions with ordinal running variables and the potential violations of said assumptions, we need to

conduct a set of sensitivity analyses to strengthen causal conclusions drawn from the analysis. These include probing the effect estimate's sensitivity to (i) the choice of the scaling function, (ii) the choice of the scaled cutoff score, and (iii) unobserved confounding due to model mis-specification.

Throughout the paper, we use the ETA example based on the 2017 NAEP data to demonstrate our proposed approach. We discuss the choices of scaling functions and cutoff scores. We also analyze the intent-to-treat and the local average treatment effect where there exists non-compliance with the assigned ETA status.

Setup

Notation

We use the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974) and its extension to multilevel/clustered data by Hong and Raudenbush (2006) to define treatment effects in sharp and fuzzy RD designs. The NAEP sampling design involves a multilevel structure where students, the study units, are nested within schools. Let $A_{ij} \in \{0, 1\}$ be a binary treatment variable where $A_{ij} = 1$ indicates that student i in school j was assigned to the ETA treatment and $A_{ij} = 0$ indicates the control condition. In a classic RD design, treatment assignment is based on a continuous running variable X_{ij} and a cutoff score x_c such that $A_{ij} = 1$ if $X_{ij} \leq x_c$ and $A_{ij} = 0$ if $X_{ij} > x_c$. Let $Z_{ij} \in \{0, 1\}$ denote the treatment received where $Z_{ij} = 1$ if student i in school j actually received ETA and $Z_{ij} = 0$ if the student did not receive ETA. Note, for a sharp RD design (without non-compliance), assignment status and treatment receipt are identical, i.e., $A_{ij} = Z_{ij}$.

$Y_{ij}(1)$ denotes the potential treatment outcome if student *i* in school *j* were to be eligible for ETA, and $Y_{ij}(0)$ denotes the potential control outcome for the same student but under the control condition. For every student, the observed outcome is linked to the potential outcomes as follows: $Y_{ij} = Z_{ij}Y_{ij}(1) + (1 - Z_{ij})Y_{ij}(0)$. The equality implies the stable unit treatment value assumption (SUTVA; Rubin, 1986), that is, (1) a student's potential outcomes are independent of other students' treatment assignment and (2) there are no different versions of the treatment. Since ETA eligibility is unlikely to have spillover effects to other students and every school uses the same ETA (e.g., fixed amount of extra time), SUTVA is plausible in our study. SUTVA would be violated (i) if some students are aware of other students' ETA eligibility status and this awareness creates negative (or positive) externalities on students' performance and (ii) if the overall fraction of ETA eligible students affect performance of students in a school. Finally, let $\mathbf{W}_{ij}$ be a set of observed pre-treatment covariates and $\mathbf{U}_{ij}$ be unobserved confounders.

Review: Sharp RD Design with a Continuous Running Variable

We first review the standard sharp RD design with a continuous running variable. Suppose our running variable, ELL English proficiency, is continuous where students scoring below or at the cutoff, $X_{ij} \leq x_c$ are eligible for ETA and students scoring above the cutoff are ineligible for ETA. Given full compliance with the assigned ETA status, this design is called a *sharp* RD design because the probability of receiving treatment jumps from one to zero when the running variable X_{ij} crosses the cutoff x_c.

Because A_{ij} is a known deterministic function of X_{ij}, the conditional unconfoundedness assumption automatically holds, that is, $Y_{ij}(1), Y_{ij}(0) \perp A_{ij}|X_{ij}$. Nonetheless, the treatment effect for the entire population is non-parametrically not identified because of a violation of the

positivity assumption ($0 < Pr(A_{ij} = 1|X_{ij}) < 1$). In fact, in a sharp RD design, treatment and control units do not share any common support on X_{ij} such that $Pr(A_{ij} = 1|X_{ij} \leq x_c) = 1$ for every individual scoring below or at the cutoff and $Pr(A_{ij} = 1|X_{ij} > x_c) = 0$ for every individual scoring above the cutoff. But the treatment effect for the population at or in the close vicinity of the cutoff score can be identified by a mild, local smoothness assumption at the limiting cutoff

(A1) Local Continuity of Potential Outcomes:

$$\lim_{x \uparrow x_c} E(Y_{ij}(1)|X_{ij} = x) = \lim_{x \downarrow x_c} E(Y_{ij}(1)|X_{ij} = x),$$

$$\lim_{x \uparrow x_c} E(Y_{ij}(0)|X_{ij} = x) = \lim_{x \downarrow x_c} E(Y_{ij}(0)|X_{ij} = x).$$

This assumption states that the mean potential treatment and control outcomes right below the cutoff are equal to the corresponding mean potential outcomes right above the cutoff. The assumption allows us to think of an RD design as a local randomized experiment where students near the cutoff are randomly assigned to treatment and control conditions (Lee & Lemieux, 2010). Under (A1), the ATE at the cutoff, $\tau(x_c)$, is identified as follows:

$$\begin{aligned} \tau(x_c) &= E\left[Y_{ij}(1) - Y_{ij}(0)\middle|X_{ij} = x_c\right] \\ &= \lim_{x \uparrow x_c} E(Y_{ij}|X_{ij} = x) - \lim_{x \downarrow x_c} E(Y_{ij}|X_{ij} = x) \end{aligned}$$

In our setting, the ATE at the cutoff represents the average effect of ETA for students scoring right at the eligibility cutoff. We can estimate $\tau(x_c)$ by comparing the average outcomes for students scoring "just below" and "just above" the cutoff using non-parametric, local polynomial regression with an optimal bandwidth parameter. We can also include baseline covariates $\mathbf{W}_{ij}$ to

improve the efficiency of the estimated effect. For more details on sharp RD designs and the non-parametric estimation of treatment effects at the cutoff, see Imbens and Lemieux (2008) and Lee and Lemieux (2010).

Review: Fuzzy RD Design with a Continuous Running Variable

In practice, study administrators frequently do not adhere to assignment rules or participants do not comply with the assigned treatment or control status. For instance, students eligible for ETA according to their English proficiency scores may not receive ETA and ineligible students might actually receive ETA due to school- or administrator-specific rules or exemptions. In the presence of noncompliance, we have a *fuzzy* RD design that can identify the intent-to-treat (ITT) and local average treatment effect (LATE) at the cutoff score. Using the ETA eligibility status, A_{ij}, as the "treatment" indicator, the ITT at the cutoff is identifiable and estimable in the same way as the ATE in the sharp RD design. To define and formalize the identification of the LATE, we now use potential outcomes notations for treatment receipt. Let $Z_{ij}(1)$ be a student's ETA receipt status if she were eligible for ETA ($A_{ij} = 1$) and $Z_{ij}(0)$ be the ETA receipt status if she were not eligible for ETA ($A_{ij} = 0$). We also assume $Z_{ij} = A_{ij}Z_{ij}(1) + (1 - A_{ij})Z_{ij}(0)$. Compared to the sharp RD design, the fuzzy RD design is characterized by a discontinuity in the treatment receipt probability of less than one but still requires a discontinuity greater than zero, i.e., $0 < \lim_{x \uparrow x_c} Pr(Z_{ij} = 1|X_{ij} = x) - \lim_{x \downarrow x_c} Pr(Z_{ij} = 1|X_{ij} = x) < 1$. Additionally, the fuzzy RD design needs the following two assumptions.

(A2) Local Monotonicity:

$$\lim_{x \uparrow x_c} Pr(Z_{ij}(1) < Z_{ij}(0)|X_{ij} = x) = \lim_{x \downarrow x_c} Pr(Z_{ij}(1) < Z_{ij}(0)|X_{ij} = x) = 0$$

(A3) Local Exclusion Restriction:

$$\lim_{x\uparrow x_c} Pr(Y_{ij}(1,z) = Y_{ij}(0,z)|X_{ij} = x) = \lim_{x\downarrow x_c} Pr(Y_{ij}(1,z) = Y_{ij}(0,z)|X_{ij} = x)$$

for each $z = 0,1$, and where the potential outcomes $Y_{ij}(a,z)$ are now functions of both the assigned treatment status (a) and the received treatment status (z).

The local monotonicity assumption rules out the presence of defiers at the cutoff, that is, students who would receive ETA if not eligible for ETA but would not receive ETA if eligible. This assumption allows us to identify the ATE for the latent subpopulation of compliers at the cutoff, that is, students who would receive ETA if they were eligible for ETA and who would not receive ETA if ineligible ($Z_{ij}(1) = 1$ and $Z_{ij}(0) = 0$). The treatment effect for the "local" complier subpopulation is referred to as the LATE. The local exclusion restriction states that the potential outcomes depend only on treatment receipt (Z_{ij}), but are unaffected by treatment assignment A_{ij} at the limiting cutoff. With assumptions (A1)-(A3), the LATE at the cutoff is identified as follows.

$$\tau_{LATE}(x_c) = E\left[Y_{ij}(1) - Y_{ij}(0)\middle|X_{ij} = x_c, Z_{ij}(1) = 1, Z_{ij}(0) = 0\right]$$

$$= \frac{\lim_{x\uparrow x_c} E(Y_{ij}|X_{ij} = x) - \lim_{x\downarrow x_c} E(Y_{ij}|X_{ij} = x)}{\lim_{x\uparrow x_c} E(Z_{ij}|X_{ij} = x) - \lim_{x\downarrow x_c} E(Z_{ij}|X_{ij} = x)}$$

In our study, $\tau_{LATE}(x_c)$ is the average effect of receiving ETA among complier students at the cutoff. We can estimate $\tau_{LATE}(x_c)$ by taking the ratio between the difference in the expected outcomes and the difference in the expected treatment receipt probabilities of students "just above" and "just below" the cutoff. Local polynomial regression is frequently used to estimate the numerator and denominator of the ratio. Also, similar to $\tau(x_c)$, we can improve the

efficiency of the estimator by using covariates $\mathbf{W}_{ij}$. Again, for more details, see Imbens and Lemieux (2008) and Lee and Lemieux (2010).

Review: A Graphical Perspective

The data generating process underlying RD designs can be formalized by a causal diagram—a directed acyclic graph (DAG) (Elwert, 2013; Morgan & Winship, 2014; Pearl, 1988, 2009; Steiner, Kim, Hall, & Su, 2017). The DAG for the ETA evaluation is shown in Figure 1.1. The graph highlights that the ELL eligibility status (A) is solely determined by a student's ELL English proficiency score (X), while the ETA receipt status (Z) depends on the eligibility status (A) and on observed and unobserved covariate sets ($\mathbf{W}, \mathbf{U}$). That is, school administrators might offer ETA to ineligible students or withhold ETA from eligible students based on variables captured by **W** and **U**. The set of measured covariates **W** may include student background variables like the number of years exposed to English education or race/ethnicity. The set of unmeasured covariates **U** may include students' academic abilities/skills or the number of English-language books read per month. Since the covariate sets **W** and **U** also affect students' outcome (Y), they confound the causal relation between ETA eligibility (A) and the outcome (Y) and the causal relation between ETA receipt (Z) and the outcome (Y). Thus, without statistical adjustments, the causal effects of ETA eligibility and receipt on student outcome are not identified.

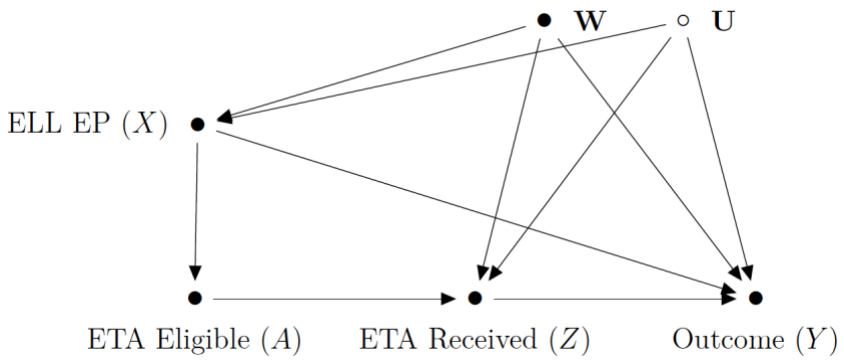

Figure 1.1. Causal directed acyclic graph for evaluating the effects of ETA. *ETA Eligible (*A*) represents students' ETA eligibility status. ELL EP (*X*) represents ELL English proficiency. ETA Received (*Z*) represents whether students received ETA or not, and Outcome represents students' math proficiency outcome.* ***W*** *represents measured covariates, and* ***U*** *represents unmeasured covariates.*

Though the graph suggests that conditioning on ELL English proficiency (X) blocks the confounding backdoor paths between ELL eligibility (A) and students' outcome (Y), the causal effect of A on Y is nonetheless not identified because the positivity assumption is not met because there is no variation in A conditional on X. That is, for each value of the running variable X, we only observe either eligible or ineligible students, but never both (complete lack of overlap). The causal effect of ETA receipt (Z) on the outcome (Y) would be identified conditional on **W** and **U** but only if all variable in **U** were observed and positivity conditional on **W** and **U** would hold. But given the lack of overlap and the presence of unobserved **U**, we exploit the discontinuity at the cutoff rather than matching methods to identify the causal effect of ETA on the math proficiency outcome.

Figure 1.2 shows the causal graph for the RD design at the limiting cutoff score, $X \rightarrow x_c$ (Steiner et al., 2017). In the limit, the running variable (X) still determines ETA eligibility (A) but neither directly affects the outcome (Y) nor is determined by variables **W** and **U**. Thus, in the close vicinity of the cutoff score, ETA eligibility is independent of **W** and **U** (local

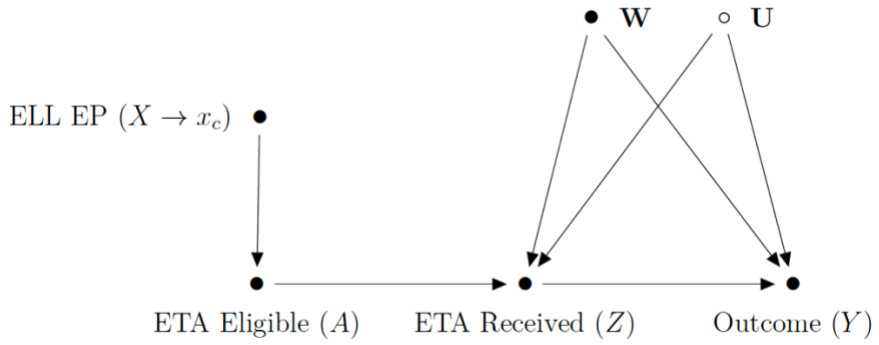

Figure 1.2. Causal graphical identification for evaluating the effects of ETA. ETA Eligible (*A*) represents students' ETA eligibility status. ELL EP (*X*) represents ELL English proficiency. ETA Received (*Z*) represents whether students received ETA or not, and Outcome represents students' math proficiency outcome. **W** represents measured covariates, and **U** represents unmeasured covariates.

randomization), and thus, the ITT at the cutoff (i.e., the effect transmitted along $A \rightarrow Z \rightarrow Y$) is identified without any covariate adjustments for **W** and **U**. Using A as an instrument for ETA receipt (Z) identifies the $\tau_{LATE}(x_c)$ at the cutoff (i.e., the effect $Z \rightarrow Y$).

RD Design with an Ordinal Running variable

So far we have assumed that the running variable X is continuous. However, in practice many RD designs rely on a discrete metric running variable such that the causal effects at the limiting cutoff are no longer non-parametrically identified. The identification of causal effects then requires parametric functional form assumptions to bridge the gap between the neighboring discrete values at the cutoff (Lee & Card, 2008). With ordinal running variables, as for our ETA study, causal identification is even more challenging because the ordinal categories first need to be mapped onto an appropriate numerical scale. In this section, we discuss the identification and estimation of causal effects from RD designs with an ordinal running variable. Drawing valid and reliable causal conclusions from such RD designs requires three main steps. First, researchers need to decide on a reasonable scaling function for the ordinal running variable.

Second, they need to correctly specify the outcome regression with respect to the chosen scaling function and estimate standard errors that reflect specification errors due to the running variable's discreteness. Third, given uncertainties about the appropriate scaling and correct model specification, researchers should always conduct a set of sensitivity analyses to probe the conclusions' robustness to (i) the choice of the scaling function, (ii) the choice of the scaled cutoff score, and (iii) unobserved confounding due to model mis-specification. The following subsections describe these steps in detail and state the causal identifying assumptions.

Scaling Function

The scaling function $S(\omega_k)$ maps the K categories $\omega_1, \ldots, \omega_K \in \Omega$ of the ordinal running variable onto the real numbers $\mathbb{R}$. In classical measurement and test theory, $S(\omega_k)$ can be interpreted as a scaling rule/function. More specifically, each category $\omega_1, \ldots, \omega_K$ is assigned to a numerical score $S(\omega_k) = x_k$ $(k = 1, \ldots, K)$, and we refer $x_1, \ldots, x_K$ as scale values. Importantly, the scaling function S preserves the rank order from the ordinal running variable (Crocker & Algina, 2006).

There is a variety of scaling functions that map ordinal categories onto real numbers. First, ordinal categories can be arranged in ascending order, and their ranks are used as scale values (Crocker & Algina, 2006). For instance, in our ETA study, ELL English proficiency has six categories ranging from No Proficiency to Never ELL. Thus, No Proficiency translates into a scale value of $S(\omega_1) = 1$, ELL Beginning to $S(\omega_2) = 2$, and up to Never ELL with a scale value of $S(\omega_6) = 6$. The rank-based scaling function presumes that there are equal distances between adjacent categorical levels, and the scale values have a meaningful metric interpretation.

Since rank-based scale values might be a poor choice when equal distances between consecutive categories are not suitable for the ordinal running variable, researchers should consider optimal scaling techniques based on observed variables that are directly related to the categories of the ordinal running variable (Bradley, Katti, & Coons, 1962). Such a variable could be the underlying continuous variable that measures the same proficiency/skill or a close proxy thereof (e.g., a continuous composite English proficiency score, domain scores like reading and writing, or English proficiency scores from previous grades). Then, the optimal scale values can be determined using optimal scaling methods for categorical data like categorical regression or categorical principal components analysis. Here, optimality is often defined by maximization of variance, maximization of pairwise linear relationships, and maximization of homogeneity among variables, to name a few. If multiple categories receive similar scale values, we can collapse these categories into one category (Bradley et al., 1962; IBM, 2019; Meulman, 1998; Meulman, van der Kooij, & Duisters, 2019).

In the absence of related continuous variables that could be used for optimal scaling techniques, it is sometimes possible to infer scale values from external sources like published cut scores used to form the categories. More specifically, if there were clearly defined cut scores on the ELL English proficiency variable for creating the ordinal categories, then the midpoints of the cut scores could be used as scale values. However, the midpoints of the lowest and highest category might not be meaningfully defined if the minimum and maximum of the ELL proficiency score are not known or are very extreme values that are rarely observed. If cut scores for the ordinal running variable are not known, we can use other related classifications based on the same or similar underlying continuous variable. For example, the Wisconsin Department of Instruction defines English language proficiency classifications with 7 ordinal levels (ELL

Beginning preproduction, ELL Beginning production, ELL Intermediate, ELL Advanced intermediate, ELL Advanced, Formerly ELL, and Never ELL) that are very similar to the observed ordinal categories of English proficiency in NAEP. After mapping the different ordinal categories, we could use their cut scores and corresponding midpoints to obtain scale values.

Sharp RD Design with an Ordinal Running Variable

As mentioned before, with an ordinal running variable, non-parametric causal identification breaks down because in the close vicinity of the cutoff score (i.e., the ELL Advanced category) only students eligible for ETA are observed but no ineligible control students. Given the identification failure at the limiting cutoff, the limiting graph in Figure 1.2 no longer applies either. Thus, we are back to the graph in Figure 1.1 which indicates that the observed and unobserved sets of covariates **W** and **U** confound the causal relations of interest. But the confounding bias can be successfully removed if the functional relation between the ordinal running variable X and the outcome Y is correctly specified such that a valid extrapolation of the average control outcome from the ETA ineligible to the eligible students at the ELL Advanced cutoff category becomes possible.

To estimate the ATE at the cutoff in the sharp RD design, we extend the approach suggested by Lee and Card (2008) for discrete running variables to ordinal running variables. Instead of the local continuity assumption (A1), we now assume a correctly specified outcome regression such that the expected control outcome can be correctly inferred by extrapolating the outcome regression for control units to the cutoff category.

(A4) Outcome Regression Function:

$$E[Y_{ij}|X_{ij} = x_k] = A_{ij}\tau(x_c) + h_S(x_k) \;\; with \;\; h_S(x_c) = E[Y_{ij}(0)|X_{ij} = x_c]$$

Here, $h_S(x_k)$ is a continuous parametric function with respect to the chosen scaling function $S(\omega_k) = x_k$. The requirement $h_S(x_c) = E[Y_{ij}(0)|X_{ij} = x_c]$ guarantees that the expected potential control outcome at the cutoff score is correctly predicted by $h_S(x_c)$. The parameter $\tau(x_c)$ is the ATE at the cutoff x_c. Due to the discreteness of X_{ij}, $h_S(\cdot)$ must be parametrized by less than K parameters; if $h_S(\cdot)$ uses K or more parameters, $\tau(x_c)$ is not identifiable because of multicollinearity.

In estimating and conducting inference for $\tau(x_c)$, we follow Lee and Card (2008) and add a random specification error with a common variance component for all the observations at any given values of the discrete running variable. This specification error reflects the potential deviation between the expected value of the true outcome at the cutoff and the predicted value based on h_S. The specification error can also be re-interpreted as adding a random effects term in a mixed effects model and hence, makes it easy to implement in practice.

Fuzzy RD Design with an Ordinal Running Variable

For fuzzy RD designs with an ordinal running variable, the ITT at the cutoff category is identified and estimated just like the ATE in the sharp RD design where the treatment assignment status A_{ij} is used as a treatment indicator. Also, the LATE is identifiable and estimable at the cutoff if a functional form assumption analogous to (A4) is met for the expectation of Z_{ij}.

(A5) Treatment Regression Function:

$$E[Z_{ij}|X_{ij} = x_k] = A_{ij}\alpha + g_S(x_k) \quad with\ g_S(x_c) = E[Z_{ij}(0)|X_{ij} = x_c]$$

As before, $g_S(x_k)$ is a continuous parametric function with respect to the scaling function $S(\omega_k) = x_k$. The term α represents the discontinuity in treatment probabilities at the cutoff and is used to obtain the LATE at the cutoff by dividing the ITT by α. As for Assumption (A4), identifiability demands less than K parameters for $g_S(\cdot)$. Also, in estimating standard errors for the LATE, we take random specification errors in the treatment regression function into account.

Sensitivity Analyses

The discussions of the assumptions for the sharp and fuzzy RD design revealed that the causal effects at the cutoff are identified only if the scaling function and outcome (or treatment) regressions are correctly specified. Given that the correct specification of the functions is uncertain in practice, researchers should always conduct sensitivity analyses to check the conclusions' robustness to (i) the choice of scaling functions, (ii) the choice of cutoff values, and (iii) the presence of unobserved confounding resulting from a mis-specified outcome or treatment regression (that lead to biased extrapolations of the control outcomes to the cutoff value).

First, in choosing different scaling functions $S(\omega_k)$ for the ordinal running variable, researchers can probe the robustness of effect estimates (ATE, ITT, and LATE at the cutoff) to alternative plausible choices of $S(\omega_k)$. As mentioned before, the scaling function for the ordinal running variable can be determined by relevant external criteria or by optimal scaling methods with respect to the underlying continuous variable or a proxy measure thereof. If the effect

estimates are rather insensitive to the choice of scaling functions, researchers can be more confident in their conclusions about the presence or absence of a treatment effect.

Second, for the preferred choice of the scaling function, one may evaluate the effect estimates at different cutoff value x_c between the scale value of the cutoff category $S(\omega_c)$ and the value of the neighboring category $S(\omega_{c+1})$. Such a sensitivity analysis reflects the fuzziness of the cutoff point due to the use of the ordinal running variable and provides further evidence for the robustness of the effect estimates. Note that differences in the effect estimates can be expected because different cutoff values refer to different local populations.

Finally, it is advisable to probe whether the conclusions drawn are sensitive to unblocked confounding due to model mis-specification of h_S (and g_S for the LATE). The effect estimates at the cutoff are unbiased only if the extrapolations are based on a correctly specified functional form h_S (and g_S) with respect to the chosen scaling function S. With a mis-specified functional form, the extrapolation to the cutoff score may become invalid and fail to completely remove confounding bias between the outcomes of treatment subjects in the cutoff category and the outcomes of control subjects in the neighboring category. For example, suppose that we simply compare the mean outcomes of the neighboring categories “Advanced ELL” (treated subjects) and “Formerly ELL” (control subjects) in our empirical example. That is, we make no attempt to remove any confounding between the treatment and control groups. Thus, the resulting unadjusted effect estimate likely suffers from confounding bias due to group differences in ELL English proficiency categories and in any other student characteristics like ability or the number of English-language books read per month. The RD design with an ordinal running variable tries to overcome differences between the neighboring treatment and control groups by a correct specification of h_S (and g_S). Since mis-specified functional forms may remove a part but not all

the confounding bias, researchers should conduct sensitivity analyses and report how large the remaining confounding would need to be to overturn the conclusions of the study. We demonstrate all these sensitivity analyses with our empirical example in the next section.

Empirical Example: Testing Accommodations in NAEP

Data and Variables

NAEP is the largest nationally representative and continuing assessment of what students in the US know and can do in various disciplines. The data has been collected by National Center for Education Statistics (NCES) within the Institute of Education Sciences. The 2017 NAEP assessments were conducted for grades 4 and 8 in mathematics, reading, and writing. The NAEP sampling procedures ensure that the students and schools selected in NAEP are representative of the target population. The NAEP assessment strives to minimize participant burden by giving students a subset of items from the total item pool (Johnson, 1992; Oranje & Kolstad, 2019). For more details of the NAEP methods and procedures, see the NAEP page of the NCES website (https://nces.ed.gov/nationsreportcard/tdw/).

In our study, we used the NAEP Grade-4 2017 restricted-use data for mathematics. For the data analysis, we excluded (i) schools with only one student, (ii) students with disabilities, and (iii) ELL students whose prior performance was below the grade level of performance of NAEP; here, ELL students' prior performance was evaluated by their teachers or school staff members through an ELL questionnaire. After sample exclusion, our final analysis sample consisted of 116,910 students from 7,450 schools (78.2% of the original reporting sample). Note that numbers are rounded to nearest tens.

In the 2017 NAEP data, we used the math proficiency as the outcome Y_{ij}. As mentioned earlier, ETA eligibility A_{ij} is binary with $A_{ij} = 1$ denoting that a student is eligible for ETA and $A_{ij} = 0$ denoting that a student is ineligible for ETA. The eligibility status is determined by ELL English proficiency, derived from students' prior ELL status and ELL English proficiency in reading. Specifically, ELL English proficiency was reported with six ordinal levels: No Proficiency, ELL Beginning, ELL Intermediate, ELL Advanced, Formerly ELL, and Never ELL; see survey questionnaires from the NAEP website (https://nces.ed.gov/nationsreportcard/experience/survey_questionnaires.aspx). However, in following school-specific rules and resources, school staff decided which students actually received ETA. The data indicate two-sided non-compliance; some students eligible for ETA did not receive ETA ($A_{ij} = 1$ but $Z_{ij} = 0$), while some ineligible students received ETA ($A_{ij} = 0$ but $Z_{ij} = 1$). For the outcome and treatment regressions, we used a set of pre-treatment covariates $\mathbf{W}_{ij}$, including gender, race/ethnicity, free lunch status, English instruction period, US school period, and primary language; these variables partially explain why some eligible students did not receive ETA. For a full list of pre-treatment covariates and their distributions, see Appendix A.

In demonstrating our proposed approach, we use only a single plausible value of math proficiency as the outcome Y_{ij} (mean=244.05, SD=28.25) and do not incorporate multiple plausible values; in the 2017 NAEP data, each student obtained 20 plausible values in math proficiency because they received only a randomly selected subset of items from the total item pool. We also ignore sampling weights and the corresponding jackknife replicate weights provided by the 2017 NAEP data. This allows us to discuss the analyses without getting

distracted by the more complex measurement and sampling design. But our results do not generalize to the target population of NAEP.

Choice of Scaling Function and Cutoff Score

To assess the effects of ETA with an RD design, we first chose the scaling function for the ordinal ELL English categories based on their ranks, $S(\omega_k) = k$, assuming that the performance differences between consecutive categories are approximately equidistant. The equidistance assumption might be justified because we found the differences in the reading proficiency scores across consecutive categories were similar except for the two extreme levels (No Proficiency and Never ELL) from the 2017 NAEP data (see Figure 1.3). Thus, we assigned 1 to No Proficiency, 2 to ELL Beginning, 3 to ELL Intermediate, 4 to ELL Advanced, 5 to Formerly ELL, and 6 to Never ELL. The cutoff score is $x_c = 4$ which corresponds to the "ELL Advanced" category.

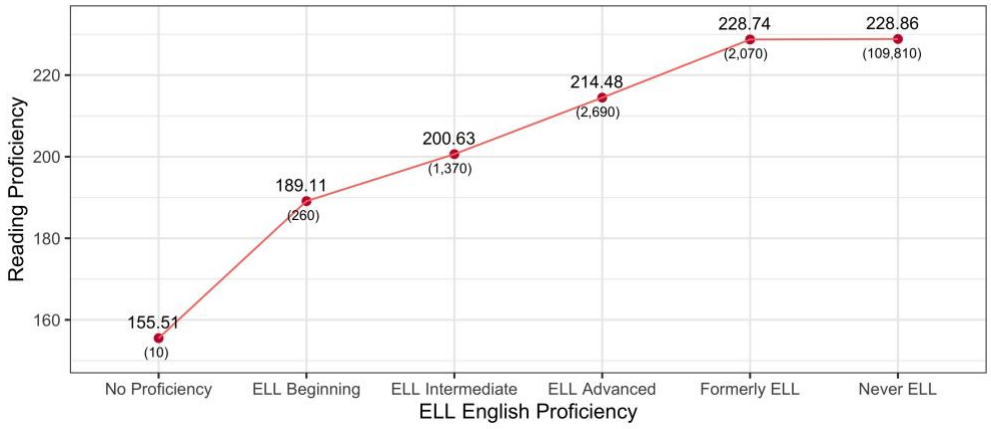

Figure 1.3. Observed means in reading by ELL English proficiency
NOTE: Numbers in parentheses represent sample sizes and are rounded to nearest tens.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

Parametric Model Specifications

Given two-sided noncompliance, we estimate the ITT and the LATE at the cutoff. Following Lee and Card (2008), we model the categories of the ordinal running variable as random effects in a hierarchical linear model that also accounts for the clustered data structure of students nested within schools. The reading proficiency Y_{ijk} for a student i, school j, and ELL category k is modeled as follows:

$$Y_{ijk} = \beta_0 + \beta_1 A_{ijk} + \beta_2 (X_{ijk} - x_c) + \beta_3 A_{ijk} (X_{ijk} - x_c) + \sum \beta_w W_{ijk} + s_j + u_k + \epsilon_{ijk} \quad (1.1)$$

$$A_{ijk} = \begin{cases} 1, & X_{ijk} \leq x_c \\ 0, & X_{ijk} > x_c \end{cases}, \quad x_c = 4$$

In the model equation, the ITT at the cutoff is given by $\beta_1 = \tau(x_c)$ where the cutoff is set to the "ELL Advanced" category, that is, the average effect of ETA eligibility for the students with the "ELL Advanced" level. The term β_0 represents the estimated average control outcome at the cutoff. The term β_2 represents the slope coefficient between scale values above the cutoff value and the outcome, and the term β_3 represents the difference in the slope between scale values below the cutoff value and those above the cutoff value. Due to limited number of ELL categories, we did not consider a polynomial term that could lead to overfitting. The model also includes a set of measured pre-treatment covariates $\mathbf{W}_{ij}$, the school random effects term s_j, and the random specification error u_k. We used the package *lme4* (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2020) to estimate the parameters and their standard errors.

For estimating the LATE effect at the cutoff, we use the following model for Z_{ijk}:

$$Z_{ijk} = \alpha_0 + \alpha_1 A_{ijk} + \alpha_2 (X_{ijk} - x_c) + \alpha_3 A_{ijk} (X_{ijk} - x_c) + \sum \alpha_w W_{ijk} + s_j + u_k + \epsilon_{ijk} \quad (1.2)$$

The term α_1 represents the discontinuity in the treatment receipt probabilities between eligible students and non-eligible students at the cutoff, and the term α_0 represents the estimated average probability of receiving ETA for non-eligible students at the cutoff. The term α_2 represents the slope coefficient between scale values above the cutoff value and the treatment, and the term α_3 represents the difference in the slope between scale values below the cutoff value and those above the cutoff value. Similar to the outcome model, we also add measured covariates $\mathbf{W}_{ij}$, the school random effect term s_j, and the random specification error u_k, and we used package lme4 (Bates et al., 2015) in R (R Core Team, 2020) to estimate the parameters of model (2). Afterwards, we run an instrumental variables regression as laid out in Appendix B to estimate the LATE. Specifically, we treat the model for Z_{ijk} as the first-stage regression and the model for Y_{ijk} as the second-stage regression and utilize two-stage least squares with random effect terms to estimate the LATE. We used cluster bootstrap sampling with 2000 replicates to estimate the standard errors[1].

Results

Table 1.1 summarizes students' ETA eligibility, as defined by ELL status, and whether they actually received the ETA. Overall, about 4.2% of the students (4,940 students) in our study sample were eligible for ETA; these are denoted as ELL students in the table. Among those who

[1] Our implementation of cluster bootstrap sampling does not account for the uncertainty associated with the random specification error. But in our empirical example, we found that the cluster bootstrap standard error (SE) for ITT at the cutoff was larger than that from a mixed effect model, and we're confident that at least the cluster bootstrap SE for LATE at the cutoff was not underestimated. For more information on variance estimation of the LATE estimate, refer to Appendix B in Lee and Card (2008).

were eligible for ETA, about 32.5% of the ETA-eligible students (i.e., ELL) actually received ETA. Also, we saw that few students received ETA even though they were not eligible for ETA (i.e., non-ELL), likely due to test irregularities; see more details on the compliance rate by ELL English proficiency categories in Appendix C.

Table 1.1
Compliance

	ETA		Total
Eligibility	Non-Received	Received	
Non-ELL	111,940	30	111,970
ELL	3,340	1,610	4,940
Total	115,270	1,640	116,910

NOTE: Numbers are rounded to nearest tens. Details may not sum to a total due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

Figure 1.4 provides a visual representation of the RD design where the x-axis represents the rank-based scale values of English proficiency categories with the cutoff point defined at ELL Advanced and the y-axis represents math proficiency scores. We see that the mean math proficiency score of Never ELL was similar to that of Formerly ELL, and the mean math proficiency score increased when ELL English proficiency increased from No Proficiency to ELL Advanced. We can also visually see that the effect of ITT at the cutoff is likely going to be small.

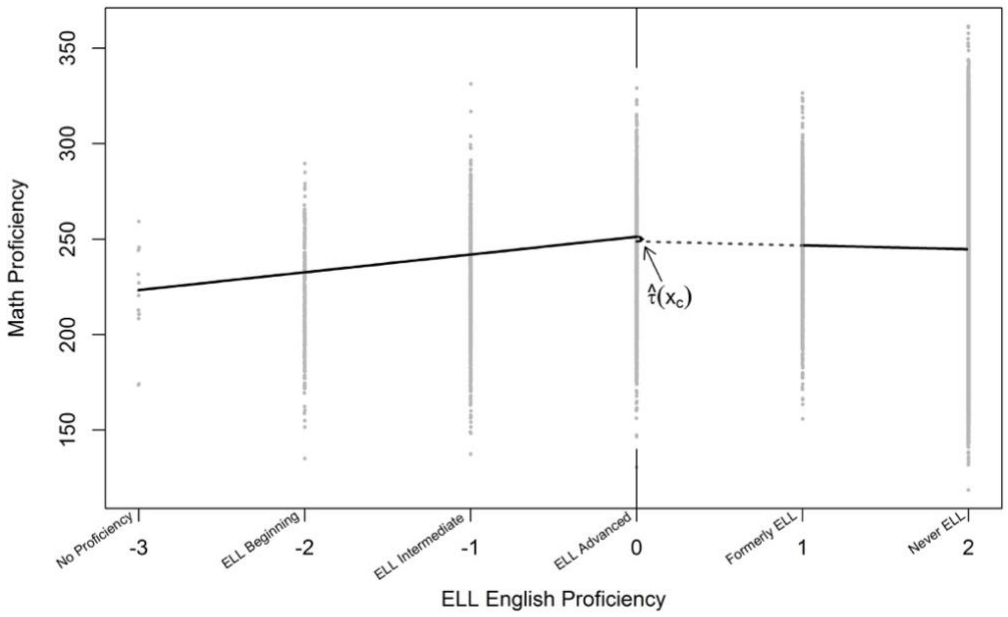

Figure 1.4. Regression discontinuity design for evaluating the effects of ETA in mathematics. The solid black lines represent the estimated regression function, and the red dotted line represents the extrapolated line from the regression function. Gray points indicate students' math scores.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

Indeed, Table 1.2 shows that the ITT estimate was small and not significant. In contrast, the effect estimate of LATE was larger than that of ITT, and it was significantly positive. To summarize, there is no evidence of the effect of ETA eligibility on math proficiency scores at the cutoff, while there is strong evidence to suggest the effect of receiving ETA on math proficiency scores among complier students at the cutoff.

Table 1.2
ITT and LATE estimates at the cutoff

Estimand	Estimate	Std. Error
ITT	2.66	2.18
LATE	12.99	3.14

NOTE: ITT represents intent-to-treatment effects, and LATE represents local average treatment effects.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

Sensitivity Analyses

Methods We assess the sensitivity of our conclusions against (i) the choice of scaling functions, (ii) the choice of the cutoff values, and (iii) unmeasured confounding. To probe the results' sensitivity to different choices of scaling functions, we used a different set of scale values based on the cut scores of the proficiency categories from the ACCESS for ELLs exam in World-Class Instructional Design and Assessment (WIDA)[2]. WIDA's English proficiency levels consist of Entering, Emerging, Developing, Expanding, Bridging, and Reaching for ELL students and former ELL students (WIDA, 2013; WIDA, 2019). From the description of WIDA's proficiency levels, we considered Entering, Emerging, Developing, Bridging, and Reaching as No proficiency, ELL Beginning, ELL Intermediate, ELL Advanced, and Formerly ELL, respectively, and we used the cut scores from grade 4 reading proficiency categorical levels (WIDA, 2013) to determine the scale values. Specifically, based on the cut scores of the WIDA's proficiency levels, we computed the mean scores (i.e., midpoints) of the first four proficiency levels and used the relative differences between consecutive proficiency levels as scale values. For non-eligible students, we assigned the same scale value to the last two categories because Formerly ELL and Never ELL's reading performance was similar from Figure 1.3. Ultimately, a set of (-5.6, -2.2, -1, 0, 1, 1) was used as new scale values that were centered at the cutoff of "ELL Advanced." Also, we varied the cutoff value by re-defining the cutoff value as the mean between the scale values of "ELL Advanced" and "Formerly ELL"; we call the new cutoff point "Between."

Regarding the sensitivity analysis against unmeasured confounding, we used the general bias formula from VanderWeele and Arah (2011), where the potential bias d arising from

[2] ACCESS stands for Assessing Comprehension and Communication in English State-to-State for English Language Learners, and it is a large-scale English language proficiency test for K–12 students (WIDA, 2019).

unmeasured confounders can be represented by the product $\delta\gamma$. δ characterizes a constant prevalence difference of a binary hypothetical confounder between the treatment groups and γ characterizes a constant outcome difference between the levels of the hypothetical confounder.

Results Figure 1.5 visualizes the sensitivity analysis for ITT estimates by changing the choice of the cutoff value and the choice of the scaling function. Table 1.3 provides the corresponding numerical results.

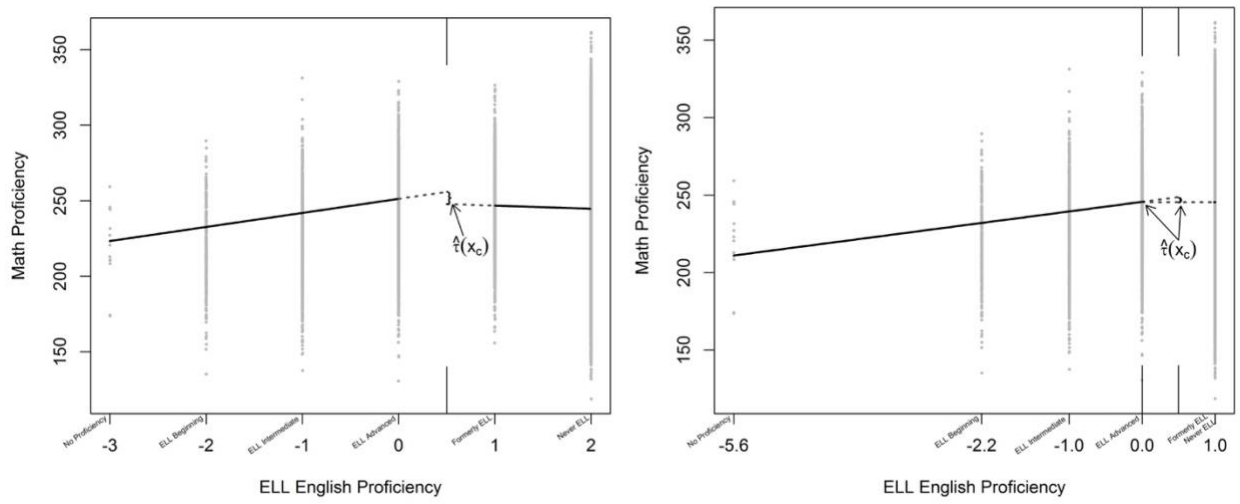

Figure 1.5. Sensitivity analysis against scaling. The solid black lines represent the estimated regression function, and the red dotted lines represent the extrapolated line from the regression function. Gray points indicate students' math scores.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

Table 1.3
Sensitivity analysis: scaling

	Scaling-Cutoff	Estimate	SE
ITT	Rank-Between	8.32	2.05
	WIDA-ELL Advanced	1.26	8.31
	WIDA-Between	2.78	8.75
LATE	Rank-Between	12.99	3.14
	WIDA-ELL Advanced	13.37	3.13
	WIDA-Between	13.37	3.13

NOTE: ITT represents intent-to-treatment effects, and LATE represents local average treatment effects.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

From Figure 1.5 and Table 1.3, we observed that the ITT estimates at the cutoff varied depending on the choice of the scaling function and the cutoff value. Specifically, when we used rank-based scale values, the ITT at the cutoff is sensitive to the choice of the cutoff value. In contrast, the ITT based on WIDA-scale values is less sensitive to the choice of the cutoff value. On the other hand, the estimated LATEs were similar across different choices of the scaling function and the cutoff. In other words, our original conclusion about the effectiveness of receiving ETA on math proficiency scores among compliers is insensitive to different specifications of the scaling function and the cutoff value.

Next, we tested the sensitivity of our results against unmeasured confounding. We evaluated (i) whether our conclusion about the non-significance of the ITT would change to a significant positive effect if an important unmeasured binary confounder U, that introduced *negative* confounding bias, were present and (ii) whether our conclusion about the significant, positive LATE would change to insignificance if an unmeasured binary confounder U, that introduced *positive* confounding, were present.

We estimated a new adjusted treatment effects $\tau^{\dagger}$ for the ITT and LATE based on the formula from VanderWeele and Arah (2011): $d = \delta\gamma = \tau - \tau^{\dagger}$. Consider the causal graph in Figure 1.1, where **U** is a set of unobserved confounders. Here, we use a single independent and binary confounder U to represent the **U**-induced confounding bias after conditioning on the observed covariates **W**. Then, δ represents the magnitude of the causally reverse effect of A on U (i.e., $A \rightarrow U$), and γ represents the causal effect of the confounder U on Y ($U \rightarrow Y$). Thus, the bias induced by this hypothetical confounder is $d = \delta\gamma$. Consequently, the bias-adjusted effect is $\tau^{\dagger} = \tau - d = \tau - \delta\gamma$. To implement the sensitivity analysis, we first need to find reasonable values for δ and γ. We determined δ by looking at the absolute mean differences in the

measured covariates **W** between the neighboring "ELL Advanced" and "Formerly ELL" groups. The largest difference we observed was $\delta = 0.07$. For γ, we used the absolute value of the largest estimated coefficient of the measured covariates in the outcome model and set $\gamma = 17.16$. For the ITT estimate, we used a negative bias of $d = -0.07 \times 17.16 = -1.20$ and computed the new adjusted effect $\tau^{\dagger} = 2.66 + 1.20 = 3.86$ with its corresponding 95% confidence interval (-0.40, 8.13), using the estimated standard error of 2.18. Since the confidence interval contains 0, we reach the same conclusion as before: there is no significant ITT effect. Thus, the ITT estimate is insensitive to unmeasured confounding (at least with regard to the magnitude of observed confounding effects). For the LATE estimate, we used a positive bias and obtained a bias-adjusted effect of $\tau^{\dagger} = 12.99 - 1.20 = 11.79$ and an adjusted 95% confidence interval of (5.64, 17.94), using the estimated standard error of 3.14. Since the confidence interval does not cover 0, our conclusion about a positive LATE does not change and is robust to unmeasured confounding.

Conclusions

In this paper, we proposed to use an RD design with an ordinal discrete running variable. We used a scale function S to convert the ordinal levels of the running variable to numeric scale values and modified Lee and Card (2008)'s framework to accommodate the ordinal running variable. We assessed the sensitivity of our results with respect to the choice of the scaling function and the cutoff value. We also assessed the sensitivity of our results to unmeasured confounding arising from an incorrect scaling function or an imperfect functional form. We demonstrated the proposed approach by investigating the effects of ETA on students' math performance based on the 2017 NAEP data. Overall, we found no evidence concerning the effect

of being eligible for ETA at the cutoff (i.e., ITT at the cutoff), but evidence for the effect of receiving ETA among compliers at the cutoff (i.e., LATE at the cutoff). Our ITT estimate is sensitive to the choice of the cutoff value. In contrast, the LATE estimate is robust to the choice of the scaling function, the choice of the cutoff value, and unmeasured confounding.

Based on our findings, we provide some suggestions for future research concerning evaluation of testing accommodations based on an RD framework. First, we didn't incorporate multiple plausible values, sampling weight, and jackknife replicate weights in the NAEP data, and future research would thoroughly consider the NAEP sampling design to generalize the results from an RD Design to the target population. Second, if the NAEP assessment provides continuous ELL English proficiency that was used to determine ETA eligibility, it will enable researchers to estimate the ITT and LATE at the cutoff with less concerns about biases arising from the choice of the scaling function or the functional form of the outcome model. Third, if an ordinal running variable is present and relevant underlying variable, such as pre-test English scores in our setting, are included in the observed data, researchers could use optimal scaling techniques to choose appropriate scale values. Lastly, we did not consider whether students made use of ETA in our study. That is, even if the student received ETA, students may have not needed the extra allotted time. Based on prior works on accommodations, about 40% of the students who received ETA made use of the extra time in the NAEP assessment (Kim & Circi, 2018, 2019). Students' actual use of ETA can be determined by process data, which are data provided by examinees' responses to the testing devices while taking the test (Bergner & von Davier, 2018). If such data is available, we may have to use sequential compliance models and it would be interesting to incorporate them into RD designs in order to assess the effect of making use of ETA.

STUDY 2 : RANDOM FORESTS APPROACH FOR CAUSAL INFERENCE WITH CLUSTERED OBSERVATIONAL DATA

Abstract

There is a growing interest in using machine learning (ML) methods for causal inference due to their (nearly) automatic and flexible ability to model key quantities such as the propensity score or the outcome model. Unfortunately, most ML methods for causal inference have been studied under single-level settings where all individuals are independent of each other and there is little work in using these methods with clustered or nested data, a common setting in education studies. This paper investigates using one particular ML method based on random forests known as Causal Forests to estimate treatment effects in multilevel observational data. We conduct simulation studies under different types of multilevel data, including two-level, three-level, and cross-classified data. Our simulation study shows that when the ML method is supplemented with estimated propensity scores from multilevel models that account for clustered/hierarchical structure, the modified ML method outperforms pre-existing methods in a wide variety of settings. We conclude by estimating the effect of private math lessons in the Trends in International Mathematics and Science Study data, a large-scale educational assessment where students are nested within schools.

Suk, Y., Kang, H., & Kim, J.-S. (2020). Random forests approach for causal inference with clustered observational data, *Multivariate Behavioral Research*. doi:10.1080/00273171.2020.1808437.

Introduction

In the social sciences, when studying treatment effects with observational data, study units are naturally clustered together. For example, in the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA), students, the study unit, are clustered/nested at the school level, and schools are clustered/nested at the country level; this type of data is known as three-level data or, more broadly, multilevel/hierarchical data. In cross-classified data, students belong to two clusters simultaneously, say student's cluster membership is defined by school and neighborhood (Raudenbush & Bryk, 2002). The goal of this paper is to estimate the average treatment effect (ATE) in multilevel observational data where the treatment is assigned at the unit level (e.g., students instead of schools or neighborhoods).

Traditionally, the most popular way to estimate the ATE in clustered observational data is by multilevel propensity score methods (Hong & Raudenbush, 2006; Kim & Seltzer, 2007; Steiner, Kim, & Thoemmes, 2012; Thoemmes & West, 2011). Briefly, a propensity score is a study unit's conditional probability of receiving treatment given observed pre-treatment covariates (Rosenbaum & Rubin, 1983) and each study unit has a propensity score ranging from 0 to 1. Propensity score methods utilize propensity scores to balance covariates between treated and control groups by using matching, stratification, or weighting. Multilevel propensity score methods use propensity scores, but also account for the underlying clustered/hierarchical structure in multilevel data, typically by fitting a multilevel propensity score model; see Section 10 of Leite (2016) for details. If selection is strongly ignorable, i.e., the treatment assignment is as-if random conditional on observed pre-treatment covariates, and the propensity score model is correctly specified, these aforementioned methods can consistently estimate the ATE. Also,

propensity score methods can be combined with models for the outcome to form doubly robust estimators; these estimators are consistent for the ATE as long as either the propensity score model or the outcome model is correctly specified.

Recently, using machine learning (ML) algorithms has become a popular way to estimate the ATE (Athey & Imbens, 2016; Chernozhukov et al., 2018; Hill, 2011; Keller, Kim, & Steiner, 2015; McCaffrey, Ridgeway, & Morral, 2004; van der Laan, Polley, & Hubbard, 2007; van der Laan & Rose, 2011; Westreich, Lessler, & Funk, 2010). An attractive feature of ML-based methods is that they flexibly and, in some cases, automatically estimate the propensity score model or the outcome regression model. However, many of these works focus on what we call single-level data settings where all study units are (i) independent of each other and (ii) come from the same population; in short, the study units are assumed to be independent and identically distributed (i.i.d.). If the i.i.d. assumption is satisfied, many ML-based methods are consistent for the ATE. In contrast, if study units are clustered or nested, such as the students in the TIMSS data, the i.i.d. assumption no longer holds and there is no guarantee that these ML-based methods produce a consistent estimate for the ATE. Indeed, as highlighted by Carvalho, Feller, Murray, Woody, and Yeager (2019), existing ML methods should be modified, say by using different tuning parameters or re-designing sampling splits, to respect the underlying clustering or hierarchical structure and to produce a more precise and consistent estimate of the treatment effect in multilevel data. However, it remains an open question as to how to exactly make such modifications for different kinds of ML methods.

The goal of this paper is to study how to modify ML-based methods in order to estimate the ATE in multilevel observational data. We focus on one ML method based on random forests called Causal Forests (Athey, Tibshirani, & Wager, 2019; Wager & Athey, 2018) which is a

popular ML-based method in causal inference. We consider three types of multilevel data: two-level, three-level, and cross-classified data. These settings, in general, violate the underlying assumptions that validate Causal Forests as a consistent and asymptotically Normal estimator of the ATE because of dependencies between subjects.

We study three simple ways to modify Causal Forests in order to account for the underlying hierarchical structure. The first modification injects Causal Forests with propensity scores from a multilevel logistic regression. The second modification uses cluster labels that denote each level of the hierarchy and changes how sample splitting is done inside Causal Forests. The third modification combines the two modifications. We compare how well these modified ML methods perform compared to ML methods without any modifications, traditional multilevel propensity score methods, and doubly robust methods where the latter two use parametric multilevel models by measuring the absolute relative bias, standard deviation, and mean-squared error. Lastly, we demonstrate our findings by studying the effect of private math lessons on students' math achievement scores from the 2015 Korea TIMSS data.

Overall, we found that in a wide range of scenarios, the modified Causal Forests using an estimated propensity score from a multilevel logistic regression was competitive to doubly robust estimators with correctly specified propensity score and outcome regression models. Also, the modified Causal Forests had smaller mean-squared error than traditional multilevel propensity score methods or the original Causal Forests without any modifications. This phenomenon generally held true even when the multilevel propensity score model in the modified Causal Forest was moderately mis-specified. More broadly, we believe that our modifications of Causal Forests based on the multilevel propensity score can serve as a template to modify other ML-

based methods in causal inference when they are used in observational studies with hierarchical/clustered structures.

Causal Inference with Clustered Observational Data

We use the Neyman-Rubin potential outcomes notation (Neyman, 1923; Rubin, 1974) and its extension to multilevel/clustered settings by Hong and Raudenbush (2006) to formalize causal effects. Suppose that there are N total individuals, indexed by ij where i indexes study units within a cluster and j indexes clusters. Let Z_{ij} denote the treatment assignment of individual i in cluster j, with $Z_{ij} = 1$ representing a treated individual and $Z_{ij} = 0$ representing an untreated individual; as noted earlier, this paper focuses on studies where the treatment is assigned at the individual level, not at the cluster level. Let $Y_{ij}(1)$ be the potential outcome if individual i in cluster j were to be treated ($Z_{ij} = 1$) and let $Y_{ij}(0)$ be the potential outcome if individual i within cluster j were to be untreated ($Z_{ij} = 0$). The observed outcome is $Y_{ij} = Z_{ij}Y_{ij}(1) + (1 - Z_{ij})Y_{ij}(0)$. We remark that the notation assumes Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1986); for more information on SUTVA in multilevel settings, see Hong and Raudenbush (2006, 2013). Finally, let $\mathbf{X}_{ij}$ and $\mathbf{W}_j$ denote individual-level and cluster-level pre-treatment covariates, respectively. Individual-level covariates $\mathbf{X}_{ij}$ are covariates whose values vary across individuals, say gender, age, and socioeconomic status, and cluster-level covariates $\mathbf{W}_j$ are covariates whose values are the same among individuals in the same cluster, say school type and school climate. Typically, cluster-level covariates define the underlying clustering or hierarchical structure in multilevel data.

The target estimand of interest is the ATE and is defined as the average linear contrast between two potential outcomes, i.e., $\tau = E[Y_{ij}(1) - Y_{ij}(0)]$. In a completely randomized

experiment, the ATE can be estimated by taking the difference between the outcomes of treated and untreated groups. However, in observational studies, this approach may lead to bias because the treated and untreated groups are no longer similar with respect to their observed and unobserved covariates.

The typical set of working assumptions for estimating the ATE in observational studies is as follows:

Assumption 1 (Unconfoundness): $Y_{ij}(1), Y_{ij}(0) \perp Z_{ij} | \mathbf{X}_{ij}, \mathbf{W}_j$

Assumption 2 (Positivity): $0 < e(\mathbf{X}_{ij}, \mathbf{W}_j) = Pr(Z_{ij} = 1 | \mathbf{X}_{ij}, \mathbf{W}_j) < 1$

Here, $\perp$ represents independence between two random variables. In a nutshell, Assumptions 1 and 2 state that conditional on the observed individual-level and cluster-level covariates $\mathbf{X}_{ij}$ and $\mathbf{W}_j$, the treatment is as-if randomly assigned to each individual with non-zero probability. Assumptions 1 and 2 combined are often referred to as *strong ignorability* (Rosenbaum & Rubin, 1983). An important idea arising from Assumptions 1 and 2 is that the propensity score $e(\mathbf{X}_{ij}, \mathbf{W}_j)$ is the coarsest balancing score (Rosenbaum & Rubin, 1983). Specifically, conditional on the propensity score, the distribution of pre-treatment covariates is independent of treatment assignment, i.e., $Y_{ij}(1), Y_{ij}(0) \perp Z_{ij} | e(\mathbf{X}_{ij}, \mathbf{W}_j)$, and the distribution of covariates is identical or "balanced" between the treated and untreated groups. This latter idea serves as the basis for propensity score methods where treated and untreated units are matched, weighed, or stratified based on their propensity scores. The next section reviews some of these propensity score methods for multilevel data.

Methods of Estimating the Average Treatment Effect in Multilevel Data

Multilevel Propensity Score Methods via Weighting

Among various propensity score methods in multilevel data, weighting methods are one of the most frequently used in practice. Briefly, weighting methods use propensity scores as sampling weights to weigh the outcomes between treated and untreated groups in order to unbiasedly estimate the ATE (Lunceford & Davidian, 2004; Schafer & Kang, 2008; Stuart, 2010). The most well-known weighting method is the inverse-propensity weighting (IPW) estimator, which is the weighted difference in the outcomes Y_{ij} of treated and untreated groups where the weights are the inverse of the estimated propensity scores:

$$\begin{aligned} \omega_{z,ij}^{IPW} &= \frac{z}{e\left(\mathbf{X}_{ij}, \mathbf{W}_j\right)} + \frac{1-z}{1-e\left(\mathbf{X}_{ij}, \mathbf{W}_j\right)} \\ \hat{\tau}_{IPW} &= \frac{1}{N}\sum_{ij} Y_{ij} Z_{ij}\, \omega_{1,ij}^{IPW} - \frac{1}{N}\sum_{ij} Y_{ij}(1-Z_{ij})\, \omega_{0,ij}^{IPW} \end{aligned} \tag{2.1}$$

Hong and Hong (2009) proposed another weighting estimator called the marginal mean weighting through stratification (MMW-S). MMW-S estimator, like the IPW estimator, takes the weighted difference in outcome between treated and untreated groups. But MMW-S stratifies the propensity score into percentiles and uses sample frequencies within each stratum as weights. A bit more formally, consider the observed frequencies O_{zs} for each treatment status $z \in \{0,1\}$ in stratum $s \in \{1,2,\ldots,S\}$ defined by percentiles of $e\left(\mathbf{X}_{ij}, \mathbf{W}_j\right)$ on the logit scale (or probability scale). Define the expected frequencies E_{zs} for each treatment status z in stratum s to be $E_{zs} = O_{(z\cdot)} \times O_{(\cdot S)}/O_{(\cdot\cdot)}$; the dots in the subscript $(\cdot)$ denote sums over z or s. Then, MMW-S weights are computed by dividing the expected frequencies by the observed frequencies, E_{zs}/O_{zs}, and the MMW-S estimator of the ATE is:

$$\omega_{z,ij}^{MMW-S} = \begin{cases} \frac{E_{z1}}{O_{z1}} & \text{if } e\left(\mathbf{X}_{ij}, \mathbf{W}_j\right) \text{ in stratum 1} \\ \vdots & \\ \frac{E_{zS}}{O_{zS}} & \text{if } e\left(\mathbf{X}_{ij}, \mathbf{W}_j\right) \text{ in stratum } S \end{cases} \tag{2.2}$$

$$\hat{\tau}_{MMW-S} = \frac{1}{N}\sum_{ij} Y_{ij} Z_{ij}\, \omega_{1,ij}^{\text{MMW-S}} - \frac{1}{N}\sum_{ij} Y_{ij}(1 - Z_{ij})\, \omega_{0,ij}^{\text{MMW-S}}$$

The use of MMW-S requires researchers to choose the number of strata *S*. Hong (2010) suggested choosing *S* to be the minimal number of strata that achieves within-stratum covariate balance in at least 95% of the observed covariates. We follow this advice in our simulation and real data studies and use covariates' standardized mean and variance within strata as our measure of within-stratum balance.

A key consideration in using weighting estimators in multilevel data is a model for the propensity score $e\left(\mathbf{X}_{ij}, \mathbf{W}_j\right)$. Broadly speaking, in two-level data, there are two main strategies for modeling the propensity score and estimating the ATE: a within-cluster strategy and an across-cluster strategy (Leite, 2016; Steiner et al., 2012). A within-cluster strategy estimates a propensity score model for each cluster with only individual-level covariates. Each propensity score model is trained using only study units within the cluster and is usually a logistic regression commonly found in single-level i.i.d. settings. Then, using one of the weighting estimators above, cluster-level treatment effects are computed and then aggregated to estimate the ATE. In contrast, an across-cluster strategy estimates a single joint propensity score model across all clusters. The joint propensity score model is typically a multilevel logistic regression with either random-effects or fixed-effects; random-effects models usually have both individual-level and cluster-level covariates, whereas fixed-effects models have individual-level covariates and cluster-level dummy variables (Leite, 2016; Steiner et al., 2012).

Generally speaking, weighting estimators based on a propensity score model from a within-cluster strategy is more robust to biases from unmeasured cluster-level covariates than those from an across-cluster strategy; one can make a joint propensity score more robust to unobserved cluster-level covariates by adding interaction terms between cluster labels and covariates. However, when there are strong selection processes at work or cluster sizes are small, overlap is hard to achieve with propensity score estimates based on a within-cluster strategy, and in such settings, an across-cluster strategy is preferred (Kim & Seltzer, 2007; Steiner et al., 2012; Thoemmes & West, 2011). Given that the size of clusters in our real data is not large and a within-cluster strategy is often not feasible in these settings (Arpino & Mealli, 2011; Steiner et al., 2012), this paper uses an across-cluster strategy where a single joint propensity score model is estimated using a multilevel logistic regression model with random effects.

Finally, we remark that there are other types of multilevel propensity score methods, such as two-stage matching (Rickles & Seltzer, 2014), preferential matching (Arpino & Cannas, 2016), and within-class matching (Kim & Steiner, 2015), and most multilevel propensity score methods are tailored for two-level data.

Doubly Robust Methods

Doubly robust (DR) methods are estimators that provide consistent estimates of the ATE as long as the propensity score or the outcome model is correctly specified, but not necessarily both (Schafer & Kang, 2008; Scharfstein, Rotnitzky, & Robins, 1999). Here, we present a DR estimator based on fitting weighted outcome regression models for treated and untreated units, i.e., $E\left[Y_{ij}\middle|\mathbf{X}_{ij},\mathbf{W}_j,Z_{ij}=z\right]=m_z\left(\mathbf{X}_{ij},\mathbf{W}_j,\boldsymbol{\beta}_z\right)$ where $\boldsymbol{\beta}_z$ is the parameter of the outcome regression function m_z for treatment group $z\in\{0,1\}$, where the weights in the regression are

from the IPW estimator or MMW-S estimator (Leite, 2016; Schafer & Kang, 2008). Specifically, the DR estimator based on weights from the IPW estimator is:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_z^{IPW} &= argmin_{\boldsymbol{\beta}_z} \sum_{ij,Z_{ij}=z} \omega_{z,ij}^{IPW}\left[Y_{ij} - m_z\left(\mathbf{X}_{ij}, \mathbf{W}_j, \boldsymbol{\beta}_z\right)\right]^2 \\ \hat{\tau}_{DR\ IPW} &= \frac{1}{N}\sum_{ij} m_1\left(\mathbf{X}_{ij}, \mathbf{W}_j, \widehat{\boldsymbol{\beta}}_1^{IPW}\right) - \frac{1}{N}\sum_{ij} m_0\left(\mathbf{X}_{ij}, \mathbf{W}_j, \widehat{\boldsymbol{\beta}}_0^{IPW}\right)\end{aligned} \tag{2.3}$$

and the DR estimator based on weights from the MMW-S estimator is:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_z^{MMW-S} &= argmin_{\boldsymbol{\beta}_z} \sum_{ij,Z_{ij}=z} \omega_{z,ij}^{MMW-S}\left[Y_{ij} - m_z\left(\mathbf{X}_{ij}, \mathbf{W}_j, \boldsymbol{\beta}_z\right)\right]^2 \\ \hat{\tau}_{DR\ MMW-S} &= \frac{1}{N}\sum_{ij} m_1\left(\mathbf{X}_{ij}, \mathbf{W}_j, \widehat{\boldsymbol{\beta}}_1^{MMW-S}\right) - \frac{1}{N}\sum_{ij} m_0\left(\mathbf{X}_{ij}, \mathbf{W}_j, \widehat{\boldsymbol{\beta}}_0^{MMW-S}\right)\end{aligned} \tag{2.4}$$

In multilevel settings, typical models for m_z are random-effects or fixed-effects linear regression models. In this paper, we will use random-effects linear regression models for m_z.

Causal Forests and Modifications for Multilevel Data

Recently, there is a growing trend in using flexible, non-parametric ML algorithms to estimate e or m_z without having to specify the functional form of these models. For example, targeted maximum likelihood estimators of the ATE (van der Laan & Rose, 2011) often utilize an ensemble learner called SuperLearner (van der Laan et al., 2007), which combines different ML algorithms such as the Lasso, K-nearest matching, generalized additive models (GAMs), generalized linear models (GLMs), random forests, and multivariate adaptive regression splines (MARS), to flexibly estimate the propensity score and the outcome model. Causal Forests is another popular ML method for estimating the ATE as well as the conditional average treatment effect (CATE); we remark that averaging across unbiased estimates of the CATE leads to an

unbiased estimator for the ATE. In this paper, we focus on Causal Forests and see how it performs in multilevel observational data.

We briefly sketch out the details of the Causal Forest algorithm; see Wager and Athey (2018), Athey et al. (2019), and Athey and Wager (2019) for details. Let $m(\mathbf{x}, \mathbf{w}) = E\left[Y_{ij} \middle| \mathbf{X}_{ij} = \mathbf{x}, \mathbf{W}_j = \mathbf{w}\right]$ be the conditional mean of the outcome given a specific value of covariates $\mathbf{x}, \mathbf{w}$. Let $\hat{m}^{-ij}(\mathbf{x}, \mathbf{w})$ be an estimate of this conditional mean at $\boldsymbol{x}$ and $\boldsymbol{w}$ where the estimate does not use study unit ij's data. Similarly, let $\hat{e}^{-ij}(\mathbf{x}, \mathbf{w})$ be an estimate of the propensity score where the estimate does not use study unit ij's data; these type of estimates are also known as out-of-bag leave-one-out estimates in machine learning. In the Causal Forests algorithm, the default estimates of $\hat{m}^{-ij}(\mathbf{x}, \mathbf{w})$ and $\hat{e}^{-ij}(\mathbf{x}, \mathbf{w})$ are computed by an honest random forest algorithm (see Procedure 1 in Wager and Athey (2018) for one example), although a consistent estimator with certain statistical properties will suffice; see Section 6.1.1 of Athey et al. (2019) for the exact conditions. For example, if the propensity score is known as in a completely randomized experiment, we can plug in the probability of treatment from the experimental design as $\hat{e}^{-ij}(\mathbf{x}, \mathbf{w})$ and this satisfies the aforementioned statistical properties. A Causal Forest estimator of the CATE at covariates $\mathbf{x}, \mathbf{w}$, denoted as $\hat{\tau}(\mathbf{x}, \mathbf{w})$, is essentially a weighted linear regression of residualized outcome $\tilde{Y}_{ij} = Y_{ij} - \hat{m}^{-ij}(\mathbf{x}, \mathbf{w})$ and a single residualized regressor $\tilde{Z}_{ij} = Z_{ij} - \hat{e}^{-ij}(\mathbf{x}, \mathbf{w})$ and can be written as follows:

$$\hat{\tau}(\mathbf{x}, \mathbf{w}) = \frac{\sum_{ij} \alpha_{ij}(\mathbf{x}, \mathbf{w}) \left(Y_{ij} - \hat{m}^{-ij}(\mathbf{X}_{ij}, \mathbf{W}_j)\right)\left(Z_{ij} - \hat{e}^{-ij}(\mathbf{X}_{ij}, \mathbf{W}_j)\right)}{\sum_{ij} \alpha_{ij}(\mathbf{x}, \mathbf{w}) \left(Z_{ij} - \hat{e}^{-ij}(\mathbf{X}_{ij}, \mathbf{W}_j)\right)^2} \tag{2.4}$$

Here, $0 \leq \alpha_{ij}(\mathbf{x}, \mathbf{w}) \leq 1$ weighs each data point ij's contribution to $\hat{\tau}(\mathbf{x}, \mathbf{w})$ based on how far ij's covariates, i.e., $\mathbf{X}_{ij}, \mathbf{W}_j$, are away from $\mathbf{x}, \mathbf{w}$. Roughly speaking, a high $\alpha_{ij}(\mathbf{x}, \mathbf{w})$ generally

indicates that data point ij contributes a large portion to $\hat{\tau}(\mathbf{x}, \mathbf{w})$, whereas a small $\alpha_{ij}(\mathbf{x}, \mathbf{w})$ generally indicates that data point ij contributes a small portion to $\hat{\tau}(\mathbf{x}, \mathbf{w})$.

Computing the weights $\alpha_{ij}(\mathbf{x}, \mathbf{w})$ is based on a modified recursive partitioning algorithm with $\tilde{Y}_{ij}, \tilde{Z}_{ij}, \mathbf{X}_{ij}, \mathbf{W}_j$. Specifically, suppose we obtain *B* bootstrap replicates and let $b = 1, \ldots, B$ be one of those replicates. For each bth replicate, take a random subsample without replacement of size *s* from *n* total samples in the bth replicate. The algorithm randomly partitions the *s* subsampled data into two equally-sized data, say subsets $\mathcal{J}_1$ and $\mathcal{J}_2$. Using $\mathcal{J}_1$, the algorithm initializes a "parent" node $\mathcal{P}$, which represents all the data in $\mathcal{J}_1$, and computes pseudo-outcomes ρ_{ij} for data in $\mathcal{P}$

$$\rho_{ij} = D_{\mathcal{P}}^{-1}\left(\tilde{Z}_{ij} - \tilde{Z}_{\mathcal{P}}\right)\left(\tilde{Y}_{ij} - \tilde{Y}_{\mathcal{P}} - \left(\tilde{Z}_{ij} - \tilde{Z}_{\mathcal{P}}\right)\hat{\beta}_{\mathcal{P}}\right), \qquad D_{\mathcal{P}}^{-1} = \frac{1}{I(ij \in \mathcal{P})}\sum_{ij \in \mathcal{P}}\left(\tilde{Z}_{ij} - \tilde{Z}_{\mathcal{P}}\right)^2$$

Here, $\tilde{Y}_{\mathcal{P}}$ and $\tilde{Z}_{\mathcal{P}}$ represent averages of $\tilde{Y}_{ij}$ and $\tilde{Z}_{ij}$, respectively, among data in the parent node $\mathcal{P}$ and $\hat{\beta}_{\mathcal{P}}$ represents a linear regression between $\tilde{Y}_{ij}$ and $\tilde{Z}_{ij}$ among data points in $\mathcal{P}$. It then uses standard Classification and Regression Trees (CART) (Breiman, Friedman, Olshen, & Stone, 1984) with the psuedo-outcomes ρ_{ij} and covariates $\mathbf{X}_{ij}, \mathbf{W}_j$ to find a partition of the parent node $\mathcal{P}$ into two non-overlapping children nodes $\mathcal{C}_1$ and $\mathcal{C}_2$ such that the following argument is maximized

$$argmax_{\mathcal{C}_1, \mathcal{C}_2} \frac{\sum_{ij \in \mathcal{C}_1} \rho_{ij}^2}{\sum_{ij \in \mathcal{P}} I(ij \in \mathcal{C}_1)} + \frac{\sum_{ij \in \mathcal{C}_2} \rho_{ij}^2}{\sum_{ij \in \mathcal{P}} I(ij \in \mathcal{C}_2)}$$

Roughly speaking, the argmax above finds partitions $\mathcal{C}_1$ and $\mathcal{C}_2$ of $\mathcal{P}$ such that the variance of the pseudo-outcomes within each partition is maximized. For example, if one of the covariates is sex

(male or female), $\mathcal{C}_1$ may be study units who are female and $\mathcal{C}_2$ may be study units who are male. It may choose these partitions based on sex if it maximizes the equation above. Once it finds a partition $\mathcal{C}_1$, $\mathcal{C}_2$, it relabels $\mathcal{C}_1$ as the parent $\mathcal{P}$ and repeats the procedure; it also does the same for the other child node $\mathcal{C}_2$. The procedure stops after it reaches a stopping criterion and the final output is a partition of the covariate space that is represented as a binary tree. It then uses the other subset $\mathcal{J}_2$ to count the proportion of data in $\mathcal{J}_2$ that fall inside each of the terminal leaf nodes in the binary tree. This process is repeated across $b = 1, \ldots, B$ bootstrap replicates, constructing *B* trees. Given a set of covariates $\mathbf{x}, \mathbf{w}$, the algorithm evaluates $\alpha_{ij,b}(\mathbf{x}, \mathbf{w})$, which is equal to one of aforementioned proportions if $\mathbf{x}, \mathbf{w}$ belongs to the same leaf node as data point ij and 0 otherwise, for each bootstrap replicate. Finally, it computes the weights $\alpha_{ij}(\mathbf{x}, \mathbf{w})$ as $\alpha_{ij}(\mathbf{x}, \mathbf{w}) = \sum_{b=1}^{B} \alpha_{ij,b}(\mathbf{x}, \mathbf{w})/B$. We remark that we can use the above recursive partitioning algorithm with bootstrap replicates to estimate the leave-one-out estimates $\hat{e}^{-ij}(\boldsymbol{x}, \boldsymbol{w})$ and $\hat{m}^{-ij}(\boldsymbol{x}, \boldsymbol{w})$. For the interested reader, the entire procedure is implemented in the *grf* package in R (R Core Team, 2020); see Tibshirani et al. (2019) for a detailed vignette.

Compared to parametric approaches in "Doubly Robust Methods", Causal Forests is a nonparametric estimator that, under some assumptions, achieves consistency and asymptotic convergence to a pivotal Gaussian distribution. The latter is crucial as it allows for construction of *p*-values and confidence intervals. Unfortunately, these statistical properties only hold if the data is generated in an i.i.d. fashion. However, Causal Forests can be fine-tuned by changing various parameters such as cluster labels, the minimum node size of each tree, and a penalty for imbalanced splits, and we will modify these parameters to improve its performance in clustered data. Specifically, we study three modifications to Causal Forests:

1. The first modification forces Causal Forests to use multilevel propensity scores discussed in "Multilevel Propensity Score Methods via Weighting" by the tuning parameter *W.hat* in the *grf* package. In other words, the estimator $\hat{e}^{-ij}(\boldsymbol{x}, \boldsymbol{w})$ in (5) is based on a multilevel logistic regression model with random effect terms instead of the default regression forests estimator based on the aforementioned recursive partitioning algorithm. This modification allows Causal Forests to recognize clustering/hierarchical structure through the multilevel propensity score and adjusts its estimate of $\hat{\tau}(\boldsymbol{x}, \boldsymbol{w})$ accordingly. The modification of this type is denoted as (Est.PS) in subsequent sections.

2. The second modification follows a suggestion from Athey and Wager (2019) where we add cluster label information to define clusters and let Causal Forests "figure out" the rest of the clustering structure from the raw cluster labels; in software, this is achieved by using the parameter *clusters* in the *grf* package. The modification forces Causal Forests to draw random subsample of clusters (instead of individuals) and then run the algorithm above within the subsampled clusters. With this modification, the out-of-bag samples are defined as samples that are not in the random subsample of clusters drawn to train the tree. The modification of this type is denoted as (ID) in subsequent sections.

3. The third modification combines the above two modifications and is denoted as (Est.PS+ID) in subsequent sections.

The next sections evaluate these modifications to Causal Forests through an extensive simulation study and a real data analysis.

Simulation Study

We conducted a large-scale simulation study to assess the performance of methods in "Methods of Estimating the Average Treatment Effect in Multilevel Data." Our simulation study can be broadly categorized into two designs and is summarized in Table 2.1. Design 1 assumes a constant treatment effect and generates data from three types of multilevel structures: two-level, three-level, and cross-classified. Design 2 is limited to two-level structures, but has varying treatment effects across clusters and mis-specifies the outcome and selection models to test the robustness of methods to model mis-specification. For both designs, we compare the performance between the default Causal Forests without any modifications, the modified Causal Forests discussed above, and traditional multilevel propensity score and DR methods discussed in "Multilevel Propensity Score Methods via Weighting" and "Doubly Robust Methods."

In particular, under Design 1, the IPW estimator, MMW-S estimator, and two DR estimators use correctly specified propensity score and outcome models and represent an ideal scenario whereby a careful investigator was nearly or completely successful in modeling. Based on asymptotic theory, the DR estimators should perform best under Design 1. However, under Design 2, these four estimators use mis-specified propensity score and outcome models with varying degrees of the model mis-specification and represent a more realistic scenario whereby a careful investigator, despite his/her best efforts, was partially successful in modeling. Under Design 2, nonparametric methods like Causal Forests have the potential to show more promise in achieving better performance than traditional methods since ML methods flexibly and (nearly) automatically capture local structure and model a wider array of functional forms, including those that are used in traditional DR and non-DR methods.

Table 2.1
A summary of simulation designs 1 and 2

	Design 1	Design 2
Clustering structure	two-level three-level cross-classified	two-level
Treatment effects	constant	cluster-specific
Model complexity	main effects	main effects and interactions
Model specification	correct	incorrect

For all simulation designs, we repeated the simulation 1000 times. We evaluated the performance of each estimator by measuring the absolute relative bias (%), standard deviation (SD), and mean squared error (MSE) defined as follows.

$$\text{Bias}(\%) = 100 \times \left| \frac{1}{1000} \sum_{m=1}^{1000} \frac{\hat{\tau}_m - \tau}{\tau} \right|$$

$$\text{SD} = \sqrt{\frac{1}{1000-1} \sum_{m=1}^{1000} \left(\hat{\tau}_m - \bar{\hat{\tau}}\right)^2}$$

$$MSE = \frac{1}{1000} \sum_{m=1}^{1000} (\hat{\tau}_m - \tau)^2$$

Here, $\hat{\tau}_m$, m=1,…,1000 is the m-th estimate of the ATE from 1000 simulations. Computer code for the simulation study is available in the supplemental materials and can also be found at the first author's github repository[3].

3 https://github.com/youmisuk/multilevelCF

Design 1 with Two-level Data

The data generating model for two-level clustered data is stated below. The specific parameter values in the model were based on our empirical data from TIMSS.

1. For each cluster $j = 1, \ldots, J$, generate the total number of individuals in each cluster n_j by drawing a number from a normal distribution with mean I and standard deviation sd and rounding it to the nearest integer.

2. For each individual $i = 1, \ldots, n_j$ in cluster j, generate cluster-level and individual-level covariates $\mathbf{W}_j = (W_{1j}, W_{2j})$ and $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij})$ as follows.

$$\begin{pmatrix} W_{1j} \\ W_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 2 & .2 \\ .2 & 2 \end{pmatrix}\right]$$

$$\begin{pmatrix} X_{1ij} \\ X_{2ij} \end{pmatrix} \sim N\left[\begin{pmatrix} 0.1W_{1j} + 0.05W_{2j} + \kappa_{1j} \\ 0.08W_{1j} + 0.1W_{2j} + \kappa_{2j} \end{pmatrix}, \quad \begin{pmatrix} 10 & 2 \\ 2 & 15 \end{pmatrix}\right]$$

$$\begin{pmatrix} \kappa_{1j} \\ \kappa_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}\right]$$

Individual-level covariates are functions of cluster-level covariates W_{1j}, W_{2j} and random errors κ_{1j}, κ_{2j}. This type of model for the covariates reflects studies where cluster-level school characteristics impact students' malleable characteristics. For example, it is plausible that students' classroom behavior, motivation, grit, and/or goal orientations (i.e., individual-level covariates $\mathbf{X}_{ij}$) are affected by location of the school, school funding, schools' vision or climate/culture (i.e., cluster-level covariates $\mathbf{W}_j$). We remark that depending on the study at hand, $\mathbf{X}_{ij}$ and $\mathbf{W}_j$ can have different relationships, such as $\mathbf{X}_{ij}$ and $\mathbf{W}_j$ being independent of each other or $\mathbf{W}_j$ being affected by $\mathbf{X}_{ij}$.

3. Generate individual treatment status Z_{ij} from the following random-effects logistic propensity score model.

$$logit(e_{ij}) = 0 + 0.1X_{1ij} + 0.03X_{2ij} + 0.16W_{1j} + 0.08W_{2j} + R_j, \quad R_j \sim N(0,1)$$

$$Z_{ij} \sim Bernoulli(e_{ij})$$

where e_{ij} is the propensity score for individual i in cluster j, and R_j is the random effect for cluster j with mean of 0 and variance of 1. The intra-class correlation (ICC) is around 0.23.

4. Generate the potential outcomes $Y_{ij}(1)$, $Y_{ij}(0)$ and observed outcome Y_{ij} from a random-effects linear regression model.

$$Y_{ij}(z) = 100 + \tau \cdot z + 2X_{1ij} + 1X_{2ij} + 2W_{1j} + 1.5W_{2j} + U_j + \epsilon_{ij}$$

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

$$U_j \sim N(0,10), \qquad \epsilon_{ij} \sim N(0,100)$$

Here, U_j is the random effect for cluster j with mean of 0 and variance of 10, and ϵ_{ij} is the random error for individual i in cluster j with mean of 0 and variance of 100. Also, the treatment effect is constant and set to $\tau = 2$. The ICC is 0.1.

Figure 2.1 summarizes the results. Each row category, denoted by ($J, I_{(sd)}$), represents three sub-types of two-level data defined by the number of clusters J and the mean size of each cluster I along with its standard deviation sd. For example, the first row category is a two-level dataset with $J = 150$ clusters and each cluster has, on average, 30 individuals with standard deviation of 2. In this condition, we observed that the two DR estimators had the smallest bias and MSE, and the MMW-S estimator had the largest bias and MSE. The performance of the MMW-S estimator was surprising given that it is the de-facto estimator for the ATE in multilevel data. Between Causal Forests and traditional methods, the performance of modified Causal Forests with multilevel propensity scores (CF+Est.PS) generally lied somewhere in between DR estimators and non-DR estimators in terms of bias and MSE. But we observed the modified

Causal Forests using multilevel propensity scores had the smallest MSE across all estimators when the mean size of clusters was greater than the number of clusters. Among Causal Forests with modifications, the modified Causal Forests using multilevel propensity scores had

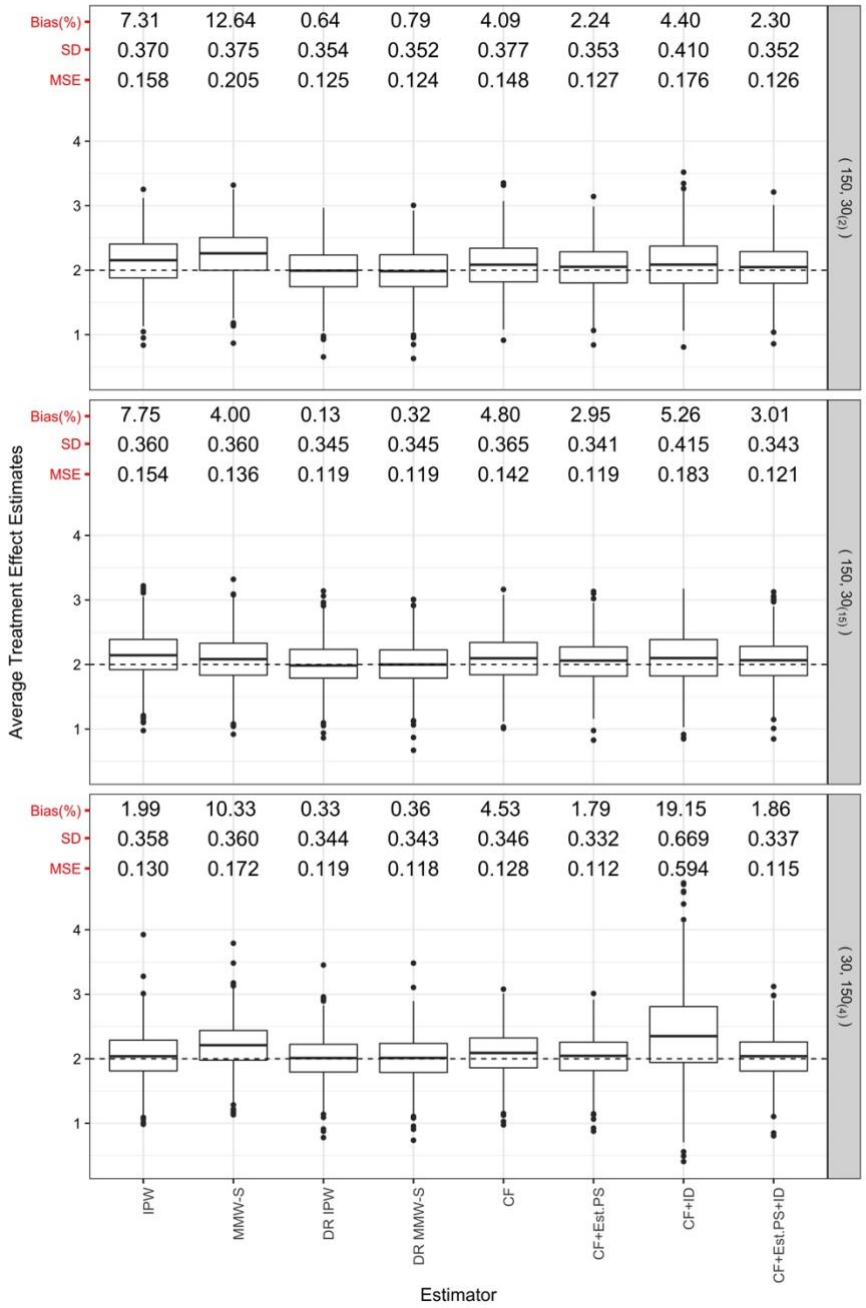

Figure 2.1. Performance of ATE estimates in two-level data under Design 1. $(J, I_{(sd)})$ represent the number of clusters and the mean size of clusters and its standard deviation, respectively. The dashed black line indicates the true average treatment effect value of 2. IPW = inverse-propensity weighting; MMW-S = marginal mean weighting through stratification; DR IPW = doubly robust estimators with IPW weights; DR

MMW-S = doubly robust estimators with MMW-S weights; CF = Causal Forests without modification; CF+Est.PS = CF with propensity scores from a multilevel logistic regression; CF+ID = CF with cluster labels. CF+Est.PS+ID = CF with propensity scores from a multilevel logistic regression and cluster labels.

the smallest bias and variance. Also, providing explicit cluster labels (CF+ID) did not improve the performance of Causal Forests. For instance, when the mean cluster size was larger than the number of clusters, Causal Forests with only cluster labels showed the worst performance across all estimators. However, Causal Forests using both multilevel propensity scores and cluster labels performed similarly to Causal Forests using multilevel propensity scores.

Intrigued by the poor performance of the modified Causal Forests based only on cluster labels, Figure 2.2 explores the properties of Causal Forests when we increase the sample size beyond what is typical in most education studies. For better visualization, Figure 2 trims the top and bottom 10% of estimates from modified Causal Forests using only cluster labels in the (J, $I_{(sd)}$) = (10, $5000_{(4)}$) setting. But the absolute relative bias, SD, and MSE are computed using all 1000 estimates from the simulation. Interestingly, when the number of clusters was 5000 and the size of each cluster was around 10, the default Causal Forests and modified Causal Forests with cluster labels performed best in terms of MSE among all Causal Forests methods. On the other hand, when the number of clusters was 10 and the size of each cluster was around 5000, Causal Forests with estimated propensity scores performed best in terms of bias and MSE among all Causal Forests methods; in fact, Causal Forests with cluster labels performed worse than the default Causal Forests.

The result of Figure 2.2 agrees with some prior theoretical results concerning clustered data and the underlying sampling splitting procedure behind Causal Forests. For example, prior works in econometrics have shown that estimators for cluster-robust standard errors are generally consistent if the number of clusters goes to infinity and the size of the cluster is relatively small

in comparison (Wooldridge, 2010). Additionally, in clustered randomized trials where the treatment is assigned at the cluster level, many methods typically require the number of clusters

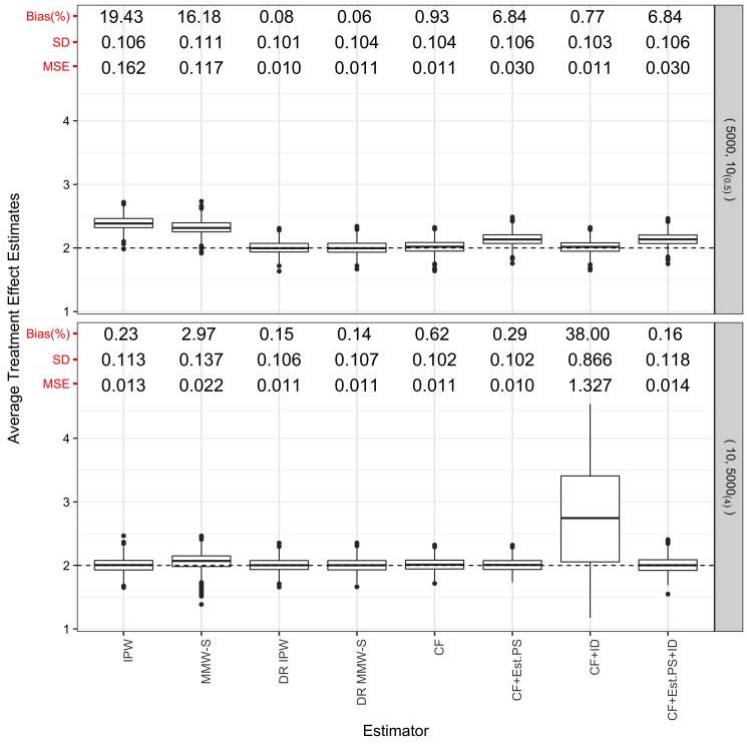

Figure 2.2. Performance of ATE estimates in two-level data: asymptotic properties. $(J, I_{(sd)})$ represent the number of clusters and the mean size of clusters and its standard deviation, respectively. The dashed black line indicates the true average treatment effect value of 2. IPW = inverse-propensity weighting; MMW-S = marginal mean weighting through stratification; DR IPW = doubly robust estimators with IPW weights; DR MMW-S = doubly robust estimators with MMW-S weights; CF = Causal Forests without modification; CF+Est.PS = CF with propensity scores from a multilevel logistic regression; CF+ID = CF with cluster labels. CF+Est.PS+ID = CF with propensity scores from a multilevel logistic regression and cluster labels.

to be much larger than the size of the clusters to achieve consistent and asymptotically Normal estimates of the ATE; see Donner and Klar (2010), Hayes and Moulton (2009), Kang and Keele (2018) and references therein. Finally, as mentioned in "Causal Forests and Modifications for Multilevel Data", when Causal Forests is provided with cluster labels, the underlying sample splitting procedure is done at the cluster level. This means that for a small number of clusters, there will be fewer clusters to train the tree, potentially leading to large biases. In contrast, for a large number of clusters, there will be many clusters to train the tree, leading to smaller biases.

Combining these insights with Figure 2.2, in two-level data, we believe that providing cluster labels to Causal Forests is only useful in a setting where the number of clusters is much larger than the size of clusters.

Design 1 with Three-level Data

Three-level data were generated in a similar way as two-level data, but with an additional set of two-dimensional continuous covariates at the highest hierarchy. Specifically, we varied the number of highest-level clusters K and the number of intermediate-level clusters where the latter was determined by drawing a number from a normal distribution with mean J and standard deviation sd and rounding it to the nearest integer. Also, the propensity score and outcome models for three-level data were:

$$\begin{aligned} logit(e_{ijk}) &= -0.2 + 0.1X_{1ijk} + 0.03X_{2ijk} + 0.1W_{1jk} + 0.08W_{2jk} + 0.1Q_{1k} + 0.05Q_{2k} \\ &\quad + R_{jk}^{W} + R_{k}^{Q}, \qquad R_{jk}^{W} \sim N(0,1), \qquad R_{k}^{Q} \sim N(0,1) \\ Y_{ijk}(z) &= 100 + 2 \cdot z + 2X_{1ijk} + 1X_{2ijk} + 2W_{1jk} + 1.5W_{2jk} + 1Q_{1k} + 0.5Q_{2k} + U_{jk}^{W} \\ &\quad + U_{k}^{Q} + \epsilon_{ijk}, \qquad U_{jk}^{W} \sim N(0,10), \qquad U_{k}^{Q} \sim N(0,7), \qquad \epsilon_{ijk} \sim N(0,100) \end{aligned}$$

We highlight three additional differences between the three-level simulation design and the two-level simulation design. First, there are two more continuous covariates, Q_{1k}, Q_{2k}, at the highest level of the hierarchy. Second, there are two random effect terms in the propensity score model (i.e., R_{jk}^{W} and R_{k}^{Q}) and in the outcome model (i.e., U_{jk}^{W} and U_{k}^{Q}). Random effect terms follow a Normal distribution with mean zero and variance σ^2. The additional random effect term in each model corresponds to the additional clustering effect at the highest level of the hierarchy. The ICCs for R_{jk}^{W} and R_{k}^{Q} in the propensity score model are around 0.19 and 0.19, respectively, and the ICCs for U_{jk}^{W} and U_{k}^{Q} in the outcome model are 0.1 and 0.07, respectively. Third, we provide

either intermediate-level or highest-level cluster labels when we modify Causal Forests.

Additional details on generating three-level data are in Appendix D.

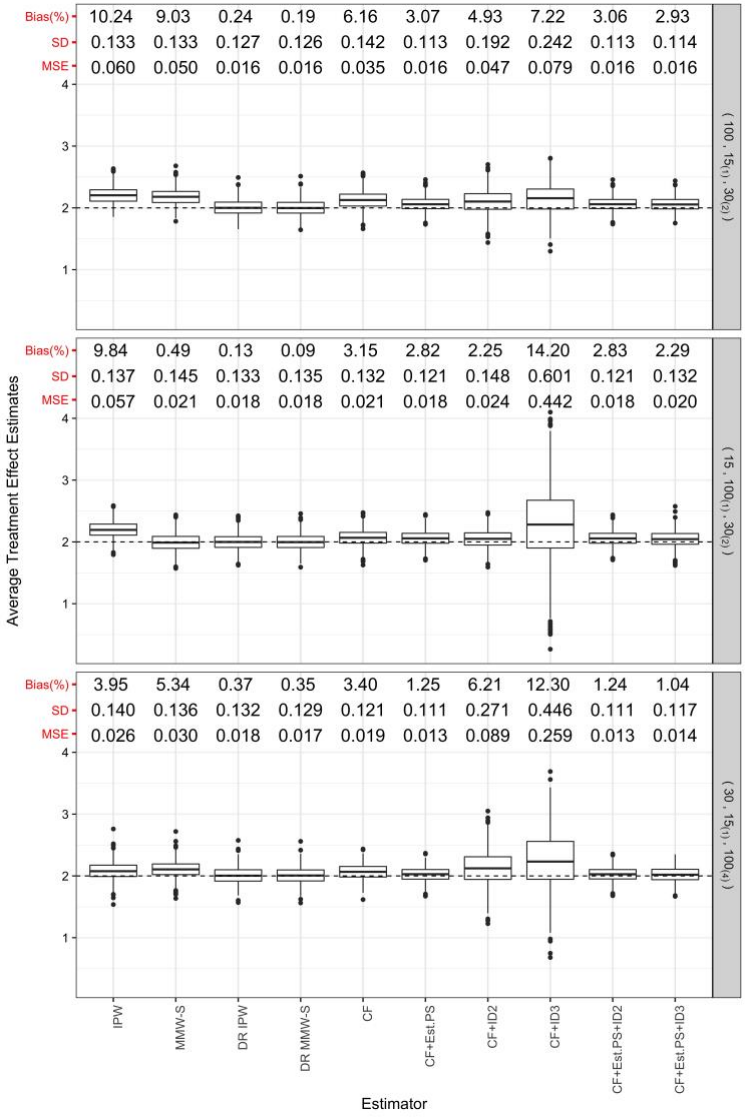

Figure 2.3. Performance of ATE estimates in three-level data. ($K, J_{(sd)}, I_{(sd)}$) represent the number of clusters and the mean size of clusters and its standard deviation, respectively. The dashed black line indicates the true average treatment effect value of 2. IPW = inverse-propensity weighting; MMW-S = marginal mean weighting through stratification; DR IPW = doubly robust estimators with IPW weights; DR MMW-S = doubly robust estimators with MMW-S weights; CF = Causal Forests without modification; CF+Est.PS = CF with propensity scores from a multilevel logistic regression; CF+ID2 = CF with level-2 cluster labels; CF+ID3 = CF with level-3 cluster labels; CF+Est.PS+ID2 = CF with propensity scores from a multilevel logistic regression and level-2 cluster labels; CF+Est.PS+ID3 = CF with propensity scores from a multilevel logistic regression and level-3 cluster labels.

Figure 2.3 shows the simulation results for three-level data. Similar to the result from two-level data, the two DR estimators had the smallest bias. However, unlike the results from two-level data, modified Causal Forests with multilevel propensity scores was able to achieve the smallest MSE among all estimators; their MSEs were equal to or slightly lower than the MSEs of the DR estimators. The bias of modified Causal Forests with multilevel propensity scores was in between the bias of the non-DR estimators and the DR estimators; the notable exception is the MMW-S estimator with 15 three-level clusters where its bias was lower than modified Causal Forests with multilevel propensity scores.

Also, similar to the result from two-level data, among different modifications to Causal Forests, we found that including the multilevel propensity scores had the largest improvement in terms of bias and variance compared to including only cluster labels into Causal Forests or using the default Causal Forests. Among Causal Forests estimators with only cluster labels, including the intermediate-level cluster labels (CF+ID2) provided more bias and variance reduction than including the highest-level cluster labels (CF+ID3). However, as seen from the two-level setting, using only cluster labels did not provide significant benefits compared to only using multilevel propensity scores, and there were no additional gains in performance from including the cluster labels into modified Causal Forests already using multilevel propensity scores.

Design 1 with Cross-classified Data

Our data generating model for cross-classified data follows closely to Meyers and Beretvas (2006) where we assume that cross-factor residuals are uncorrelated and there are three feeders from one factor to the other factor. The two factors' numbers, denoted as F1 and F2,

varied simultaneously. The exact propensity score and outcome models in cross-classified data were:

$$logit\left(e_{i(jk)}\right) = -0.2 + 0.1X_{1i(jk)} + 0.03X_{2i(jk)} + 0.1W_{1j} + 0.08W_{2j} + 0.1Q_{1k} + 0.05Q_{2k} + R_j^W + R_k^Q, \quad R_j^W \sim N(0,1), \quad R_k^Q \sim N(0,0.5)$$

$$Y_{i(jk)}(z) = 100 + 2 \cdot z + 2X_{1i(jk)} + 1X_{2i(jk)} + 2W_{1j} + 1.5W_{2j} + 1Q_{1k} + 0.5Q_{2k} + U_j^W + U_k^Q + \epsilon_{i(jk)}, \quad U_j^W \sim N(0,10), \quad U_k^Q \sim N(0,7), \quad \epsilon_{i(jk)} \sim N(0,100)$$

Compared to two-level data, in cross-classified data, there are two more continuous covariates to indicate an additional cluster level, Q_{1k}, Q_{2k}. Also, both the propensity score and the outcome models have two random effect terms for the two factors (i.e., R_j^W, R_k^Q and U_j^W, U_k^Q). Random effect terms follow a Normal distribution with mean zero and variance σ^2. The ICCs for R_j^W and R_k^Q in the propensity score model are around 0.21 and 0.10, respectively, and the ICCs for U_j^W and U_k^Q in the outcome model are 0.1 and 0.07, respectively. Finally, for modified Causal Forests based on cluster labels, we use either the first, second, or combined factor labels. Additional details on generating cross-classified data are in Appendix E.

Figure 2.4 provides the results under cross-classified data setting. While the two DR estimators had the smallest bias in most scenarios, the modified Causal Forests with multilevel propensity scores had the smallest bias and MSE when the cluster size (here, factor 1 clusters) was larger than the number of the clusters. In the opposite setting where the size of the clusters was smaller than the number of clusters, the bias and MSE of modified Causal Forests with multilevel propensity scores was somewhere in between those from DR and non-DR estimators. The bias of modified Causal Forests with only cluster labels was comparable to or sometimes worse than that of the IPW estimator or MMW-S estimator. Among modified Causal Forests using cluster labels, Causal Forests using combined cluster identifiers (CF+F12ID) had the

smallest MSE than those using only one factor identifier (CF+F1ID or CF+F2ID). Similar to two-level and three-level settings, having both estimated propensity scores and cluster IDs in Causal Forests provided no additional benefits compared to having only multilevel propensity scores.

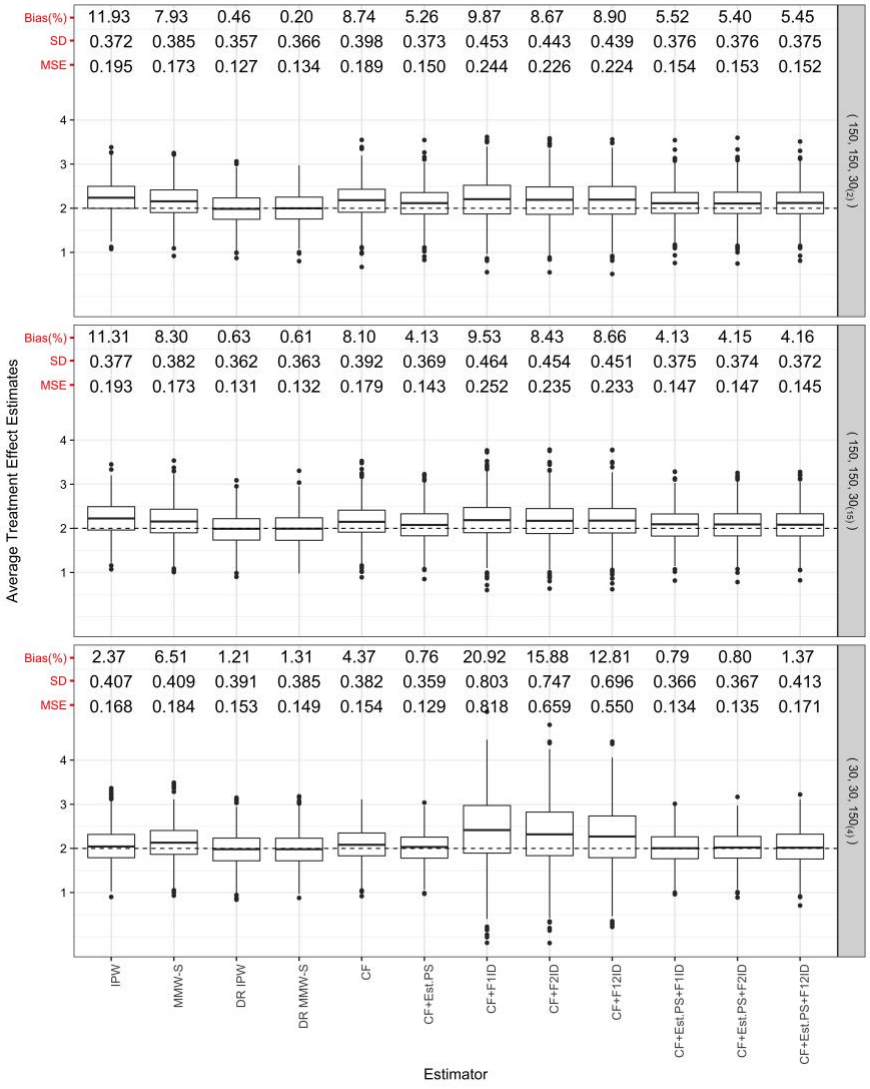

Figure 2.4. Performance of ATE estimates in cross-classified data. (*F1, F2,* $I_{(sd)}$) represent the number of clusters and the mean size of clusters and its standard deviation, respectively. The dashed black line indicates the true average treatment effect value of 2. IPW = inverse-propensity weighting; MMW-S = marginal mean weighting through stratification; DR IPW = doubly robust estimators with IPW weights; DR MMW-S = doubly robust estimators with MMW-S weights; CF = Causal Forests without modification; CF+Est.PS = CF with propensity scores from a multilevel logistic regression; CF+F1ID, CF+F2ID, CF+F12ID = CF with the first factor labels, the second factor labels, or the two factors' combined labels; CF+Est.PS+F1ID, CF+Est.PS+F2ID, CF+Est.PS +F12ID = CF with propensity scores from a multilevel logistic regression and the first factor labels, the second factor labels, or the two factors' combined labels.

Design 2 with Two-level Data

In Design 2, we generated two-level data similar to Design 1, except we changed the propensity score and the outcome model as follows:

$$logit(e_{ij}) = 0 + 0.1X_{1ij} + 0.03X_{2ij} + 0.16W_{1j} + 0.08W_{2j} + \beta_1 X_{1ij} W_{2j} + R_j, \quad R_j \sim N(0,1)$$

$$Y_{ij}(z) = 100 + z\,(\tau + 0.5W_{1j}) + 2X_{1ij} + 1X_{2ij} + 2W_{1j} + 1.5W_{2j} + \beta_2 X_{1ij} W_{2j} + U_j + \epsilon_{ij}$$

$$U_j \sim N(0,10), \qquad \epsilon_{ij} \sim N(0,100), \qquad \boldsymbol{\beta} = (\beta_1, \beta_2) \in \{(0.02, 0.3), (0.04, 0.6), (0.06, 1)\}$$

There are two major differences between the previous two-level simulation design and the new two-level simulation design. First, there is now an interaction term between the individual-level covariate X_{1ij} and the cluster-level covariate W_{2j} in both the propensity score and outcome models. The interaction term's magnitude is controlled by parameter $\boldsymbol{\beta}$. We use this extra interaction term to intentionally mis-specify the propensity score and the outcome model by only fitting the main effects; note that this type of mis-specification can also be seen as a form of omitted variable bias where we "omitted" the interaction terms. Second, in the outcome model, there is an interaction term between the treatment *z* and the cluster-level covariate W_{1j}. This interaction term creates heterogeneous cluster-specific treatment effects where the cluster-specific CATE is $\tau + 0.5W_{1j}$. However, the mean of W_{1j} is zero and hence, the overall ATE remains the same as before, $\tau = 2$.

Figure 2.5 summarizes the results with different values of $\boldsymbol{\beta}$. As the propensity score and outcome models became more mis-specified, the DR estimators behaved similarly to non-DR estimators and both were equally biased. When the model mis-specification was moderate to large, Causal Forests methods almost always performed better in terms of bias and MSE than any of the traditional methods (IPW, MMW-S, and DR estimators) across different conditions on the

number of clusters and the mean size of clusters. Even the default Causal Forests that did not incorporate any clustering information did better than traditional methods that incorporated

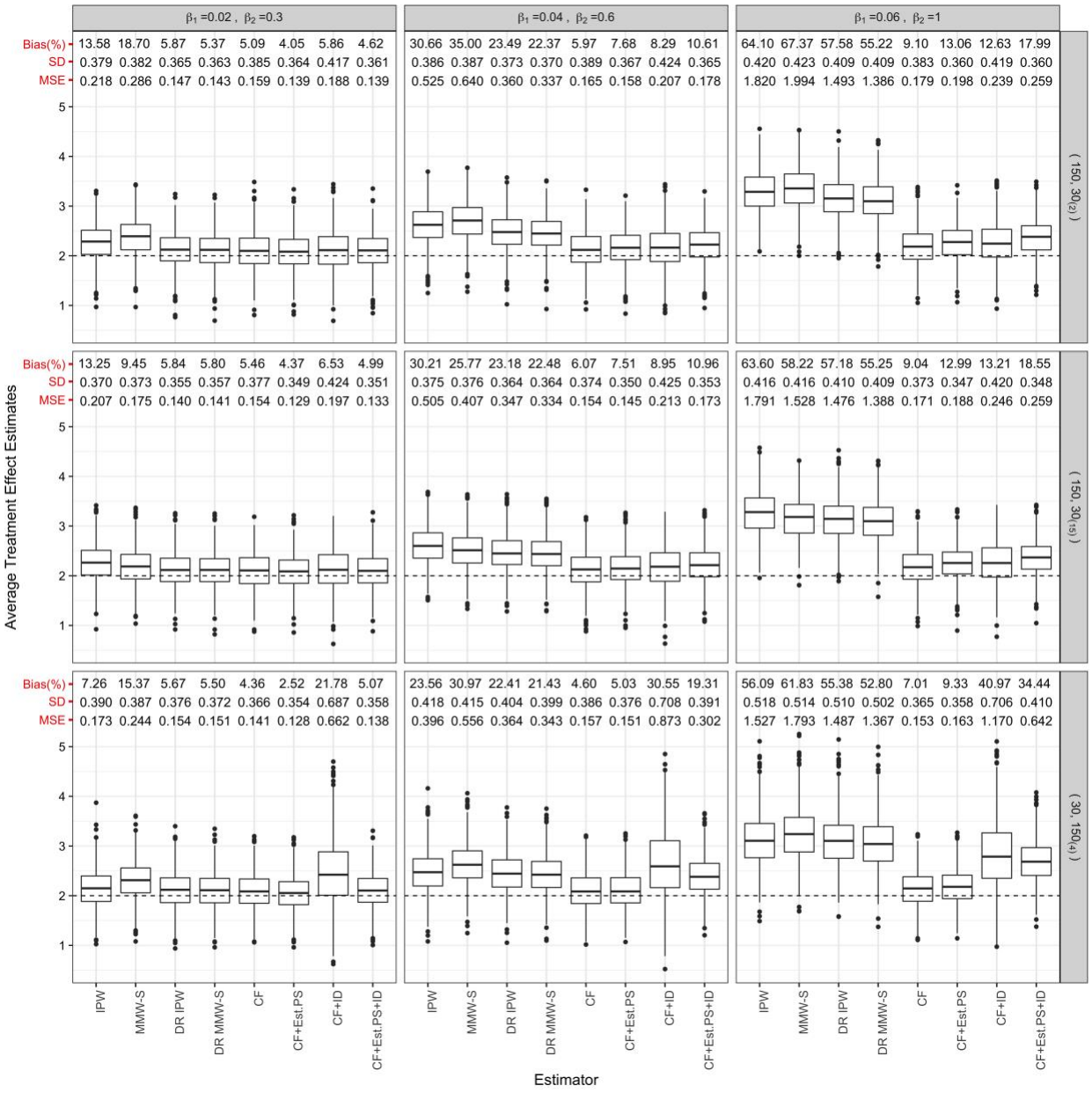

Figure 2.5. Performance of ATE estimates in two-level data under Design 2. β_1 and β_2 are coefficients of the omitted interaction terms in parametric selection and outcome models, respectively. (J, I(sd)) represent the number of clusters and the mean size of clusters and its standard deviation, respectively. IPW represents inverse-propensity weighting and MMW-S represents marginal mean weighting through stratification. DR IPW and DR MMW-S represent doubly robust estimators with IPW and MMW-S weights, respectively. CF represents Causal Forests without modification. CF+Est.PS represents CF with propensity scores from a multilevel logistic regression. CF+ID represents CF with cluster labels. CF+Est.PS+ID represents CF with propensity scores from a multilevel logistic regression and cluster labels. The dashed black line indicates the true average treatment effect value of 2.

clustering. This suggests that when parametric models are mis-specified, nonparametric methods like Causal Forests perform better than traditional parametric methods, even if the ML methods were not initially designed for clustered data, because they can automatically capture the omitted interaction terms that potentially lead to more significant bias compared to bias arising from ignoring clustering structure inside ML methods. Also, the modified Causal Forests with multilevel propensity scores outperformed the default Causal Forests in terms of variance across all settings, and they did better than the default Causal Forests in terms of MSE when the model mis-specification was small to moderate. Finally, Causal Forests using only cluster labels generally performed worse than Causal Forests using multilevel propensity scores.

Takeaways from Simulation Studies

Overall, the simulation studies above suggest some guidelines on how to modify ML methods for clustered/hierarchical data. While these are not meant to encompass every ML method in causal inference or every possible type of clustered data in practice, we hope that the suggestions provided below based our simulation study with Causal Forests can serve as useful guidelines for empirical or theoretical analyses of clustered data with ML methods.

1. For Causal Forests, incorporating multilevel, hierarchical structure through multilevel propensity score models typically had the largest improvement in terms of bias and MSE compared to Causal Forests that directly used cluster labels. In particular, Causal Forests using cluster labels should only be reserved to settings when the number of clusters is much larger than the size of clusters.

2. In general, there was no significant benefit in terms of bias and MSE when using both multilevel propensity scores and cluster labels inside Causal Forests compared to just using multilevel propensity scores.

3. Causal Forests with multilevel propensity scores almost always performed better in terms of bias and MSE than multilevel propensity score estimators (IPW and MMW-S) with correctly specified parametric propensity scores, but generally performed worse than DR methods with correctly specified parametric propensity score and outcome models, with some notable exceptions. In particular, Causal Forests with multilevel propensity scores had similar or slightly lower MSE than those from DR methods with correctly specified propensity score and outcome models whenever the size of the clusters is larger than the number of clusters.

4. If DR or non-DR estimators were using parametric propensity score or outcome models that were moderately mis-specified, Causal Forests with multilevel propensity scores outperformed them in terms of bias and MSE.

5. Causal Forests with multilevel propensity scores performed better in terms of MSE than the default Causal Forests without any modifications, even if the multilevel propensity score model inside modified Causal Forests was moderately mis-specified.

Real Data Study

Data and Variables

TIMSS is an international educational study about students' achievement progresses in mathematics and science and is sponsored by the International Association for the Evaluation of Educational Achievement (IEA). Since 1995, TIMSS has collected data among students in

Grades 4 and 8 every four years. To do this, TIMSS uses a two-stage stratified cluster sampling design where in the first stage, each country selects schools based on important demographic variables (e.g., school location and/or school gender type), and in the second stage, each school randomly selects one or more intact classrooms (Martin, Mullis, & Hooper, 2016). The most recent completed wave of TIMSS was in 2015 and conducted across 60 countries.

We used the 2015 Korea TIMSS Grade-8 data to investigate the effect of private math lessons. The data contained 5309 students from 150 middle schools where the school sizes ranged from 6 to 75 students; the mean size of schools was about 30. While the original data resembled a three-level structure based on a student-class-school hierarchy, we found that most of the schools selected one classroom; 130 schools selected one classroom, and 20 schools selected two classrooms. Since there was a near one-to-one correspondence between school and class levels, we analyzed the data as a two-level data where students are nested within schools.

The treatment was whether students received private math lessons, with 1 indicating that the student did and 0 otherwise; the treatment was assigned at the student level. The outcome was the first plausible value of students' math achievement scores; in the 2015 TIMSS data, each student obtained five plausible values because they took a subset of items from a full battery of assessment items. In addition, we used 12 covariates that were thought to influence the treatment and outcome variables. Six of 12 covariates were student-level covariates and the other six covariates were school-level covariates. Student-level covariates included 1) gender (*sexM*, male and female), 2) fathers' highest education level (*dad.edu*, with three levels including no college, college *dad_cll*, and don't know *dad_q*), 3) the number of books at home (*books25*, with two levels defined as more than 25 books or less than or equal to 25 books), 4) the number of home study supports (*hspprt*, with three levels including neither own room nor Internet connection, one

of them *hspprt_1*, and both *hspprt_2*), 5) students' confidence in math (*M.stuconf*, continuous), and 6) value in math (*M.value*, continuous). School-level covariates included 1) whether the school is gendered (*gender.type*, with three levels of all-boys, all-girls *girlsch*, and co-education *coedu*), 2) the percentage of economically disadvantaged students (*pct.disad*, with four levels of 0 to 10%, 11 to 25% *disad_11*, 26 to 50% *disad_26*, and more than 50% *disad_M50*), 3) school location (*city.size*, with four levels of urban *city_U*, suburban *city_Sub*, medium size city *city_M*, and small town), 4) emphasis on academic success (*acad.emph*, continuous), 5) math instruction affected by resource shortage (*M.resshort*, continuous), and 6) discipline problems (*dscpn*, continuous).

We excluded students whose responses were inconsistent with the following two questions regarding their participation in private math lessons: (Q1) whether students received private math lessons and (Q2) for how many months they received these lessons. In particular, we excluded students who answered "Yes" to (Q1) and "did not attend" to (Q2) and students who answered "No" to (Q1) and answered something other than "did not attend" to (Q2). Additionally, we excluded students who were missing 7 out of 12 covariates. The final sample consisted of 4943 students (93.1% of the original sample) from 149 schools. In the final sample, the outcome's mean was 606.08 and its standard deviation was 84.19; its minimum and maximum were 306.66 and 859.86, respectively. Data analyzed in this study is included in the supplementary materials and can also be found at the first author's github repository.

Methods

We followed our simulation study and estimated a joint propensity score model based on random-effects logistic regression with main effect terms. The outcome model (for DR methods)

was based on random-effects linear regression model with main effect terms. We also implemented Causal Forests as mentioned in "Design 1 with Two-level Data." We used *grf* package to run Causal Forests. For the IPW, MMW-S, DR IPW, and DR MMW-S estimators, we used *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) to estimate propensity scores and outcome regression models. Standard errors were estimated using cluster bootstrap sampling with 5000 replicates. All analyses were conducted in R. We evaluated covariate balance between treated and untreated units before estimating the ATE. As a rule of thumb, a good balance for a covariate is when its absolute standardized mean difference is smaller than 0.1 and its variance ratio is in between 4/5 and 5/4 (Rubin, 2001; Shadish, Clark, & Steiner, 2008; Steiner, Cook, Shadish, & Clark, 2010). For simplicity, we ignored sampling weights that weighed each individual in the data. Since we did not incorporate sampling weights and multiple plausible values, our proposed analysis plan does not generalize to the study population outside of TIMSS.

Results

Figure 2.6 provides covariate balance plots before and after propensity score adjustments. Before adjustment, there were imbalances in most of the 12 covariates: *dad.edu*, *books25, hspprt, M.stuconf, M.value, pct.disad, city.size,* and *acad.emph*. After IPW or MMW-S adjustment, we improved balance between private-lessons takers (the treated) and non-takers (the untreated). Weights from MMW-S with 3 strata achieved more successful balance than weights from the IPW estimator; weights from MMW-S achieved almost perfect mean and variance balance, whereas weights from the IPW estimator still left two covariates slightly imbalanced.

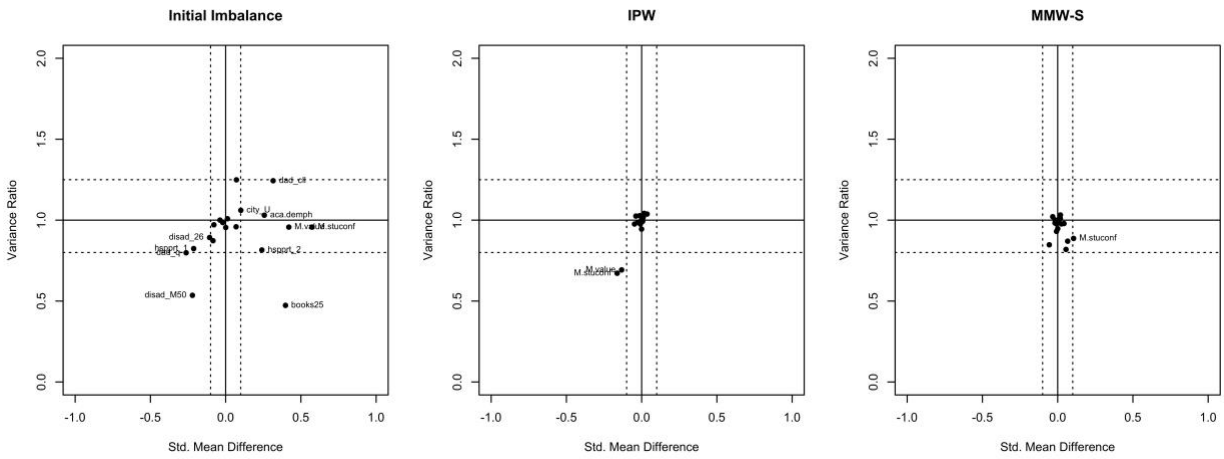

Figure 2.6. Covariate balance plots before and after propensity score adjustment.

Figure 2.7 summarizes the estimates, standard errors (in parentheses), and 95% confidence intervals of the ATE using different estimators. The prima facie effect is the unadjusted mean difference in students' math achievement scores between private-lessons takers and non-takers. The unadjusted estimate was 66.88 points and the effect was reduced to 27.67 points and 33.88 points using the IPW estimator and MMW-S estimator, respectively. The DR IPW and DR MMW-S estimators produced estimates of 32.96 and 29.58 points, respectively. The ATE estimates obtained with Causal Forests ranged from 25.54 (CF+Est.PS) to 30.30 (CF+ID). Between Causal Forests, the IPW estimator, and MMW-S estimator, we found that the point estimates of the ATE from Causal Forests using multilevel propensity scores (CF+Est.PS or CF+Est.PS+ID) were slightly smaller compared to the IPW estimator and MMW-S. Regarding the 95% confidence intervals of the ATE estimates, we found that all the confidence intervals were overlapping. The IPW estimator yielded the widest confidence interval (i.e., the most conservative standard error), while Causal Forests using multilevel propensity scores (CF+Est.PS) produced the narrowest confidence interval (i.e., the most liberal standard error).

Overall, all estimates reached similar conclusions about the treatment effect, that private math lessons had a positive effect on math achievement scores.

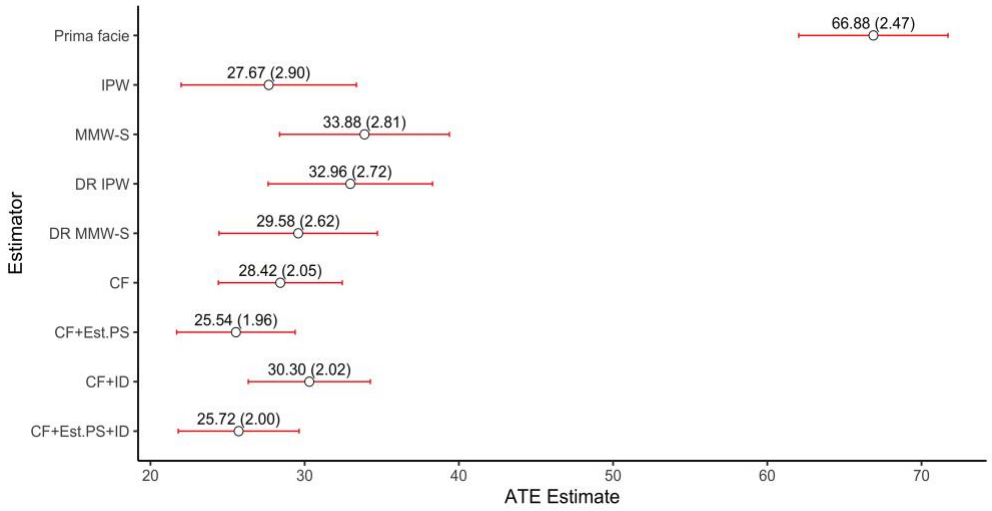

Figure 2.7. The estimates, standard errors (in parentheses), and 95% confidence intervals of the ATE of taking private math lessons.

Discussion and Conclusions

The goal of this paper was to study how to properly modify ML methods originally designed for i.i.d. data to estimate treatment effects in clustered/multilevel data settings. In particular, we explored how to account for cluster or hierarchical structures in Causal Forests by including propensity scores from multilevel logistic regression or cluster labels. As stated in "Takeaways from Simulation Studies", our simulation studies showed that among all modifications to Causal Forests, using Causal Forests with multilevel propensity scores had the best performance in terms of bias and MSE, and providing cluster labels was only useful in settings where the number of clusters is far larger than the size of the clusters. Also, modified Causal Forests with multilevel propensity scores performed somewhere in between DR and non-

DR estimators that had correctly specified propensity score and outcome regression models. However, if these parametric models were mis-specified, modified Causal Forests with multilevel propensity scores performed better than the DR and non-DR estimators in our simulation.

We also conducted a real data study to examine the effects of private math lessons from the 2015 TIMSS data. All methods generally revealed that private math lessons improved math scores, though each estimator exhibited slightly different point estimates. Since each method carries different underlying assumptions about the data generating model, differences in these assumptions likely affected the specific estimated values of the treatment effect. Nevertheless, comparing their results allowed us to assess the plausibility of the underlying assumptions and reinforce our evidence for a causal effect. As such, while Causal Forests with multilevel propensity scores is generally recommended when researchers suspect more complex data generating processes or when there is insufficient subject-matter knowledge to justify parametric models, we recommend using both Causal Forests and traditional DR and non-DR approaches to strengthen the causal conclusion from the study.

There are some limitations of the paper. First, we assumed SUTVA in a clustered setting where the treatment, assigned at the individual level, is hypothesized to not have spillover effects through interference. Second, our simulation study is limited to three types of clustered data structures and we did not analyze more complex clustering structures, such as clustering through spatiotemporal processes. Third, though we explored the consequence of omitting an interaction term when all the confounders were measured in Design 2, we did not consider omitted variable bias arising from unmeasured pre-treatment covariates. Fourth, we primarily used estimated propensity scores from random-effects logistic regression models as part of our modification

strategy for Causal Forests. However, it has been shown that fixed effects models with cluster dummies are robust against unmeasured cluster-level variables (Arpino & Mealli, 2011; Wooldridge, 2010) and it would be interesting to examine whether using multilevel propensity scores with fixed effects inside ML methods can provide additional benefits in estimating the ATE. Fifth, since treatments are often multi-valued rather than binary, future work may extend the applicability of our results to multi-valued treatment settings by utilizing the results in Imbens (2000) and Imai and Van Dyk (2004). Sixth, this paper primarily focuses on estimating the ATE, but we believe that the insights from this work can be used to address challenges in optimal treatment assignment (Dimakopoulou, Zhou, Athey, & Imbens, 2017; Kosorok & Moodie, 2015; Li, Lu, & Zhou, 2017).

Despite these limitations, this paper provides a simple set of ways to modify pre-existing ML methods in causal inference in order to estimate the ATE in multilevel settings. Our proposed modification of Causal Forests by using propensity scores from multilevel regression models helped minimize bias and MSE compared to directly providing cluster labels or using Causal Forests without any modifications. More broadly, this type of modification via the propensity score can serve as a template to modify a wide variety of ML methods when used in multilevel observational studies.

STUDY 3 : HYBRIDIZING MACHINE LEARNING METHODS AND FINITE MIXTURE MODELS FOR ESTIMATING HETEROGENEOUS TREATMENT EFFECTS IN LATENT CLASSES

Abstract

There has been increasing interest in exploring heterogeneous treatment effects using machine learning (ML) methods such as Causal Forests, Bayesian Additive Regression Trees (BART), and Targeted Maximum Likelihood Estimation (TMLE). However, there is little work on applying these methods to estimate treatment effects in latent classes defined by well-established finite mixture/latent class models. This paper proposes a hybrid method, a combination of finite mixture modeling and ML methods from causal inference to discover effect heterogeneity in latent classes. Our simulation study reveals that hybrid ML methods produced more precise and accurate estimates of treatment effects in latent classes. We also use hybrid ML methods to estimate the differential effects of private lessons across latent classes from the Trends in International Mathematics and Science Study (TIMSS) data.

Suk, Y., Kim, J.-S., & Kang, H. (2020). Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes *Journal of Educational and Behavioral Statistics*. doi:10.3102/1076998620951983

Introduction

Motivation

There has been a growing interest in causal inference to estimate conditional average treatment effects (CATEs) using machine learning (ML) methods (Athey & Imbens, 2016; Hill, 2011; Imai & Ratkovic, 2013; Künzel, Sekhon, Bickel, & Yu, 2019; Su, Tsai, Wang, Nickerson, & Li, 2009; Wager & Athey, 2018). These methods show great promise in understanding treatment effect heterogeneity based on observable characteristics of the study population. However, in some settings, observable characteristics are thought to emerge from meaningful latent processes. For example, many studies in education and psychology posit the existence of latent classes defined by parameters in latent class/finite mixture models in order to better understand observed student behaviors such as internet and smartphone addiction or teen smoking (Clogg, 1995; McLachlan & Peel, 2000; Mok et al., 2014; Sutfin, Reboussin, McCoy, & Wolfson, 2009). Differences in observed behaviors are hypothesized to arise due to differences in latent classes, and, as such, there is a strong emphasis on understanding the differences between latent classes by examining the parameters of latent class/profile models or latent class regression models; see Magidson and Vermunt (2004) for details. In such cases where latent classes play a vital role in the scientific understanding of observed phenomena, understanding how the effects of a new treatment, program, or policy vary across these latent classes is of great interest.

To provide a concrete example that motivated this work, consider an observational study estimating the effect of taking private science lessons (i.e., treatment) on science test scores (i.e., outcome) among middle school students. Each student's choice to have private tutors is based on a number of observable characteristics, such as their previous grades and the location of their

schools, as well as unobservable, latent characteristics, such as students' motivations, academic resilience[4], or science self-efficacy[5]. For example, some students who are academically resilient may seek private tutors to supplement classroom instruction compared to those who are less resilient. Some students may opt for a private tutor because they are self-motivated, while others may not seek a tutor because they are less motivated. Or, the driving factor for private lessons may be similar for all students in the same school because of deficiencies (or lack thereof) in school resources. Regardless, these characteristics may not be directly observable, but rich latent class models exist in psychology to help us better understand them; see McLachlan and Peel (2000), Kaplan, Kim, and Kim (2009), and Masyn (2013) for examples. More importantly, variation in these latent classes may lead to differential effects of having a private tutor. For instance, a private tutor may be more helpful in raising test scores among students who are academically resilient or self-motivated compared to students who are less resilient or less motivated.

If an investigator uses one of the aforementioned ML-based estimator to study effect heterogeneity of private tutoring, these methods will only reveal variations in treatment effects among observable characteristics of the student; they would not be able to reveal variations in treatment effects among latent classes representing resilience and self-motivation. To better illustrate this point, consider Figure 3.1, where we constructed a hypothetical two-class latent structure and students belong to either one of the two latent classes; say one class represents strong academic resilience, while another class represents weak academic resilience. The average

[4] Academic resilience is defined as "the heightened likelihood of success in school and inother life accomplishments, despite environmental adversities brought about by early traits, conditions, and experiences" (Wang & Gordon, 2012).
[5] Self-efficacy is defined as "people's beliefs in their ability to influence events that affecttheir lives" (Bandura, 2010).

treatment effect of having private tutors in the first latent class (in yellow) is two, whereas the average treatment effect in the second latent class (in green) is zero. When we use Causal Forests (Athey, Tibshirani, & Wager, 2019), an ML-based causal inference method based on random forests, to estimate treatment effects, the Causal Forests masks these two latent classes' treatment effects. In contrast, our hybrid methods, which we explain below, are able to reveal the two latent classes and their respective treatment effects. Specifically, Figure 3.1 uses our hybrid methods based on Causal Forests, which we refer to as Hybrid Causal Forests.

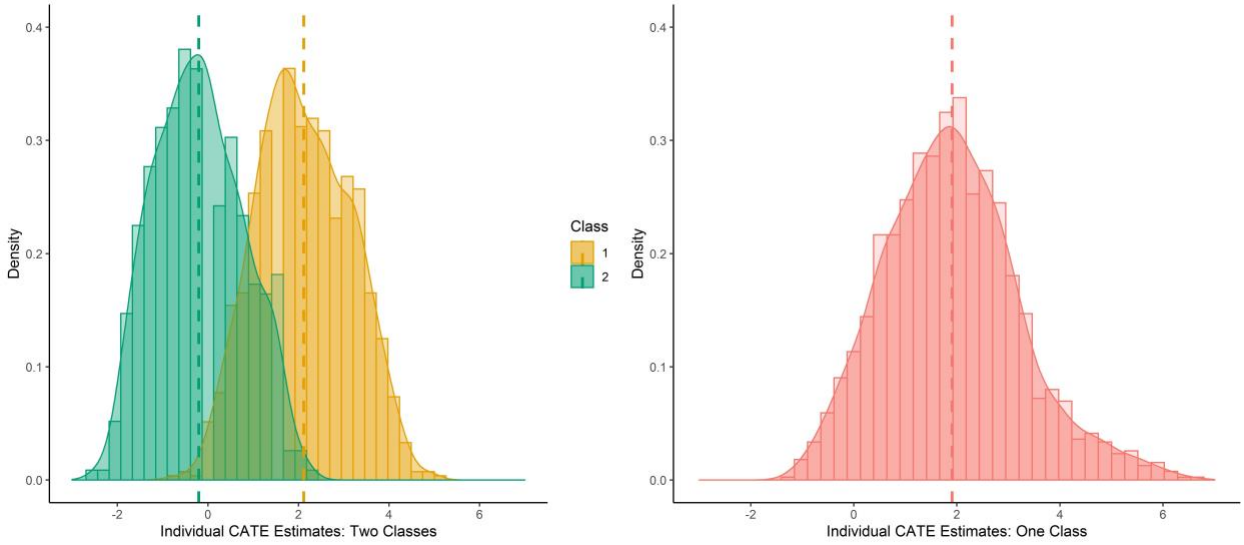

Figure 3.1. Distributions of individual CATE estimates. The left plot shows our method, Hybrid Causal Forests, while the right plot shows Causal Forests. Dashed lines represent class-specific treatment effect estimates.

Prior Work and Our Contribution

Prior works on treatment effect estimation in latent classes are diverse. Kang and Schafer (2010) and Schuler, Leoutsakos, and Stuart (2014) used latent class models to identify latent treatment classes based on manifest/observed items. Butera, Lanza, and Coffman (2014) and Lanza, Coffman, and Xu (2013) discussed estimating treatment effects when the outcome

variables are latent classes. Specifically, the underlying construct of the outcomes was measured using observed items and the goal was to estimate treatment effects on latent class membership of the outcomes. Jo, Wang, and Ialongo (2009) discussed latent trajectory structures in three outcome measures of attention deficit among children and revealed heterogeneity in longitudinal outcomes across latent classes. The work most related to ours is by Kim and Steiner (2015) who used a latent class regression model to model different latent representations of students' selection into treatment (or control) and used multilevel propensity score matching to estimate the treatment effect within each latent class.

The goal of this paper is to complement these prior works and provide a general "hybrid" framework to study treatment effect variation within latent classes by combining latent class modeling with ML-based methods in causal inference. Specifically, in a two-level setting common in education, we propose a two-step hybrid procedure that first uses latent class/finite mixture modeling to identify latent class structures and second, uses modern ML methods in causal inference to estimate treatment effects within each latent class. Our rationale for using ML methods in the second step is to leverage ML's flexibility in modeling potentially complex propensity score and outcome regression models in each latent class. More broadly, our approach to this problem follows a growing trend of combining ML methods with well-established models in psychology to capitalize on the advantages of each approach (Ma, 2018; Suk, Kang, & Kim, 2020). The paper focuses on three popular ML methods in causal inference—Causal Forests (Athey et al., 2019; Wager & Athey, 2018), Bayesian additive regression trees (BART) (Hill, 2011), and Targeted Maximum Likelihood Estimation (TMLE) (Van Der Laan & Rubin, 2006)— but our framework can be extended to other ML methods. We validate our proposed methods through a simulation study and a large-scale educational assessment study concerning

the effect of private science lessons on science achievement scores. We show that our ML-based hybrid methods have more precise and accurate estimates of the variations in treatment effects associated with latent classes than other parametric methods used in this research.

Review: Notation, Causal Assumptions, and the Propensity Score

We use the Neyman-Rubin causal model (Neyman, 1923; Rubin, 1974) and its extension to multilevel data to define causal effects (Hong & Raudenbush, 2006). Let $Y_{ij}(1)$ be the potential outcome if individual i at cluster j were to be treated ($Z_{ij} = 1$). Let $Y_{ij}(0)$ be the potential outcome if individual i at cluster j were to be untreated ($Z_{ij} = 0$). The notation assumes the stable unit treatment value assumption (SUTVA; Rubin, 1986) where the potential outcomes of each individual are not affected by others' treatment assignments and there is only a single version of treatment. This allows us to write the observed outcome Y_{ij} as $Y_{ij} = Z_{ij}Y_{ij}(1) + (1 - Z_{ij})Y_{ij}(0)$. Let $\mathbf{X}_{ij}$ and $\mathbf{W}_j$ denote pre-treatment covariates for individual *i* in cluster *j*, where $\mathbf{X}_{ij}$ are individual-specific covariates and $\mathbf{W}_j$ are cluster-specific covariates.

Under the potential outcomes framework, we assume *strong ignorability*:

$$Y_{ij}(1), Y_{ij}(0) \perp Z_{ij} | \mathbf{X}_{ij}, \mathbf{W}_j \quad \text{and} \quad 0 < e(\mathbf{X}_{ij}, \mathbf{W}_j) = Pr(Z_{ij} = 1 | \mathbf{X}_{ij}, \mathbf{W}_j) < 1$$

where $\perp$ denotes independence between two random variables and $e(\mathbf{X}_{ij}, \mathbf{W}_j)$ is the propensity score (Rosenbaum & Rubin, 1983). In single-level data, propensity scores are typically estimated with logistic regression. In multilevel data, propensity scores are typically estimated with random or fixed effects logistic regression (Leite, 2016). Propensity scores are often used in matching methods to match treated and control units or to weigh individuals' outcomes via inverse probability weighing.

We conclude the section by defining the causal estimand of interest. Let $K_{ij} \in \{1,2,\ldots,C\}$ denote the latent class membership of individual *i* in cluster *j* from a latent class model. The goal in this paper is to estimate the CATE for individuals who belong to latent class $K_{ij} = k$ and is formalized as follows:

$$\tau(k) = E[Y_{ij}(1) - Y_{ij}(0)|K_{ij} = k]$$

The estimand $\tau(k)$ cannot be directly estimated with observed data since latent class membership K_{ij} is unobserved. More precisely, the function that relates the observable characteristics $\mathbf{X}_{ij}, \mathbf{W}_j$ to their latent counterparts K_{ij} is unknown and must be modeled based on context-specific finite mixture/latent class models. In contrast, the usual CATE formalized below

$$\tau(\mathbf{x}, \mathbf{w}) = E[Y_{ij}(1) - Y_{ij}(0)|\mathbf{X}_{ij} = \mathbf{x}, \mathbf{W}_j = \mathbf{w}]$$

can be directly estimated from the observed data under strong ignorability and SUTVA; see Imbens and Rubin (2015) for more details about identification of CATE. Modern ML methods in causal inference (e.g., Causal Forests) provide consistent and asymptotically Normal estimates of $\tau(\mathbf{x}, \mathbf{w})$. However, $\tau(\mathbf{x}, \mathbf{w})$ only reveals treatment heterogeneity among observable characteristics and masks treatment variability in latent classes. The next section discusses our proposed approach, hybrid ML methods, which sequentially integrate latent class modeling and ML methods to estimate $\tau(k)$.

Hybridizing Latent Class Modeling and ML for Causal Inference

Our hybrid approach has two steps. The first step estimates latent classes via context-specific latent class/finite mixture modeling, and the second part uses ML-based methods to estimate treatment effects within each latent class. Subsequent sections elaborate on each step.

Step 1: Latent Class/Finite Mixture Modeling

Latent class or finite mixture models have been frequently used to group individuals or data into unobserved latent classes that can be inferred from the observed data (McLachlan & Peel, 2000; Vermunt & Magidson, 2003). In a standard latent class model, latent classes are identified using categorical latent class indicators e.g., dichotomous survey items, and parameters defining latent classes are response probabilities (Muthén & Muthén, 2017; Wang & Wang, 2012). More generally, there are many types of latent class/finite mixture models based on regression analysis, path analysis, and factor analysis; see Magidson and Vermunt (2004), McLachlan and Peel (2000), Kaplan et al. (2009), and Masyn (2013) for more details. The choice of which latent model to use is context-specific and so long as researchers choose an identifiable finite mixture model to estimate latent classes and each class meets the aforementioned casual assumptions, our methodology will work.

In our real data study of private science tutoring and school achievement scores, we focus on a type of latent class models that describe how students in each school select themselves into treatment (i.e., private science tutoring), also referred to as latent selection/propensity score models, and latent class membership applies at the cluster level (i.e., at the school level) so that $K_{ij} = k$ for everyone in the same cluster. In particular, each cluster belongs to one of $k = 1, \ldots, C$ propensity score models that govern how students within each school select themselves

into treatment, and the parameters of each propensity score define the latent classes; in other words, students' selection behaviors in private tutoring are homogeneous within each cluster, but there are hidden heterogeneous structures across schools that can be inferred from the data. As an illustration of the chosen latent class model, suppose that school principals emphasize academic achievements. In such schools, students may seek private lessons to receive higher achievement scores. Additionally, although a school principal's emphasis on academic achievement can play a role in students seeking private tutors, its importance may differ depending on the location of the school; private education services are more readily available in urban areas than in rural areas. We remark that this model is also called a restricted multiple group latent class model (Vermunt, 2003) and Kim and Steiner (2015) provides additional interpretations as well as some limitations of the model.

The overall goal of the latent class selection model in our real data example is to understand which of the C selection models govern students' choices to select private tutors. Formally, let $\pi_k = P(K_j = k)$, $k = 1, \ldots, C$ be the marginal probability of being in latent class *k*. For this latent class model, we drop the subscript i in K_{ij} for clarity, but we can define $K_{ij} = K_j$ to fit it into the general notation and we use the two notations interchangeably. Consider a latent class random effects logistic regression model where each student's choice to seek treatment (e.g., private tutors) is a mixture of C different selection models. Specifically, each latent class k has its own selection model $e_k(\mathbf{X}_{ij}, \mathbf{W}_j, \theta_k) = P(Z_{ij} = 1|\mathbf{X}_{ij}, \mathbf{W}_j, \theta_k)$ where θ_k parameterizes the model; for two-level data, the selection model for each latent class k is a random effects logistic model. Then, a latent class selection model assumes that $P(Z_{ij} = 1|\mathbf{X}_{ij}, \mathbf{W}_j)$ is C mixtures of selection models with mixing probabilities π_k, $k = 1, \ldots, C$:

$$P\left(Z_{ij}=1\middle|\mathbf{X}_{ij},\mathbf{W}_j\right)=\sum_{k=1}^{C}\pi_k e_k\left(\mathbf{X}_{ij},\mathbf{W}_j,\theta_k\right),\qquad 0\le\pi_k\le 1,\qquad \sum_{k=1}^{C}\pi_k=1 \tag{3.1}$$

where

$$\pi_k=P\left(K_j=k\right)=\frac{exp(\gamma_k)}{\sum_{k=1}^{C}exp(\gamma_k)} \tag{3.2}$$

That is, π_k is modeled by a multinomial logistic model with γ_k representing a class-specific multinomial intercept. The parameters in the models (i.e., θ_k, π_k, $k=1,\ldots,C$) are estimated by an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; Leisch, 2004; McLachlan & Peel, 2000). Using Bayes rule, we can compute the posterior probability that each cluster belongs to latent class k. Specifically, individual ij is assigned to one of C latent classes with the highest posterior probability, also known as modal assignment. Some alternatives to modal assignment are proportional assignment and random assignment (Goodman, 2007; Vermunt, 2010). Here, we use modal assignment due to its simplicity and optimality under certain assumptions about Bayes classification error rates (Bakk, Tekle, & Vermunt, 2013). Regardless, let $\widehat{K}_j$ (or $\widehat{K}_{ij}=\widehat{K}_j$) denote the estimated latent class membership for each cluster. We will use the estimated membership in the second step of our proposed algorithm.

There are some important implementation details in estimating latent class models and we briefly summarize four issues that are most relevant to our setting; see Everitt and Hand (1981), Titterington, Smith, and Makov (1985), and McLachlan and Peel (2000) for detailed discussions. First, typically, the number of latent classes C is initially specified based on subject-matter theories about latent class structure, but later verified by a data-driven approach based on various measures of model fit, such as the likelihood ratio statistic, Pearson's Chi-square, the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) (Kaplan et al.,

2009). While there is always a risk of incorrectly specifying the number of latent classes, either through over-extraction (i.e., more latent classes were specified than the true number of latent classes) or under-extraction (i.e., fewer latent classes were specified than the true number of latent classes), it is generally preferable to have more latent classes than fewer latent classes for estimating treatment effects, as the former would be able to identify diverse structures of latent classes. Second, latent class models are only identifiable up to labeling permutations of latent classes because estimated class labels are arbitrary. For example, if there are three latent classes, the estimated parameters for Class 1 can equally be labeled as Class 2 or Class 3 and the data will be observationally equivalent. This issue primarily affects label assignment, but not estimation of model parameters (Leisch, 2004). Also, many algorithms have been developed to detect label switching issues and to relabel latent classes using ordering constraints and Stephens' methods (e.g., Stephens, 2000; Tueller, Drotar, & Lubke, 2011). Third, in general, mixtures of univariate Normal, gamma, exponential, Cauchy, and Poisson distributions are identifiable, whereas mixtures of uniform distributions are not identifiable. Mixtures of binomial and multinomial distributions can be identified under some assumptions on the number of latent classes and the size of the support of $Z_{ij}, \mathbf{X}_{ij}$, and $\mathbf{W}_j$ (Allman, Matias, Rhodes, et al., 2009; Everitt & Hand, 1981; Grün & Leisch, 2008; Titterington et al., 1985). Finally, to prevent over-fitting the latent selection model, we can set the prior class probabilities to be far away from zero and set θ_k to be sufficiently different (Leisch, 2004). Adding random effects in e_k can also help avoid overfitting (Lenk & DeSarbo, 2000).

For software to implement step 1, we used the software *Mplus8* (Muthén & Muthén, 2017) and the *MplusAutomation* package (Hallquist & Wiley, 2017) in R (R Core Team, 2019)

to estimate the latent class $\widehat{K}_{ij}$. We also applied a "class assignment based algorithm" to resolve potential label switching issues (Tueller et al., 2011).

Step 2: Machine Learning Methods for Causal Inference

This section describes some methods in causal inference that utilize ML to estimate heterogeneous treatment effects. We remind readers that the specific choice of the ML method is not critical so long as they provide consistent point estimators and valid confidence intervals. Also, for one of the ML methods, Causal Forests, we follow a suggestion from recent work (Suk et al., 2020) to improve its performance in two-level data.

General Approach. At a high level, almost all ML-based methods for causal inference require estimating either the outcome model, the propensity score model, or both. Briefly, the outcome model is the conditional expectation of the outcome given the observed covariates and treatment assignment, say $m(\mathbf{x}, \mathbf{w}, z) = E[Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}, \mathbf{W}_j = \mathbf{w}, Z_{ij} = z]$; some methods also estimate the conditional expectation of the outcome given only the covariates, say $m(\mathbf{x}, \mathbf{w}) = E[Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}, \mathbf{W}_j = \mathbf{w}]$. As mentioned above, the propensity score model is the probability of being assigned to treatment given observed covariates, i.e., $e(\mathbf{x}, \mathbf{w}) = P(Z_{ij} = 1|\mathbf{X}_{ij} = \mathbf{x}, \mathbf{W}_j = \mathbf{w})$. Each ML-based method in causal inference estimates the outcome model or the propensity score model using different supervised ML algorithms. For example, BART uses Bayesian regression trees to estimate $m(\cdot)$. TMLE, combined with SuperLearner (van der Laan, Polley, & Hubbard, 2007), uses an ensemble of supervised learning algorithms to estimate $m(\cdot)$ and $e(\cdot)$. Causal Forests uses a modified random forest to estimate $m(\cdot)$ and $e(\cdot)$. Also, each ML-based method aggregates estimates of $m(\cdot)$ and $e(\cdot)$ differently to arrive at the final estimate of CATE. For example, typical BART only uses $m(\cdot)$ to estimate CATE. Causal Forests, which we

describe below in "Vanilla Causal Forests and Modified Causal Forests", uses $m(\cdot)$ and $e(\cdot)$ through a weighted linear regression approach. TMLE uses both $m(\cdot)$ and $e(\cdot)$ through a "clever covariate" to estimate the CATE. If the underlying supervised ML algorithm can consistently estimate $m(\cdot)$ and $e(\cdot)$, these methods not only provide a consistent estimate of the CATE but also, under additional assumptions, provide valid p-values and confidence intervals.

To incorporate ML-based methods into latent class estimation in two-level settings, we outline the following approach. First, for each estimated latent class k, use any of the aforementioned ML-based CATE estimators to estimate the CATE within each k by only using the data from the latent class and denote this as $\tau(\mathbf{x}, \mathbf{w}, k)$; note that if the encompassing ML method requires estimation of $e(\cdot)$, one can use a random effects logistic regression model instead of the associated superivsed learning algorithm to improve performance in clustered/multilevel data. Second, average $\tau(\mathbf{x}, \mathbf{w}, k)$ among individuals with the same k to arrive at the final estimator for $\tau(k)$. We show an example of this general recipe based on Causal Forests below.

Vanilla Causal Forests and Modified Causal Forests. Causal Forests (Athey et al., 2019; Wager & Athey, 2018) is a type of random forests (Breiman, 2001) that is used to estimate the CATE as well as the average treatment effect. Specifically, a Causal Forest estimator of the CATE is a weighted linear regression of residualized outcome $\tilde{Y}_{ij} = Y_{ij} - \hat{m}^{-ij}(\mathbf{x}, \mathbf{w})$ and a single residualized regressor $\tilde{Z}_{ij} = Z_{ij} - \hat{e}^{-ij}(\mathbf{x}, \mathbf{w})$.

$$\hat{\tau}(\mathbf{x}, \mathbf{w}) = \frac{\sum_{ij} \alpha_{ij}(\mathbf{x}, \mathbf{w}) \left(Y_{ij} - \hat{m}^{-ij}(\mathbf{X}_{ij}, \mathbf{W}_j)\right)\left(Z_{ij} - \hat{e}^{-ij}(\mathbf{X}_{ij}, \mathbf{W}_j)\right)}{\sum_{ij} \alpha_{ij}(\mathbf{x}, \mathbf{w}) \left(Z_{ij} - \hat{e}^{-ij}(\mathbf{X}_{ij}, \mathbf{W}_j)\right)^2} \tag{3.3}$$

Here, $0 \leq \alpha_{ij}(\mathbf{x}, \mathbf{w}) \leq 1$ weighs how much each unit ij contributes to the estimate of CATE, $\tau(\mathbf{x}, \mathbf{w})$. The $^{-ij}$-superscript represents out-of-bag leave-one-out estimates in machine learning, i.e., the estimates of functions when unit ij's data is not used. In Causal Forests, the estimates of $\hat{m}^{-ij}(\mathbf{x}, \mathbf{w})$ and $\hat{e}^{-ij}(\mathbf{x}, \mathbf{w})$ are obtained by an honest random forest algorithm in Procedure 1 of Wager and Athey (2018). Wager and Athey (2018) and Athey et al. (2019) showed that the Causal Forests estimator is consistent for the CATE and has an asymptotic pivotal Gaussian distribution under some assumptions; the latter property allows researchers to construct valid p-values and confidence intervals for the CATE.

To estimate class-specific treatment effects in two-level data using Causal Forests, we do the following. First, instead of using a random forest to estimate the propensity score, we use a multilevel logistic regression in step 1 to account for clustering structures inside Causal Forests (Suk et al., 2020). Second, we run Causal Forests among units that are in the same latent class *k*. Combined, the modified CATE estimator using Causal Forests can be formalized as:

$$\hat{\tau}(\mathbf{x}, \mathbf{w}, k) = \frac{\sum_{ij:\hat{K}_{ij}=k} \alpha_{ij}(\mathbf{x}, \mathbf{w}) \left(Y_{ij} - \hat{m}_k^{-ij}(\mathbf{X}_{ij}, \mathbf{W}_j)\right)\left(Z_{ij} - \hat{e}_k(\mathbf{X}_{ij}, \mathbf{W}_j)\right)}{\sum_{ij:\hat{K}_{ij}=k} \alpha_{ij}(\mathbf{x}, \mathbf{w}) \left(Z_{ij} - \hat{e}_k(\mathbf{X}_{ij}, \mathbf{W}_j)\right)^2}$$

Note that $\hat{m}_k^{-ij}$ bears a subscript k to denote that it has been estimated using data from individuals who belong to latent class k. Also, $\hat{e}_k$ no longer has the $^{-ij}$-superscript to denote that it has been estimated using a multilevel logistic regression instead of the default regression forests. Averaging $\hat{\tau}(\mathbf{x}, \mathbf{w}, k)$ across all individuals in the same latent class *k* is our estimate of the average treatment effect within latent class k, i.e.,

$$\hat{\tau}(k) = \frac{1}{N_k} \sum_{ij:\hat{K}_{ij}=k} \hat{\tau}(\mathbf{x}, \mathbf{w}, k)$$

where N_k denotes the sample size in each latent class, k.

Finally, we briefly remark that instead of the proposed approach, an alternative approach to combine latent class estimates with ML methods is to use the estimated latent class as a "covariate" in ML methods; see Appendix F. We show in the Appendix that our approach has better finite sample performance than the alternative approach in terms of bias and mean squared error (MSE).

Simulation Study

Simulation Design and Evaluation

We conducted a simulation study to investigate the performance of hybrid ML methods. Our data generating models follow Kim and Steiner (2015), Kim, Steiner, and Lim (2016), and our motivating data, which had a two-level structure with one continuous outcome and one binary treatment. We consider four continuous covariates, two of which are individual-level covariates and the other two are cluster-level covariates. We also assume two latent classes defined by the latent selection model discussed before and each latent class has its own unique treatment effect. The details of our data generating procedure are stated below.

1. Let *nC*1 and *nC*2 represent the number of clusters in latent class $k = 1$ and latent class $k = 2$. For each cluster in each latent class, we generate the number of individuals n_j based on drawing samples from a Normal distribution with mean *nS* set to either 30 or 50, variance *v* set to either 4 or 16, and round them to the nearest
2. For each individual i in cluster j, randomly sample two cluster-level covariates, $\mathbf{W}_j = (W_{1j}, W_{2j})$ and two individual-level covariates, $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij})$, from the following distributions

$$\begin{pmatrix} W_{1j} \\ W_{2j} \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 2 & .2 \\ .2 & 2 \end{pmatrix} \right]$$

$$\begin{pmatrix} X_{1ij} \\ X_{2ij} \end{pmatrix} \sim \left[\begin{pmatrix} 0.1W_{1j} + 0.05W_{2j} + \kappa_{1j} \\ 0.08W_{1j} + 0.1W_{2j} + \kappa_{2j} \end{pmatrix}, \quad \begin{pmatrix} 10 & 2 \\ 2 & 15 \end{pmatrix} \right]$$

$$\begin{pmatrix} \kappa_{1j} \\ \kappa_{2j} \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix} \right]$$

We remark that the individual-level covariates' means μ_j are a function of the cluster-level covariates $\mathbf{W}_j$ and random errors $\boldsymbol{\kappa}_j$. We set larger variances for individual-level covariates than cluster-level covariates to reflect typically higher variations in individual-level covariates than cluster-level covariates.

3. For each latent class $k = 1, 2$, define the propensity score model $e_k(\cdot)$ based on a random effects logistic selection model.

$$logit\left(e_1(\mathbf{X}_{ij}, \mathbf{W}_j)\right) = 0 + 0.15X_{1ij} + 0.1X_{2ij} + 0.1W_{1j} + 0.2W_{2j} + 0.2X_{1ij}W_{2j} + R_{j1}$$
$$logit\left(e_2(\mathbf{X}_{ij}, \mathbf{W}_j)\right) = -0.05 + 0.05X_{1ij} - 0.05X_{2ij} + 0.2W_{1j} + 0.05W_{2j} + R_{j2}$$
$$R_{j1} \sim N(0, 0.5), \qquad R_{j2} \sim N(0, 0.2)$$

Here, $e_k(\cdot)$ is the propensity score for individual i in cluster j which belongs to latent class k. R_{jk} is the random effect for cluster j in class k. The slope coefficients for Class 1 and Class 2 differ where Class 1 has a stronger selection than Class 2. The intra-class correlations for Class 1 and Class 2 are around 0.13 and 0.06, respectively.

4. For each individual i in cluster j which belongs to latent class k, generate individual treatment status, Z_{ij} (0 = untreated; 1=treated) from a Bernoulli distribution with the propensity score specified above.

$$Z_{ij} \sim \begin{cases} Bernoulli\left(e_1(\mathbf{X}_{ij}, \mathbf{W}_j)\right), & \text{if } i \text{ belongs to latent class } k = 1 \\ Bernoulli\left(e_2(\mathbf{X}_{ij}, \mathbf{W}_j)\right), & \text{if } i \text{ belongs to latent class } k = 2 \end{cases}$$

5. For each individual i in cluster j which belongs to latent class k, generate potential outcomes and observed outcomes based on random effects linear regression models.

$$Y_{ij1}(z) = 100 + 2.5z + 2X_{1ij} + 1X_{2ij} + 2W_{1j} + 1.5W_{2j} + 0.5X_{2ij}W_{1j} + 0.3X_{2ij}^2 + U_{j1} + \epsilon_{ij1}$$
$$Y_{ij2}(z) = 80 + 0z + 1X_{1ij} + 0.5X_{2ij} + 1W_{1j} + 0.5W_{2j} + 0.2X_{2ij}W_{1j} + 0.2W_{1j}W_{2j} + U_{j2} + \epsilon_{ij2}$$
$$Y_{ij} = Z_{ij}Y_{ijk}(1) + \big(1 - Z_{ij}\big)Y_{ijk}(0), \qquad U_{j1} \sim N(0,10), \qquad U_{j2} \sim N(0,7), \qquad \epsilon_{ijk} \sim N(0,100)$$

The term U_{jk} is the random effect for cluster j in latent class k, and ϵ_{ijk} is the random error for individual i in cluster j which belongs to latent class k. The treatment effect is positive for Class 1, but zero for Class 2 so that each latent class has distinct treatment effects. The intra-class correlations are 0.10 and 0.07 for Classes 1 and 2, respectively. Additionally, there are non-linear and/or interaction terms in the outcome model.

In our simulation study, we varied the following simulation parameters: the size of each latent class, *nC*1 and *nC*2, and the mean size of each cluster *nS*. We examined the performance of hybrid ML methods—Hybrid Causal Forests, Hybrid BART, and Hybrid TMLE—in estimating latent class average treatment effects τ(*k*). In Appendix I, we examined the performance of the estimated individual CATE. As for software, we use the *grf* package (Tibshirani et al., 2019) for Causal Forests, *bartCause* package (Dorie & Hill, 2019) for BART, and *tmle* package (Gruber & van der Laan, 2012) for TMLE, all implemented in R (R Core Team, 2019). As a comparison, we ran within-class matching (Kim & Steiner, 2015; Kim et al., 2016) as an alternative to hybrid ML methods which estimate average treatment effects within each latent class via propensity score within-class matching. In brief, within-class matching is a type of multilevel matching that matches treated and control units across clusters, but within the same latent classes defined by latent selection models. Within-class matching uses the same latent selection model as above to identify latent classes and requires specifying a weighing

function that depends on the estimated propensity score to obtain estimates of the treatment effect within each latent class. For our simulation, the propensity score for within-class matching was estimated using random effects logistic regression. For weighing, we used inverse probability weighting (IPW) and marginal mean weighing through stratification (MMW-S) (Hong & Hong, 2009). Specifically, the IPW estimator for latent class *k* using within-class matching is

$$\hat{\tau}_{IPW}(k) = \frac{1}{N_k} \sum_{ij:\widehat{K}_{ij}=k} \frac{Y_{ij}Z_{ij}}{e_k(\mathbf{X}_{ij}, \mathbf{W}_j)} - \frac{1}{N_k} \sum_{ij:\widehat{K}_{ij}=k} \frac{Y_{ij}(1-Z_{ij})}{1-e_k(\mathbf{X}_{ij}, \mathbf{W}_j)}$$

and the MMW-S estimator for latent class *k* using within-class matching is

$$\omega_{z,ij(k)} = \begin{cases} \frac{E_{z1(k)}}{O_{z1(k)}} & \text{if } e(\mathbf{X}_{ij}, \mathbf{W}_j) \text{ in stratum 1 of latent class } k \\ \vdots & \\ \frac{E_{zS(k)}}{O_{zS(k)}} & \text{if } e(\mathbf{X}_{ij}, \mathbf{W}_j) \text{ in stratum } S \text{ of latent class } k \end{cases}$$

$$\hat{\tau}_{MMW-S}(k) = \frac{1}{N} \sum_{ij:\widehat{K}_{ij}=k} Y_{ij}Z_{ij}\, \omega_{1,ij(k)} - \frac{1}{N} \sum_{ij:\widehat{K}_{ij}=k} Y_{ij}(1-Z_{ij})\, \omega_{0,ij(k)}$$

where $O_{zs(k)}$ is the observed frequency of individuals in treatment status $z \in \{0,1\}$ and stratum $s \in \{1,2,\ldots,S\}$ of the distribution of the propensity score in latent class k, and $E_{zs(k)}$ is the expected frequency assuming the distributions between treated and untreated units are the same across strata. We created 10 strata of propensity scores for MMW-S.

We also computed the doubly robust (DR) estimator as follows:

$$\hat{\tau}_{DR}(k) = \frac{1}{N_k} \sum_{ij:\widehat{K}_{ij}=k} \left[\frac{Y_{ij}Z_{ij}}{e_k(\mathbf{X}_{ij}, \mathbf{W}_j)} - \frac{Z_{ij} - e_k(\mathbf{X}_{ij}, \mathbf{W}_j)}{e_k(\mathbf{X}_{ij}, \mathbf{W}_j)} m_k(\mathbf{X}_{ij}, \mathbf{W}_j, 1) \right]$$

$$-\frac{1}{N_k}\sum_{ij:\widehat{R}_{ij}=k}\left[\frac{Y_{ij}(1-Z_{ij})}{1-e_k(\mathbf{X}_{ij},\mathbf{W}_j)}+\frac{Z_{ij}-e_k(\mathbf{X}_{ij},\mathbf{W}_j)}{1-e_k(\mathbf{X}_{ij},\mathbf{W}_j)}m_k(\mathbf{X}_{ij},\mathbf{W}_j,0)\right]$$

Here, $e_k(\mathbf{X}_{ij},\mathbf{W}_j)$ is the propensity score estimated by random-effects logistic regression models within each latent class k. $m_k(\mathbf{X}_{ij},\mathbf{W}_j,0)$ and $m_k(\mathbf{X}_{ij},\mathbf{W}_j,1)$ are outcome models based on random-effects linear regression models within each latent class *k*. Both models are specified to be the same as those from the data generating models. For more details on other propensity score techniques, see Schafer and Kang (2008), Austin (2011), and Steiner and Cook (2013).

Each method was evaluated based on the absolute bias and MSE of class-specific average treatment effect estimates. Specifically, given $m = 1,\ldots,400$ simulation replications and their corresponding estimates $\hat{\tau}_m(k)$ $(m = 1,\ldots,400)$, the absolute bias and MSE within each class are defined as:

$$|\text{Bias}(k)| = \left|\frac{1}{400}\sum_{m=1}^{400}(\hat{\tau}_m(k)-\tau(k))\right|, \quad MSE = \frac{1}{400}\sum_{m=1}^{400}(\hat{\tau}_m(k)-\tau(k))^2$$

We also evaluate the overall performance across latent classes by computing the overall bias and MSE as follows:

$$|\text{Bias}| = \left|\frac{1}{400}\sum_{m=1}^{400}\sum_{k=1}^{2}\frac{N_k^*}{N_m}(\hat{\tau}_m(k)-\tau(k))\right|, \qquad MSE = \frac{1}{400}\sum_{m=1}^{400}\sum_{k=1}^{2}\frac{N_k^*}{N_m}(\hat{\tau}_m(k)-\tau(k))^2$$

The term N_k^* denotes the true sample size in each latent class k and N_m denotes the total sample size in each simulation replication.

Simulation Results

Table 3.1 summarizes the mean percentage of correctly identifying latent class membership. Mean percentages for modal assignment were calculated by comparing the true class membership of each individual with the estimated class membership from the mixture model, while for proportional assignment, the mean percentages were computed by using a weighted average of the latent class posterior probabilities in each true class. We found that modal assignment classified the latent classes more accurately than proportional assignment. Also, classification rates were affected by cluster sizes and the number of clusters. In particular, we found that increasing the size of the clusters had a larger impact on classification rates than increasing the number of clusters.

Table 3.1
Classification rate (%) in class membership

(nC1, nC2, nS)	Modal Assignment	Proportional Assignment
(25, 25, 30)	73.42	71.04
(25, 25, 50)	84.04	81.22
(50, 50, 30)	78.81	74.92

Note: nC1, nC2, and nS represent the number of clusters for the first latent class, the number of clusters for the second latent class, and average cluster sizes, respectively.

Figure 3.2 displays results of class-specific average treatment effect estimates; see Table G1 in Appendix G for numerical results. Across simulation conditions, hybrid ML methods generally performed better than DR and non-DR methods in terms of overall bias and MSE. MSE. Indeed, it is not surprising that with the misclassification of the latent classes, the outcome model and/or the propensity scores are inherently incorrect inside the DR and non-DR estimators and thus, traditional parametric methods—IPW, MMW-S, and DR estimators—are directly affected by misclassified units in latent classes. In contrast, ML methods are often “local” non-parametric methods which are more robust to model mis-specifications. Of course, if the

misclassification rate is fairly high, then it is unlikely that any method will work properly. Also, as seen from Figure 3.2, when the sample sizes increased from (25, 25, 30) to (25, 25, 50) or (50, 50, 30), we saw that overall bias and overall MSE decreased across different estimators, but the magnitudes varied depending on the exact sample proportions within each latent class.

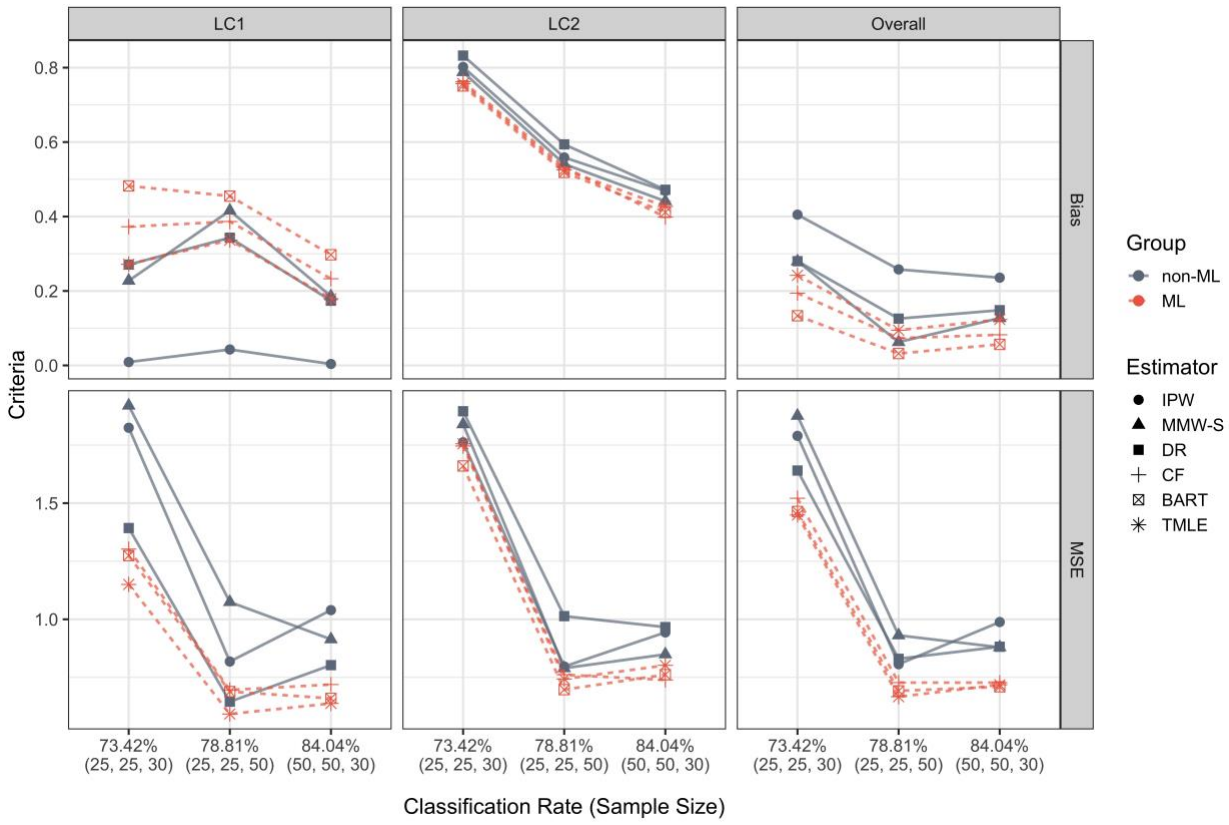

Figure 3.2. Performance of class-specific treatment effect estimates with classification rates and sample sizes. The three values in parentheses represent the number of clusters for the first latent class, the number of clusters for the second latent class, and the average cluster sizes, respectively. IPW represents inverse-propensity weighting, and MMW-S represents marginal mean weighting through stratification. DR represents the doubly robust estimator. BART represents Bayesian additive regression trees, and TMLE represents targeted maximum likelihood estimation. The true treatment effects are 2.5 and 0 for the first and second latent classes, respectively.

We remark that the bias of the IPW estimator in latent class 1 was surprisingly small, but the overall bias was still larger than other methods. This suggests that the IPW estimator traded

off a large bias reduction in latent class 1 at the expense of an increase in bias in latent class 2. But this trade-off was not "balanced" and led to a large overall bias. In contrast, hybrid methods like BART had a large bias in latent class 1 and a small bias in latent class 2 but achieved the smallest overall bias. In Appendix H, we further examined the bias trade-off between latent classes and found that the IPW estimator exhibited this phenomenon in other settings. Overall, our results demonstrate that hybrid ML methods provides accurate and precise estimates of the treatment effect and are an attractive alternative to those based on parametric propensity score techniques, IPW and MMW-S, or parametric DR methods.

TIMSS Data Study: The Effects of Private Science Lessons

Data and Variables

We revisit the question in the introduction and study the heterogeneous effects of private science lessons on students' science achievement scores where we suspected distinct latent selection processes across clusters of students. The data comes from the 2015 Trends in International Mathematics and Science Study (TIMSS) data. TIMSS is an international educational assessment that examines the progression of students' performance in mathematics and science and it was first conducted in 1995 by the International Association for the Evaluation of Educational Achievement (IEA). Since 1995, TIMSS has been conducted for 4-th and 8-th graders every four years in more than 40 countries. The recent data collection, which took place in 2015, was conducted in 60 countries and a new data collection was planned in 2019. The data are based on a two-stage stratified cluster sampling; schools are chosen first according to each country's important demographic variables (e.g., in Korea, school location, and/or whether schools are gendered), and then at least one intact classroom is randomly chosen from each

school (Martin, Mullis, & Hooper, 2016). We used the Korea TIMSS 2015 data of 8th graders for our analysis.

The original data included 5,309 students from 150 middle schools with varying school sizes (a range of 6 to 75; mean of 35.4 students per school; median of 32 students per school). We removed students with 1) inconsistent responses about their attendance of private science lessons and 2) missing information in 7 out of 12 covariates (see below for a list of covariates). Our final sample was 4,874 students (91.81% of the original data) from 149 schools. For simplicity and to demonstrate the new methodology, we did not consider multiple plausible values of student achievements in the sciences and ignored sampling weights. However, to rigorously evaluate the effects of private lessons and generalize these results, it is necessary to consider five different plausible values and sampling weights; see Rutkowski, Gonzalez, Joncas, and von Davier (2010) and Foy, Arora, and Stanco (2017) for details.

The treatment variable was whether a student received private science lessons ($Z_{ij} = 1$) or not ($Z_{ij} = 0$). The outcome Y_{ij} was the first plausible value of achievement in the sciences. We included 12 covariates that affected the selection and outcome processes, including six student-level covariates $\mathbf{X}_{ij}$ and six school-level covariates $\mathbf{W}_j$. The student-level covariates were student's gender (*male*), fathers' highest education levels (*dad.edu*, with three levels; no college, college graduates *dad.cll*, and don't know *dad.q*), the number of books at home (*books25*, with two levels; more than 25, and less than or equal to 25), the number of home study supports (*hspprt*, with three levels; neither own room nor Internet connection, one of them *hspprt.1*, and both *hspprt.2*), student's confidence in science (*sci.conf*), and student's perceived value of science (*sci.value*). The school-level covariates were school's gender type (*gender.type*, with

three levels; all-boys, all-girls *girl.sch*, and co-education *coedu*), the percentage of economically disadvantaged students (*pct.disad*, with four levels; 0 to 10%, 11 to 25% *disad.11*, 26 to 50% *disad.26*, and more than 50% *disad.M50*), school location (*city.size*, with four levels; urban *city.U*, suburban *city.Sub*, medium size city *city.M*, and small town), science instruction affected by resource shortage (*res.short*), school's emphasis on academic success (*aca.emph*), and school discipline problems (*dscpn*).

Results

In the first step of our hybrid ML methods, we determined the optimal number of latent classes by comparing the AIC measures under different numbers of latent classes. The two-class model had the lowest AIC and Latent Class 1 had 1,556 students from 44 schools, and Latent Class 2 had 3,318 students from 105 schools. Latent Class 1 was about 60% smaller than Latent Class 2; Latent Class 1 had 31.9% of the total students from 29.5% of the schools, and Latent Class 2 had 68.1% of the total students from 70.5% of the schools.

Figure 3.3 plots the estimated selection models $\hat{e}_k$ from each latent class as a function of two observed covariates, value in science and resource shortages. The figure contains the line of best fit to guide visualization. For Latent Class 2, the propensity of taking private lessons was linearly increasing with how much value students placed in the sciences (*sci.value*). However, for Latent Class 1, the propensity remained flat and there was no discernable relationship between *sci.value* and selection probabilities. We also observed that a cluster-level covariate *res.short* increased the selection probabilities in Class 1, but there was no increasing pattern in Class 2.

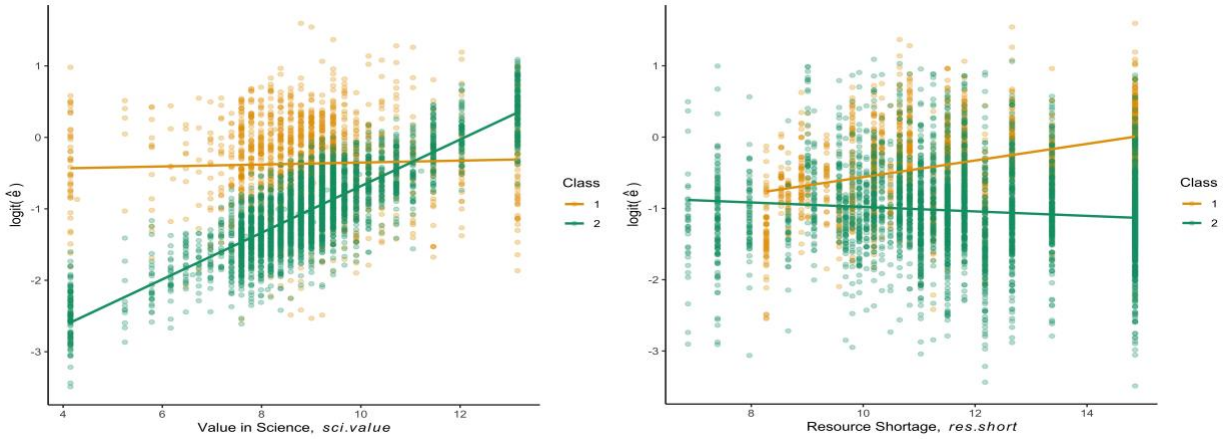

Figure 3.3. Class-specific selection models with respect to individual-level and cluster-level covariates. Each dot indicates a students' estimated logit propensity score, logit($\hat{e}_k$). Line of best fit is plotted to guide visualization.

We also summarize the student-level and school-level variables in Table 3.2. For most covariates, the two latent classes showed similar descriptive statistics. However, we found statistically significant differences in the propensity of taking private science lessons and father's educational level. Students in Class 1 were more likely to have a higher probability of taking private lessons and come from families whose father did not hold a college degree than those in Class 2. Looking at both Table 3.2 and Figure 3.3, students in Class 1 likely sought private lessons because they may receive inadequate lessons at schools, whereas those in Class 2 likely sought private lessons because their families had stronger education backgrounds and may had placed a high value in science. Overall, as suspected from our subject-matter expertise, the latent class selection model revealed different latent structures where students in each latent class had different propensities to seek private tutors.

Figure 3.4 shows the distributions of individual CATE estimates from Hybrid Causal Forests. The figure also shows vanilla Causal Forests that did not consider latent class membership. Using Hybrid Causal Forests, we see two different distributions centering around

one and ten, respectively, to reflect variation in treatment effects between latent classes. In contrast, the vanilla Causal Forests only shows variation in treatment effects in the observed covariates and centers around four. Appendix J shows the distributions of CATE based on other ML methods.

Table 3.2
Descriptive statistics of the two latent classes

	Class 1		Class 2	
	Mean or Percent	Std. Dev	Mean or Percent	Std. Dev
Student-level Variables	(N=1,556 students)		(N=3,318 students)	
science.score	557.84	75.94	555.97	76.25
math.score	608.24	82.74	605.13	84.63
propensity.score	**0.42**		**0.28**	
sci.conf	8.68	2.09	8.61	2.11
sci.value	8.98	1.64	8.92	1.64
male	51.6%		48.6%	
dad.cll	**34.4%**		**37.5%**	
dad.q	28.2%		28.1%	
books25	86.1%		86.0%	
hssprt.2	70.2%		71.9%	
School-level Variables	(J=44 schools)		(J=105 schools)	
res.short	11.71	2.04	11.79	2.00
aca.demph	11.14	1.84	11.10	1.87
dscpn	10.74	2.07	11.16	1.99
girl.sch	15.9%		21.9%	
coedu	65.9%		58.1%	
disad.11	29.5%		37.1%	
disad.26	27.3%		23.8%	
disad.M50	11.4%		10.5%	
city.U	40.9%		35.2%	
city.Sub	6.8%		9.5%	
city.M	31.8%		27.6%	

Note: Values in bold are significant differences between classes at α=0.05.

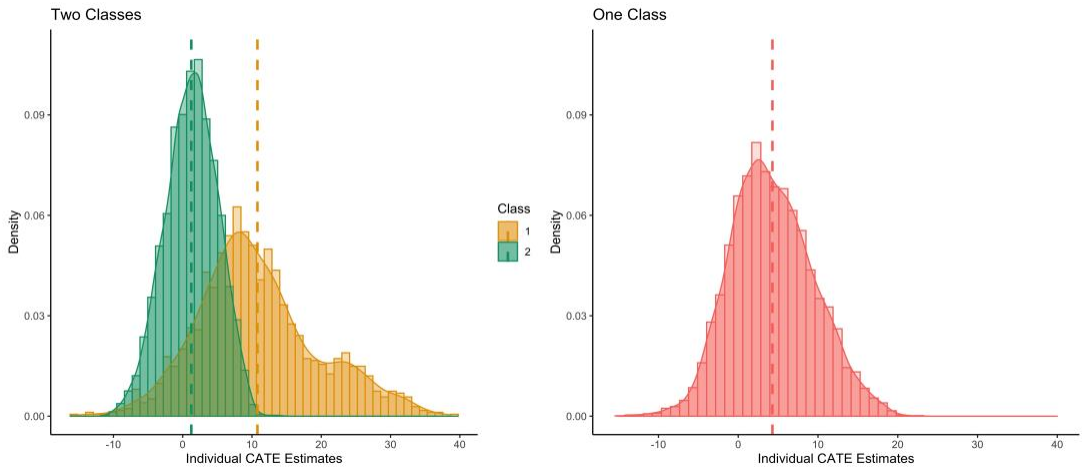

Figure 3.4. Distributions of individual CATE estimates from Causal Forests. The left shows Hybrid Causal Forests discovering two latent classes, while the right shows vanilla Causal Forests without consideration for latent classes. Dashed lines represent class-specific treatment effect estimates.

Table 3.3 summarizes average treatment effect estimates of private science lessons within each latent class. As a comparison, we used within-class IPW estimator, MMW-S, and DR estimator with parametric propensity score or outcome models to estimate class-specific average treatment effects; we remark that hybrid ML methods do not require a priori specification of the propensity score or the outcome model. We included covariate balance plots in Appendix K. After estimating the average treatment effect within latent classes, we found that the prima facie effects amounted to 16.99 and 20.24 for Class 1 and Class 2, respectively. The prima facie effect is the unadjusted mean difference in science achievement scores between the treated and untreated groups. The treatment effects with IPW, MMW-S, and DR estimators varied depending on the latent class; there were significantly positive effects in Class 1, while no significant effects existed in Class 2. When we implemented hybrid ML methods, we observed that the average treatment effect estimates in both classes were similar to parametric methods except that the estimates for Class 2 were generally smaller than parametric methods.

For the one class model (far right column) where we assumed no latent class structure, the prima facie effect amounted to 18.96 points. After applying MMW-S, the effect decreased to 1.73 points, which was not statistically significant. However, one-class IPW and DR methods produced positive, significant, but reduced effects. Also, the average treatment effects with hybrid ML methods were positive, but slightly smaller than IPW and DR estimators. However, as mentioned before, none of these effects uncovered the potential effect heterogeneity within latent classes.

Table 3.3
Comparisons of the class-specific average treatment effect estimates

	Two Classes				One Class	
	Class 1		Class 2			
	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Prima facie (unadjusted)	**16.99**	(3.88)	**20.24**	(2.93)	**18.96**	(2.32)
IPW	**11.62**	(2.58)	2.93	(2.94)	**5.36**	(2.06)
MMW-S	**12.28**	(2.69)	2.33	(2.78)	1.73	(2.08)
DR	**11.78**	(2.64)	3.02	(2.89)	**5.39**	(1.99)
Hybrid Causal Forests	**10.79**	(2.11)	1.24	(2.14)	**4.28**	(1.58)
Hybrid BART	**12.24**	(3.26)	0.73	(2.43)	**4.36**	(1.90)
Hybrid TMLE	**12.19**	(3.13)	1.40	(2.39)	**4.54**	(1.87)

Note: Standard errors (SE) were estimated using bootstrap sampling with 5,000 repetitions. Estimates in bold are significant at α=0.05. IPW represents inverse-propensity weighting, and MMW-S represents marginal mean weighting through stratification. DR represents the doubly robust estimator. BART represents Bayesian additive regression trees, and TMLE represents targeted maximum likelihood estimation.

Discussion and Conclusions

We propose hybrid ML methods to estimate heterogeneous treatment effects between latent classes. Our proposed hybrid approach uses context-specific finite mixture models to identify different latent classes and ML-based causal inference methods to estimate treatment effects within each class. Broadly speaking, hybrid ML methods extend the capacities of ML

methods to capture treatment effect heterogeneity defined by latent class mixture models. Our simulation study revealed that hybrid ML methods are an attractive alternative to existing propensity score methods. Finally, in our data analysis, we demonstrated that hybrid ML methods were able to capture heterogeneous effects and the average treatment effect for each latent class.

We make three concluding remarks about hybrid ML methods. First, ensuring sufficient sample sizes is important when using multilevel latent class mixture models. We observed that increasing cluster sizes affected the proportions of correctly identifying class membership and we generally recommend using hybrid ML methods when the number of clusters is more than 50 and the mean cluster size is more than 30 so that the total sample size is at least 1500, the minimum sample size in our simulation design. Hybrid ML methods did perform well in our simulation study even when the maximum misclassification rate was about 27%. However, in general, if there are insufficient samples, there is an increased likelihood of misclassifying units and consequently, an increased risk of biasing the average treatment effect. Second, though ML methods can flexibly fit the outcome model and the propensity score model, this does not give a free pass for mis-classification in latent class models, and it would be an interesting topic of future research to design ML methods to be robust to biases arising from mis-classification in latent class models. Third, we believe that our work here provides a more systematic approach of applying ML methods for causal inference to education and psychology. In particular, we hope that the work provides a template for researchers to combine other types of latent class modeling with any ML-based causal inference methods to better understand the nature of treatment effect heterogeneity and the underlying latent structures in the data.

CONCLUSIONS

The three studies of this dissertation discussed different methodological challenges associated with estimating treatment effects in educational assessment data. Each study proposed appropriate methods and demonstrated the proposed methods in large-scale educational assessment data such as NAEP and TIMSS data. Specifically, in the first study, the proposed regression discontinuity design with an ordinal running variable was used to assess the effects of extended time accommodations on students' math proficiency from the 2017 NAEP data. The second study investigated optimal modifications for Causal Forests to enhance its performance in multilevel/clustered observational data, and the proposed modifications including injecting multilevel propensity scores (the most effective strategy) were applied to TIMSS data for estimating the effects of private math lessons on students' math achievement scores. The third study proposed a two-step hybrid procedure for ML-based causal inference methods to estimate heterogeneous treatment effects between latent classes; the first step uses context-specific finite mixture models to identify different latent classes, and the second step uses ML-based causal inference methods to estimate treatment effects within each class. The proposed approach was performed to assess the effects of private science lessons on students' science achievement scores. Overall, these three studies provide useful guidelines and suggestions for applied researchers who wish to use regression discontinuity designs or ML-based causal inference methods in educational assessment data.

There are some limitations to this study that may impact the interpretation of the results from empirical examples. First, we did not incorporate multiple plausible values of students' proficiency and ignored sampling weights (and jackknife replicate weights). Thus, our empirical

results did not generalize to the target population of NAEP or TIMSS. Second, the second and third studies assume that there are no unmeasured confounders. But the real data examples in the two studies are based on the 2015 TIMSS data that were cross-sectional data and lacked the variable of the pre-test score. Since the pre-test score is the most important predictor of the post-test score, our effect estimates may still have the remaining bias from unmeasured confounders. Therefore, it would be better to conduct a sensitivity analysis to check the robustness of our effect estimates against unmeasured confounding or to use longitudinal datasets like ECLS-K that contain pre-test scores. Despite these limitations, this dissertation provides researchers with modern tools to estimate causal effects in increasingly large and complex educational assessment data.

REFERENCES

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, *37*(6A), 3099–3132.

Arpino, B., & Cannas, M. (2016). Propensity score matching with clustered data. an application to the estimation of the impact of caesarean section on the apgar score. *Statistics in Medicine*, *35*(12), 2074–2091. doi: https://doi.org/10.1002/sim.6880

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780. doi: https://doi.org/10.1016/j.csda.2010.11.008

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. doi: https://doi.org/10.1073/pnas.1510489113

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

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological methodology*, *43*(1), 272–311.

Bandura, A. (2010). Self-efficacy. In I. B. Weiner & W. E. Craighead (Eds.), *The corsini encyclopedia of psychology* (4th ed., pp. 1–3). Hoboken, NJ : Wiley.

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

Bergner, Y., & von Davier, A. A. (2018). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*. doi:https://doi.org/10.3102/1076998618784700

Bradley, R., Katti, S., & Coons, I. J. (1962). Optimal scaling for ordered categories. *Psychometrika*, *27*(4), 355–374.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.

Butera, N. M., Lanza, S. T., & Coffman, D. L. (2014). A framework for estimating causal effects in latent class analysis: Is there a causal link between early sex and subsequent profiles of delinquency? *Prevention science*, *15*(3), 397–407.

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*. 5, 21-35.

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

Clogg, C. C. (1995). Latent class models. In S. M. E. Arminger G. Clogg C. C. (Ed.), *Handbook of statistical modeling for the social and behavioral sciences* (p. 311-359). Springer.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* CENGAGE Learning.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dimakopoulou, M., Zhou, Z., Athey, S., & Imbens, G. (2017). Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*.

Donner, A., & Klar, N. (2010). *Design and analysis of cluster randomization trials in health research*. New York: Wiley.

Dorie, V., & Hill, J. (2019). bartcause: Causal inference using bayesian additive regression trees [Computer software manual]. Retrieved from https://github.com/vdorie/bartCause (R package version 1.0-0)

Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). Springer.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman and Hall.

Foy, P., Arora, A., & Stanco, G. (2017). *Timss 2015 user guide for the international database*. TIMSS & PIRLS International Study Center, Boston College.

Goodman, L. A. (2007). On the assignment of individuals to latent classes. *Sociological Methodology*, *37*(1), 1–22.

Gregg, N., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, *45*(2), 128–138.

Gruber, S., & van der Laan, M. J. (2012). tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, *51*(13), 1–35. Retrieved from http://www.jstatsoft.org/v51/i13/ (doi:10.18637/jss.v051.i13)

Grün, B., & Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of classification*, *25*(2), 225–247.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209.

Hallquist, M., & Wiley, J. (2017). Mplusautomation: Automating mplus model estimation and interpretation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=MplusAutomation (R package version 0.7)

Hayes, R. J., & Moulton, L. H. (2009). *Cluster randomised trials*. Chapman & Hall/CRC. doi: https://doi.org/10.1201/9781584888178

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

Hong, G. (2010). Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, *35*(5), 499–531. doi: https://doi.org/10.3102/1076998609359785

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis*, *31*(1), 54–81. doi: https://doi.org/10.3102/0162373708328259

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

Hong, G., & Raudenbush, S. W. (2013). Heterogeneous agents, social interactions, and causal inference. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 331–352). Springer. doi: https://doi.org/10.1007/978-94-007-6094-3$_1$6

IBM. (2019). *IBM SPSS Categories 26.* IBM Corporation. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/26.0/en/client/Manuals/IBM_SPSS_Categories.pdf

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443–470.

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

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*(467), 854–866. doi: https://doi.org/10.1198/016214504000001187

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710. doi: https://doi.org/10.1093/biomet/87.3.706

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, *142*(2), 615–635.

Jo, B., Wang, C.-P., & Ialongo, N. S. (2009). Using latent outcome trajectory classes in causal inference. *Statistics and its Interface*, *2*(4), 403.

Johnson, E. G. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement*, *29*(2), 95–110.

Jonson, J. L., Trantham, P., & Usher-Tate, B. J. (2019). An evaluative framework for reviewing fairness standards and practices in educational tests. *Educational Measurement: Issues and Practice*, *38*(3), 6–19.

Kang, H., & Keele, L. (2018). Estimation methods for cluster randomized trials with noncompliance: A study of a biometric smartcard payment system in india. *arXiv preprint arXiv:1805.03744*.

Kang, J., & Schafer, J. L. (2010). *Estimating average treatment effects when the treatment is a latent class* (Tech. Rep. No. 1005). Department of Statistics, The Pennsylvania State University.

Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-scale Assessments in Education*, *4*(1), 1-24.

Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). Multilevel latent variable modeling: Current research and recent developments. In R. Millsap & A. Maydeu-Olivares (Eds.), *Handbook of quantitative methods in psychology* (pp. 592–612). Sage.

Keller, B., Kim, J.-S., & Steiner, P. M. (2015). Neural networks for propensity score estimation: Simulation results and recommendations. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research: The 80th annual meeting of the psychometric society* (pp. 279–291). Springer. doi: https://doi.org/10.1007/978-3-319-19977-1_20

Kim, J., & Seltzer, M. (2007). Causal inference in multilevel settings in which selection processes vary across schools. cse technical report 708. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. doi: https://doi.org/10.1037/e644002011-001

Kim, J.-S., & Steiner, P. M. (2015). Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research: The 80th annual meeting of the psychometric society* (pp. 293–306). Springer. doi: https://doi.org/10.1007/978-3-319-19977-1_21

Kim, J.-S., Steiner, P. M., & Lim, W. C. (2016). Mixture modeling methods for causal inference with multilevel data. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research* (pp. 335–359). Information Age Publishing.

Kim, Y. Y., & Circi, R. (2018). *The extended time accommodation (ETA) and performance of students with eta.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, NY. U.S.

Kim, Y. Y., & Circi, R. (2019). *Effects of the extended time accommodation on performance in naep mathematics.* Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

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

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

Lanza, S. T., Coffman, D. L., & Xu, S. (2013). Causal inference in latent class analysis. *Structural equation modeling: a multidisciplinary journal*, *20*(3), 361–383.

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, *48*(2), 281–355.

Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent glass regression in R. *Journal of Statistical Software*, *11*.

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

Lenk, P. J., & DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, *65*(1), 93–119.

Li, L., Lu, Y., & Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 2071–2080).

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, *23*(19), 2937–2960. doi: https://doi.org/10.1002/sim.1903

Ma, X. (2018). *Using classification and regression trees: A practical primer*. Information Age Publishing.

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 175–198).

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and procedures in timss 2015*. TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html

Masyn, K. E. (2013). latent class analysis and finite mixture modeling. In T. Little (Ed.), *The oxford handbook of quantitative methods* (p. 551-611). New York, NY: Oxford University Press.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403. doi: https://doi.org/10.1037/1082-989X.9.4.403

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. Wiley.

Meulman, J. J. (1998). *Optimal scaling methods for multivariate categorical data analysis.* SPSS White Paper: Chicago.

Meulman, J. J., van der Kooij, A. J., & Duisters, K. L. W. (2019). Ros regression: Integrating regularization with optimal scaling regression. *Statistical science*, *34*(3), 361–390.

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, *41*(4), 473–497. doi: https://doi.org/10.1207/s15327906mbr4104_3

Mok, J.-Y., Choi, S.-W., Kim, D.-J., Choi, J.-S., Lee, J., Ahn, H., ... Song, W.-Y. (2014). Latent class analysis on internet and smartphone addiction in college students. *Neuropsychiatric disease and treatment*, *10*, 817-828.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: methods and principles for social research. cambridge univ*. Cambridge University Press.

National Center for Education Statistics. (2017a). *2017 mathematics grades 4 and 8 assessment report cards: Summary data tables for national and state sample sizes, participation rates, proportions of SD and ELL students identified, and types of accommodations.* Retrieved from https://www.nationsreportcard.gov/math_2017/files/2017_Technical_Appendix_Math_State.pdf

National Center for Education Statistics. (2017b). *2017 reading grades 4 and 8 assessment report cards: Summary data tables for national and state sample sizes, participation rates, proportions of SD and ELL students identified, and types of accommodations.* Retrieved from https://www.nationsreportcard.gov/reading_2017/files/2017_Technical_Appendix_Reading_State.pdf

National Research Council. (2002). *Reporting test results for students with disabilities and english-language learners: Summary of a workshop*. National Academies Press.

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.

Oranje, A., & Kolstad, A. (2019). Research on psychometric modeling, analysis, and reporting of the national assessment of educational progress. *Journal of Educational and Behavioral Statistics*, *44*(6), 648–670.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo: Morgan Kaufman.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rickles, J. H., & Seltzer, M. (2014). A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics, 39*(6), 612–636. doi: https://doi.org/10.3102/1076998614559748

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*(427), 846–866. doi: https://doi.org/10.1080/01621459.1994.10476818

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961–962.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169–188. doi: https://doi.org/10.1023/A:1020363010465

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151.

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

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association, 94*(448), 1096–1120. doi: https://doi.org/10.1080/01621459.1999.10473862

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334–1344. doi: https://doi.org/10.1198/016214508000000733

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. Little (Ed.), *The oxford handbook of quantitative methods* (p. 236-258). New York, NY: Oxford University Press.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250. doi: https://doi.org/10.1037/a0018719

Steiner, P. M., Kim, J.-S., & Thoemmes, F. (2012). Matching strategies for observational multilevel data. In *Joint statistical meeting proceedings* (pp. 5020–5032). Alexandria, VA: American Statistical Association.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Singer, J. D., Braun, H. I., & Chudowsky, N. (2018). *International education assessments: Cautions, conundrums, and common sense*. Washington, DC: National Academy of Education.

Steiner, P. M., Kim, Y., Hall, C. E., & Su, D. (2017). Graphical models for quasi-experimental designs. *Sociological methods & research*, *46*(2), 155–188.

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

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. Journal of Machine Learning Research, 10(5), 141–158.

Suk, Y., Kang, H., & Kim, J.-S. (2020). Random forests approach for causal inference with clustered observational data. *Multivariate Behavioral Research*. doi: 10.1080/00273171.2020.1808437.

Sutfin, E. L., Reboussin, B. A., McCoy, T. P., & Wolfson, M. (2009). Are college student smokers really a homogeneous group? a latent class analysis of college student smokers. *Nicotine & Tobacco Research*, *11*(4), 444–454.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*(3), 514–543. doi: https://doi.org/10.1080/00273171.2011.569395

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Miner, L., Wager, S., & Wright, M. (2019). grf: Generalized random forests (beta) [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=grf (R package version 0.10.3)

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester: Wiley.

Tueller, S. J., Drotar, S., & Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling*, *18*(1), 110–131.

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

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

VanderWeele, T. J., & Arah, O. A. (2011). Unmeasured confounding for general outcomes, treatments, and confounders: bias formulas for sensitivity analysis. *Epidemiology*, *22*(1), 42–52.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological methodology*, *33*(1), 213–239.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450–469.

Vermunt, J. K., & Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis, 41*(3-4), 531–537.

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

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using mplus*. John Wiley & Sons.

Wang, M. C., & Gordon, E. W. (2012). *Educational resilience in inner-city america: Challenges and prospects*. Routledge.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826–833. Doi https://doi.org/10.1016/j.jclinepi.2009.11.020

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

World-Class Instructional Design, & Assessment. (2013). *ACCESS for ELLs: Interpretive guide for score reports, spring 2013.* Retrieved from http://www.wrsdcurriculum.net/ACCESSInterpretiveGuide2013_1_.pdf

World-Class Instructional Design, & Assessment. (2019). *ACCESS for ELLs: Interpretive guide for score reports, grades k–12 spring 2020.* Retrieved from https://wida.wisc.edu/sites/default/files/resource/Interpretive-Guide.pdf

APPENDICES

Appendix A

Table A1 provides a list of the variables used for this study from the NAEP 2017 data.

Table A1
Description of variables to evaluate the effects of ETA

Variable	Dataset	Description
ETA Eligible	LEP	Whether this student is ELL or not
ETA Received		Whether this student receives ETA or not; constructed by two variables:
	ACCOM2	Whether this student receives any types of accommodations
	ACCEXT	Whether this student receives ETA
ELL EP		Discrete ELL English proficiency with 6 levels: No Proficiency, ELL Beginning, ELL Intermediate, ELL Advanced, Formerly ELL, and Never ELL; constructed by two variables:
	ELL	Student has limited English proficiency
	XL04303	Student's English proficiency: Reading English
Proficiency	MTHCM1-20	20 plausible values of math proficiency
	RRPCM1-20	20 plausible values of reading proficiency
SD	SD3	Whether this student has disabilities or not
Gender	DSEX	Whether this student is male or female
Race/Ethnicity	SDRACEM	Student' race/ethnicity: White, Black, Hispanic, Asian/Pacific Islander, American Indian/Alaska Native, and Unclassified
Free Lunch	SLUNCH1	Student's eligibility for National School Lunch Program: eligible, not eligible, and info not available
ELL Grade Level	XL04202	ELL's grade level of performance in NAEP subject: at/above, 1 year below, 2 or more years below, no instruction, and I don't know
English Instruction	XL04101	How long this student has been receiving instruction in English: No instruction in English, less than 1 year, 1-2 years, 2-3 years, 3 years or more, and I don't know
US School	XL04801	How long this student has been in US schools: 1 year or more and less than 1 year
Primary Language	XL04601	Student's primary language: Spanish and Other

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

Table A2 summarizes the covariate distributions depending on students' eligible status and the receipt status of ETA in mathematics. The responses of English instruction, US school, and primary language were measured from the ELL contextual questionnaire.

Table A2
Descriptive statistics

	Eligible status		Treatment receipt status	
	ELL	Non-ELL	Treated	Untreated
Variable	N	N	N	N
Total	4,940	111,970	1,640	115,270
Gender: Male	2,620	54,310	860	56,070
Race: White	290	60,000	80	60,200
Race: Black	190	21,220	50	21,360
Race: Hispanic	3,720	19,010	1,320	21,410
Race: Asian/PI	640	5,080	160	5,560
Race: AI/AN	80	1,990	10	2,050
Free Lunch: Eligible	4,130	57,540	1,420	60,260
Eng. Instr: No	320	—	110	220
Eng. Instr.: < 1 yr	220	—	120	100
Eng. Instr.: 1-2 yrs	460	—	210	250
Eng. Instr.: 2-3 yrs	450	—	200	260
Eng. Instr.: > 3 yrs	3,410	—	960	2,450
Eng. Instr.: don't know	80	—	20	60
US School : < 1 yr	60	—	30	30
US School : ≥ 1 yr	4,890	—	1,580	3,310
Primary language: Spanish	3,520	—	1,260	2,260
Primary language: Other	1,420	—	340	1,070

— Not available.
NOTE: English Instruction, US School, and Primary languages are from an ELL questionnaire. Numbers *N* are rounded to nearest tens, and details may not sum to a total due to rounding. The percentages in parentheses are calculated based on unrounded numbers.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

Appendix B

We use an instrumental variable regression or two-stage least squares (TSLS) regression to estimate LATE. We run the model for Z_{ijk} (the endogenous regressor) as the first-stage regression that regresses Z_{ijk} on instrument A_{ijk} and all the other covariates. Then, we run the following model as the second-stage regression:

$$Y_{ijk} = \gamma_0 + \gamma_1 Z_{ijk} + \gamma_2 (X_{ijk} - x_c) + \gamma_3 A_{ijk} (X_{ijk} - x_c) + \sum \gamma_w W_{ijk} + s_j + u_k + \epsilon_{ijk} \quad (1.3)$$

We extract predicted values of Z_{ijk} from model (2) and substitute them into model (3) as is standard in the TSLS procedure. Then, we can estimate $\gamma_1 = \tau_{LATE}(x_c)$ that represents the LATE at the cutoff of "ELL Advanced", that is, the effect of receiving ETA among the complier students in the "ELL Advanced" category.

Appendix C

Table C1 provides the compliance rate by ELL English proficiency categories.

Table C1
Compliance by ELL English Proficiency Categories

Eligibility	Non-Received	Received
No Proficiency	10	#
ELL Beginning	200	120
ELL Intermediate	1,040	570
ELL Advanced	2,090	920
Formerly ELL	1,980	#
Never ELL	109,950	30

NOTE: Numbers are rounded to nearest tens. #s are rounds to zero. Details may not sum to a total due to rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2017.

Appendix D

The data generating model for three-level data is stated below.

1. For each level-2 cluster $j = 1, \ldots, J_k$ (e.g., classes) of level-3 cluster $k = 1, \ldots, K$ (e.g., schools), generate number of individuals per cluster n_{jk} by drawing a number from a normal distribution with mean I and standard deviation sd and rounding it to the nearest integer.

2. For each individual $i = 1, \ldots, n_{jk}$ in level-2 cluster j and level-3 cluster k, generate level-3, level-2, and level-1 covariates $\mathbf{Q}_k = (Q_{1k}, Q_{2k})$, $\mathbf{W}_{jk} = (W_{1jk}, W_{2jk})$ and $\mathbf{X}_{ijk} = (X_{1ijk}, X_{2ijk})$ as follows.

$$Q_{1k} \sim U[0,1],\ Q_{2k} \sim U[0,1]$$

$$\begin{pmatrix} W_{1jk} \\ W_{2jk} \end{pmatrix} \sim N\left[\begin{pmatrix} 0.1Q_{1k} + 0.05Q_{2k} + \kappa_{1k} \\ 0.08Q_{1k} + 0.1Q_{2k} + \kappa_{2k} \end{pmatrix},\ \begin{pmatrix} 2 & .2 \\ .2 & 2 \end{pmatrix}\right]$$

$$\begin{pmatrix} X_{1ijk} \\ X_{2ijk} \end{pmatrix} \sim N\left[\begin{pmatrix} 0.1W_{1jk} + 0.05W_{2jk} + 0.1Q_{1k} + 0.02Q_{2k} + \kappa_{1jk} \\ 0.08W_{1jk} + 0.1W_{2jk} + 0.05Q_{1k} + 0.01Q_{2k} + \kappa_{2jk} \end{pmatrix},\ \begin{pmatrix} 10 & 2 \\ 2 & 15 \end{pmatrix}\right]$$

$$\begin{pmatrix} \kappa_{1k} \\ \kappa_{2k} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}\right], \qquad \begin{pmatrix} \kappa_{1jk} \\ \kappa_{2jk} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}\right]$$

Note that level-2 covariates form a hierarchical model with level-3 covariates Q_{1k}, Q_{2k}, random errors κ_{1k}, κ_{2k}, and random variances. Similarly, the means of level-1 covariates are a function of upper-level covariates $W_{1jk}, W_{2jk}, Q_{1k}, Q_{2k}$.

3. Generate individual treatment status Z_{ijk} from the following random-effects logistic propensity score model.

$$logit(e_{ijk}) = -0.2 + 0.1X_{1ijk} + 0.03X_{2ijk} + 0.1W_{1jk} + 0.08W_{2jk} + 0.1Q_{1k} + 0.05Q_{2k} + R_{jk}^{W} + R_{k}^{Q}, \qquad R_{jk}^{W} \sim N(0,1), \qquad R_{k}^{Q} \sim N(0,1)$$

$$Z_{ijk} \sim Bernoulli(e_{ijk})$$

where e_{ijk} is the propensity score for individual i in level-2 cluster j and level-3 cluster k, and R_{jk}^{W} and R_{k}^{Q} are normally distributed random effects for level-2 clusters and level-3 clusters, respectively.

4. Generate the potential outcomes $Y_{ijk}(1)$, $Y_{ijk}(0)$ and observed outcome Y_{ijk} from a random-effects linear regression model.

$$Y_{ijk}(z) = 100 + 2 \cdot z + 2X_{1ijk} + 1X_{2ijk} + 2W_{1jk} + 1.5W_{2jk} + 1Q_{1k} + 0.5Q_{2k} + U_{jk}^{W} + U_{k}^{Q} + \epsilon_{ijk}, \quad U_{jk}^{W} \sim N(0,10), \quad U_{k}^{Q} \sim N(0,7), \quad \epsilon_{ij} \sim N(0,100)$$

$$Y_{ijk} = Z_{ijk}Y_{ijk}(1) + \left(1 - Z_{ijk}\right)Y_{ijk}(0)$$

where U_{jk}^{W} and U_{k}^{Q} are normally distributed random effects for level-2 cluster j of level-3 cluster k, respectively. ϵ_{ijk} is the random error for individual i in level-2 cluster j of level-3 cluster k.

Appendix E

The data generating model for cross-classified data is stated below.

1. For each factor-1 cluster $j = 1, \ldots, J$, generate number of individuals (e.g., students) per cluster n_j (e.g., schools) by drawing a number from a normal distribution with mean nI and standard deviation *sd* and rounding it to the nearest integer.

2. Create factor-2 Cluster labels $k = 1, \ldots, K$, (e.g., neighborhoods) according to Meyers and Beretvas (2006). Here, suppose there is no correlation between residuals and there are three feeders from one factor to the other factor. Each factor-1 cluster sends 70% of its individuals to the most adjacent factor-2 cluster, 15% of its individuals to the next closest factor-2 cluster, and 15% of its individuals to the third closest. The first factor-1 cluster sends 80% of its individuals to the most adjacent factor-2 cluster, and the rest of 20% to the next closest. Individuals of the last factor-1 cluster are distributed in the same way as the first factor-1 cluster.

3. For each individual $i = 1, \ldots, n_{(jk)}$ in factor-1 cluster j and level-3 cluster k, generate level-3, level-2, and level-1 covariates $\mathbf{W}_j = (W_{1j}, W_{2j})$, $\mathbf{Q}_k = (Q_{1k}, Q_{2k})$, and $\mathbf{X}_{i(jk)} = (X_{1i(jk)}, X_{2i(jk)})$ as follows.

$$\begin{pmatrix} W_{1j} \\ W_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 2 & .2 \\ .2 & 2 \end{pmatrix}\right], \quad \begin{pmatrix} Q_{1j} \\ Q_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}\right]$$

$$\begin{pmatrix} X_{1i(jk)} \\ X_{2i(jk)} \end{pmatrix} \sim N\left[\begin{pmatrix} 0.1W_{1j} + 0.05W_{2j} + \kappa_{1j} \\ 0.08W_{1j} + 0.1W_{2j} + \kappa_{2j} \end{pmatrix}, \quad \begin{pmatrix} 10 & 2 \\ 2 & 15 \end{pmatrix}\right]$$

$$\begin{pmatrix} \kappa_{1j} \\ \kappa_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}\right]$$

Note that individual-level covariates form a hierarchical model with factor-1 covariates W_{1j}, W_{2j}, random errors κ_{1j}, κ_{2j}, and random variances.

4. Generate individual treatment status $Z_{i(jk)}$ from the following cross-classified random-effects logistic propensity score model.

$$logit\left(e_{i(jk)}\right) = -0.2 + 0.1X_{1i(jk)} + 0.03X_{2i(jk)} + 0.1W_{1j} + 0.08W_{2j} + 0.1Q_{1k} + 0.05Q_{2k} + R_j^W + R_k^Q, \qquad R_j^W \sim N(0,1), \qquad R_k^Q \sim N(0,0.5)$$

$$Z_{i(jk)} \sim Bernoulli\left(e_{i(jk)}\right)$$

where $e_{i(jk)}$is the propensity score for individual i in cluster j. R_j^W and R_k^Q are normally distributed random effects for Factor-1 cluster j and Factor-2 cluster k, respectively.

5. Generate the potential outcomes $Y_{i(jk)}(1)$, $Y_{i(jk)}(0)$ and observed outcome $Y_{i(jk)}$ from a cross-classified random-effects linear regression model.

$$Y_{i(jk)}(z) = 100 + 2 \cdot z + 2X_{1i(jk)} + 1X_{2i(jk)} + 2W_{1j} + 1.5W_{2j} + 1Q_{1k} + 0.5Q_{2k} + U_j^W + U_k^Q + \epsilon_{i(jk)}, \qquad U_j^W \sim N(0,10), \qquad U_k^Q \sim N(0,7), \qquad \epsilon_{i(jk)} \sim N(0,100)$$

$$Y_{i(jk)} = Z_{i(jk)}Y_{i(jk)}(1) + \left(1 - Z_{i(jk)}\right)Y_{i(jk)}(0)$$

where U_j^W and U_k^Q are normally distributed random effects for factor-1 cluster j and factor-2 cluster k, respectively. $\epsilon_{i(jk)}$ is the random error for individual i in factor-1 cluster j and factor-2 cluster k.

Appendix F

We compared the performance between our proposed sequential/two-step approach and an alternative "covariate" approach where the estimated class membership is used as an additional covariate. Specifically, the proposed approach implements ML methods within each latent class and estimates the ATE within each class, while the covariate approach includes the estimated class membership variable as another covariate in ML methods and estimates the conditional ATE defined by this covariate. We suspect that under some assumptions, both are asymptotically equivalent, but they may have different finite-sample properties and we investigate them through a small simulation study below.

Table F1 provides the performance of class-specific ATE estimates between the two approaches using the data generating model from the main text. We saw that biases from our two-step approach were generally smaller than those from the covariate approach. But as the cluster size increased, biases became smaller for both approaches and the differences between the two were generally negligible. Also, the MSEs were consistently smaller under our approach across different cluster sizes. While the simulation study is small, the result gives some confidence that our approach outperforms the covariate approach in terms of finite-sample bias and MSE.

Table F1

Comparison between the proposed sequential approach and the covariate approach

	Latent Class 1		Latent Class 2	
(nC1, nC2, nS)	|Bias|	MSE	|Bias|	MSE
Our Approach				
(25, 25, 50)	0.141	0.538	0.088	0.403
(25, 25, 100)	0.039	0.259	0.073	0.186
(25, 25, 200)	0.001	0.111	0.018	0.101
(25, 25, 400)	0.014	0.062	0.013	0.051
Covariate Approach				
(25, 25, 50)	0.669	0.844	0.711	0.828
(25, 25, 100)	0.562	0.560	0.525	0.470
(25, 25, 200)	0.309	0.247	0.210	0.179
(25, 25, 400)	0.111	0.092	0.075	0.067

Appendix G

Table G1

Performance of class-specific treatment effect estimates: estimated class membership

	Latent Class 1		Latent Class 2		Overall	
(nC1, nC2, nS)	|Bias|	MSE	|Bias|	MSE	|Bias|	MSE
(25, 25, 30)						
IPW	0.009	1.825	0.802	1.762	0.405	1.790
MMW-S	0.228	1.920	0.788	1.840	0.280	1.876
DR	0.271	1.393	0.832	1.896	0.280	1.640
Hybrid Causal Forests	0.373	1.302	0.762	1.746	0.194	1.521
Hybrid BART	0.482	1.274	0.751	1.660	0.134	1.464
Hybrid TMLE	0.272	1.150	0.757	1.756	0.242	1.449
(25, 25, 50)						
IPW	0.004	1.039	0.469	0.944	0.236	0.988
MMW-S	0.186	0.914	0.441	0.849	0.127	0.879
DR	0.174	0.803	0.472	0.967	0.148	0.882
Hybrid Causal Forests	0.233	0.719	0.399	0.739	0.082	0.727
Hybrid BART	0.297	0.660	0.411	0.762	0.057	0.710
Hybrid TMLE	0.179	0.638	0.427	0.802	0.123	0.718
(50, 50, 30)						
IPW	0.043	0.818	0.559	0.797	0.258	0.806
MMW-S	0.416	1.074	0.541	0.790	0.063	0.931
DR	0.343	0.645	0.594	1.014	0.126	0.830
Hybrid Causal Forests	0.387	0.696	0.533	0.762	0.073	0.728
Hybrid BART	0.455	0.689	0.519	0.697	0.032	0.692
Hybrid TMLE	0.337	0.592	0.525	0.743	0.095	0.667

Note: nC1, nC2, and nS represent the number of clusters for the first latent class, the number of clusters for the second latent class, and average cluster sizes, respectively. IPW represents inverse-propensity weighting, and MMW-S represents marginal mean weighting through stratification. DR represents a doubly robust estimator. BART represents Bayesian additive regression trees, and TMLE represents targeted maximum likelihood estimation. The true treatment effect values are 2.5 and 0 for the first and second latent classes, respectively.

Appendix H

We replicated the simulation study from the main text and measured the performance of class-specific treatment effect estimates when we use the true class labels. In this setting, all the parametric methods have correctly specified outcome and propensity score models and should perform well. Specifically, we expect the DR estimator to perform the best followed by the non-DR estimators (IPW and MMW-S). Finally, we expect the performance of ML methods to be somewhere in between the performance of the DR and non-DR estimators, but the performance of ML methods will become similar to the performance of the DR estimator as the sample size increases.

Table H1 shows the results. As expected, the DR estimators performed best in terms of bias and MSE. The non-DR estimators—IPW and MMW-S—performed worse than the DR estimator and ML methods with one exception: absolute bias of the MMW-S in the sample size condition (25, 25, 50). Also, similar to what we observed in the main text, we saw that the non-DR estimators, especially the IPW estimator, achieved more bias reduction in one latent class over another latent class in the current data generating model, but ended up having a relatively large amount of overall bias. In contrast, DR and hybrid methods achieved bias reduction in both latent classes and had overall bias reductions. Finally, when the sample size increased, the performance of ML methods was competitive to the performance of the DR estimator.

Table H1
Performance of class-specific treatment effect estimates: true class membership

	Latent Class 1		Latent Class 2		Overall	
(nC1, nC2, nS)	|Bias|	MSE	|Bias|	MSE	|Bias|	MSE
(25, 25, 30)						
IPW	0.371	1.521	0.008	0.586	0.182	1.054
MMW-S	0.138	1.455	0.064	0.625	0.101	1.040
DR	0.032	0.732	0.018	0.595	0.007	0.663
Hybrid Causal Forests	0.065	0.848	0.010	0.634	0.028	0.741
Hybrid BART	0.048	0.736	0.023	0.595	0.036	0.665
Hybrid TMLE	0.114	0.741	0.007	0.609	0.054	0.675
(25, 25, 50)						
IPW	0.227	0.964	0.003	0.342	0.115	0.652
MMW-S	0.005	0.787	0.025	0.349	0.016	0.568
DR	0.037	0.601	0.000	0.344	0.018	0.473
Hybrid Causal Forests	0.063	0.553	0.003	0.353	0.030	0.453
Hybrid BART	0.001	0.524	0.008	0.343	0.005	0.434
Hybrid TMLE	0.086	0.533	0.007	0.347	0.046	0.440
(50, 50, 30)						
IPW	0.313	0.747	0.025	0.281	0.144	0.514
MMW-S	0.222	0.737	0.083	0.297	0.153	0.517
DR	0.019	0.374	0.034	0.269	0.027	0.322
Hybrid Causal Forests	0.072	0.417	0.038	0.289	0.017	0.353
Hybrid BART	0.034	0.384	0.042	0.282	0.038	0.333
Hybrid TMLE	0.046	0.362	0.034	0.279	0.006	0.320

Note: nC1, nC2, and nS represent the number of clusters for the first latent class, the number of clusters for the second latent class, and average cluster sizes, respectively. IPW represents inverse-propensity weighting, and MMW-S represents marginal mean weighting through stratification. DR represents a doubly robust estimator. BART represents Bayesian additive regression trees, and TMLE represents targeted maximum likelihood estimation. The true treatment effect values are 2.5 and 0 for the first and second latent classes, respectively.

Appendix I

We also assessed the performance of ML methods based on the root mean squared error (RMSE) of estimating individual CATE estimates in simulation replication m, denoted as $\hat{\tau}_{ij,m}(k)$. Specifically, let N_m denote the sample size of each simulation replication. We evaluated the following quantity:

$$RMSE_m(k) = \sqrt{\frac{1}{N_m}\sum_{ij}(\hat{\tau}_{ij,m}(k) - \tau_{ij}(k))^2}$$

and took averages of $RMSE_m(k)$ across simulation replicates.

Figure I1 summarizes the performance of individual CATE estimates within each latent class across different ML methods. Though TMLE tends to have slightly large RMSEs, the performance across methods was comparable and our simulation results in the main manuscript are generally not sensitive to the choice of ML methods.

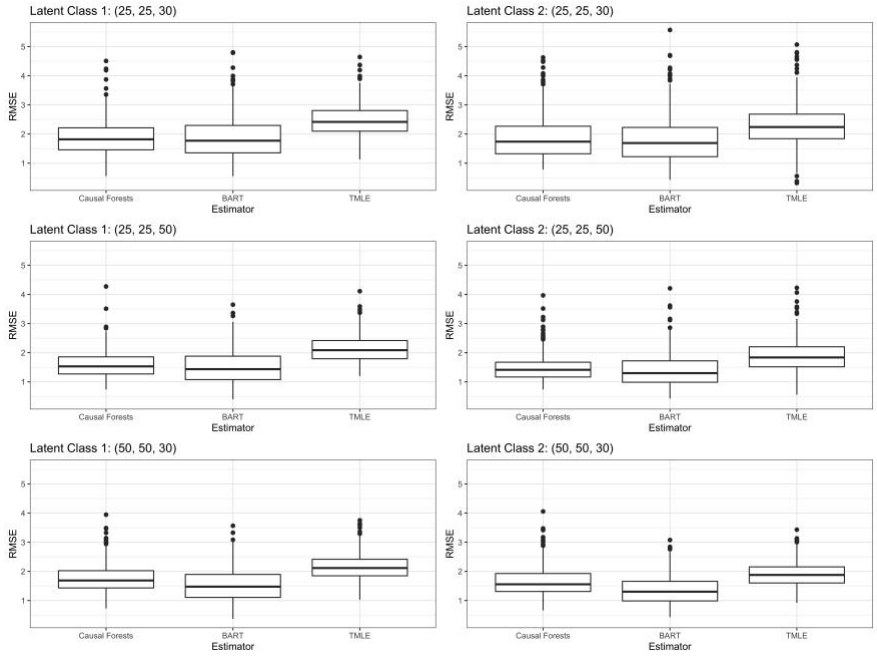

Figure I1. Performance of individual CATE estimates: root mean squared error.

Appendix J

Figure J1 displays the distributions of individual CATE estimates when BART and TMLE are used. Our hybrid approach with different ML methods produced similar estimates of $\tau(k)$; the dotted lines across methods align closely with each other. However, there were some differences in the distributions of the individual CATE estimates, with BART producing "spiky" distributions, whereas TMLE producing smoother distributions. This suggests that different ML methods make different assumptions about how to locally smooth across the observed covariates.

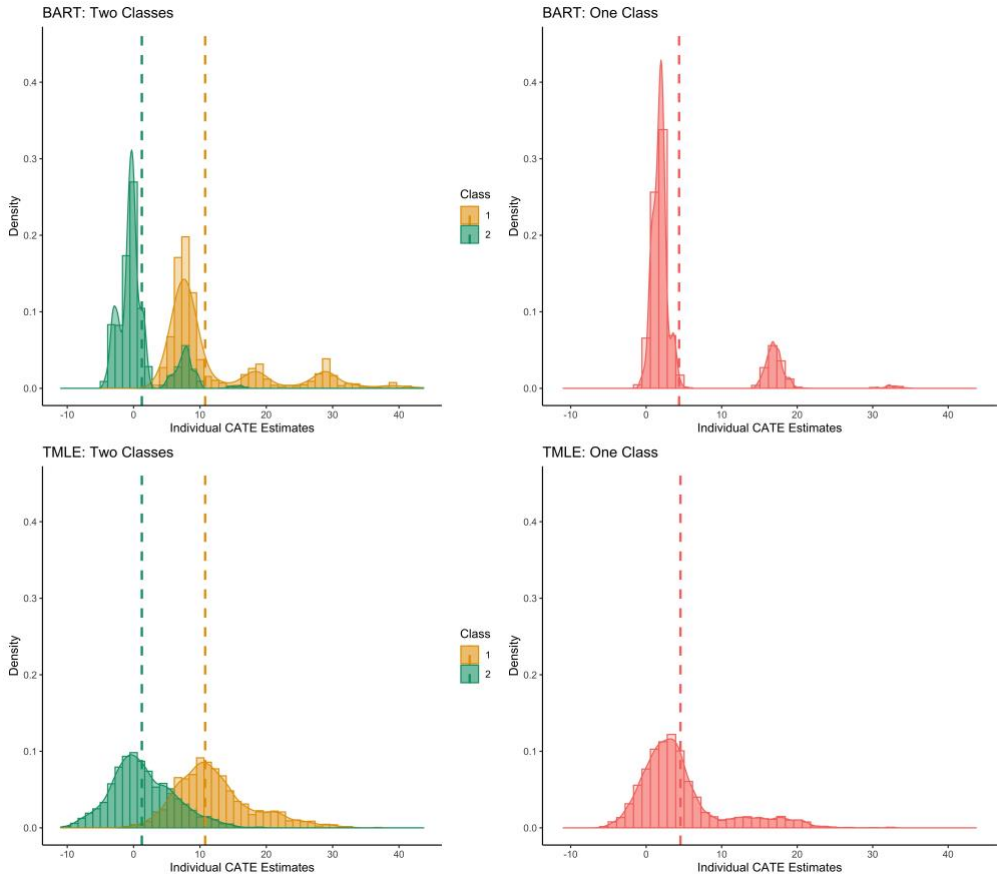

Figure E1. Distributions of individual CATE estimates with BART and TMLE. The left shows hybrid ML methods discovering two latent classes, while the right shows the usual ML methods without latent classes. Dashed lines represent class-specific treatment effect estimates.

Appendix K

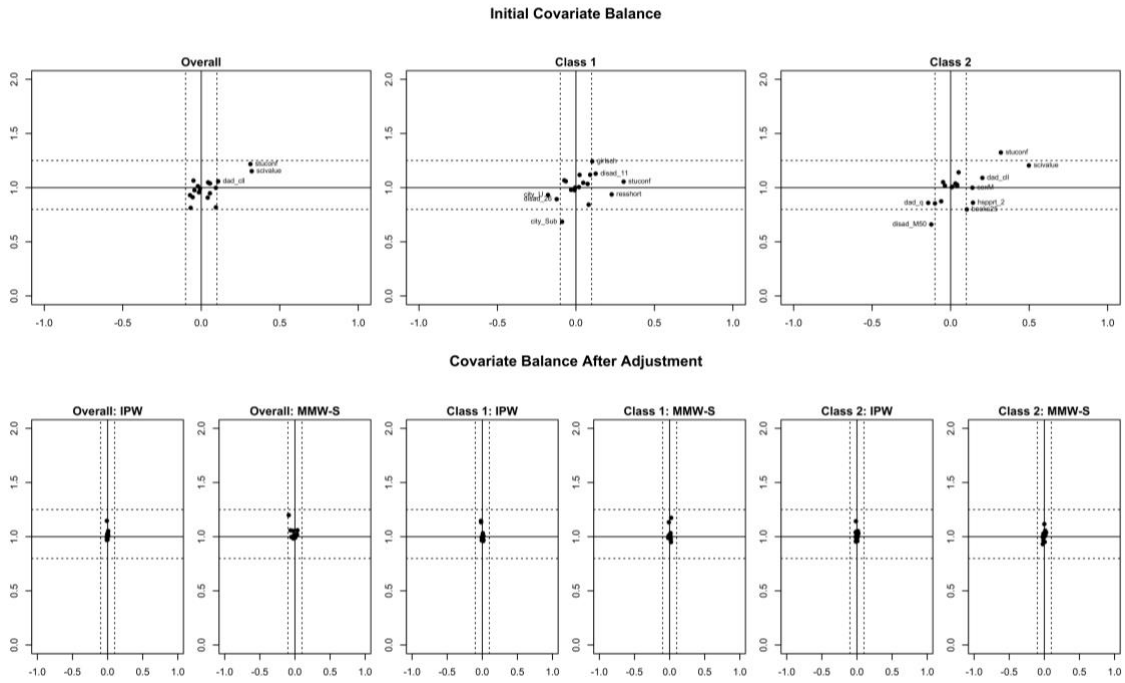

Figure K1. Covariate balance plots before and after propensity score adjustment (Standardized mean differences on the x-axis and variance ratios on the y-axis)

We checked covariate balance in within-class matching by computing the absolute standardized mean differences and variance ratios between treated units and control units. As a rule of thumb, if the mean difference of each covariate is less than 0.1 standard deviation and the variance ratio is more than 4/5 and less than 5/4, we can provide evidence for good balance of the covariates. Figure F1 displays covariate balance plots before and after propensity score adjustment for two classes as well as for one homogeneous class. One homogeneous class assumed no subpopulations or multiple latent classes, and its propensity scores were estimated via random effects logistic regression. There was less initial imbalance in covariates for one homogeneous class, and we achieved good covariate balance between the treated and untreated groups after applying IPW and MMW-S. For the two-class approach, the covariates imbalanced differed in

each class. However, after applying IPW and MMW-S, we achieved acceptable covariate balance within each class.