Heterogeneous Subgroup Identification in Observational Studies

Jianshen Chen and Bryan Keller

College Board and Columbia University

Draft – Manuscript Under Review

Author Note

The authors contributed equally. Part of this work was carried out when Jianshen Chen was employed by Educational Testing Service.

Abstract

Despite the overwhelming focus on the overall average treatment effect in the methodological and statistical literature, in many cases the efficacy of an educational program or intervention might vary based on unit background characteristics. The identification of subgroups for which an educational intervention is particularly effective or, on the other hand, has no effect or is possibly harmful, may have important practical implications, especially in terms of allocation of resources. We propose a five-step approach using propensity score matching and regression trees to identify subgroups with heterogeneous treatment effects in observational studies. Results of two Monte Carlo simulation studies demonstrate that the proposed approach can accurately identify heterogeneous subgroups while maintaining Type I error rate. In a case study with ECLS-K data we find that the effect of exposure to special education services on 5th grade mathematics achievement varies based on kindergarten mathematics achievement and student gender.

Heterogeneous Subgroup Identification in Observational Studies

Introduction

Much effort has been devoted to the identification and estimation of overall average treatment effects with observational data through conditioning strategies such as propensity score analysis, nonparametric regression estimation, or doubly robust combinations of selection and outcome models. Despite the overwhelming focus on the overall average treatment effect in the methodological and statistical literature, in many cases the efficacy of an educational program or intervention might vary based on unit background characteristics. The identification of subgroups for which a particular educational intervention is particularly effective or, on the other hand, has no effect or is possibly harmful, may have important practical implications, especially in terms of allocation of resources. Recent advances in statistics and causal inference have focused on understanding such *conditional* average treatment effects. Nevertheless, methods for detecting and describing treatment effect heterogeneity with observational data have not entered the mainstream in education or behavior sciences. In this paper, we describe an exploratory approach for identification of treatment effect heterogeneity based on propensity score matching, a method that is familiar to many educational researchers. We then use regression trees, a simple nonparametric regression technique, to identify subgroups based on background characteristics that best explain how units respond to treatments.

The organization of this paper is as follows. After a brief review of propensity score analysis and regression trees, we describe the proposed approach for subgroup identification. Next, we run a Monte Carlo simulation study to examine (a) the sensitivity and specificity of the method for detecting treatment effect heterogeneity and (b) the accuracy of estimated conditional average effects. We then apply the approach in a case study to search for heterogeneous subgroups in the context of an observational study designed to examine the average causal effect of special education on mathematics achievement in fifth grade. We conclude with some discussion.

Propensity Score Analysis

In this paper, we consider the simple case of a binary treatment condition, in which individuals self-select into a “treatment group" or a “control group". Following the general

notation of the Rubin causal model (Rubin, 1974), let Y_i^1 and Y_i^0 be the potential outcomes for individual i under the treatment ($T_i = 1$) and control ($T_i = 0$) conditions. The propensity score for unit i, given a vector of observed covariates X, is defined as the conditional probability that a subject is assigned to the treatment group, $e_i(X) = P(T_i = 1|X_i)$. The average treatment effect (ATE) for the population is defined as the average of the individual treatment effects: ATE $= E(Y_i^1 - Y_i^0) = E(Y_i^1)$ - $E(Y_i^0)$. Conditional average treatment effects may be defined as average treatment effect for units with a particular value, or ranges of values, on one or more covariates: CATE $= E(Y_i^1 - Y_i^0|X_i = x_i)$.

For identification of the ATE and CATE, propensity score techniques and other conditioning strategies rely on an important and multipart assumption known as *strong ignorability* (Rosenbaum & Rubin, 1983):

1. *ignorability*: the potential outcomes are independent of the treatment assignment given observed covariates X; that is, $\{Y^0, Y^1\} \perp\!\!\!\perp T|X$,

2. *reliable measurement*: observed covariates X have been reliably measured (Steiner, Cook, & Shadish, 2011), and

3. *positivity*: the propensity score for each unit lies strictly between zero and one; that is, $0 < e_i(X) < 1$ for all i.

When the strong ignorability assumption holds, participants with the same propensity score have identical distributions on the set of covariates; that is, $X \perp\!\!\!\perp T|e(X)$. Thus, by conditioning on the propensity score, rather than the full set of covariates, it is possible to restore covariate balance across treatment and control groups to what would have been expected under random assignment. Furthermore, under strong ignorability, the potential outcomes are independent of treatment assignment, $\{Y^0, Y^1\} \perp\!\!\!\perp T|e(X)$, so that unbiased treatment effect estimates can be obtained.

Methods for conditioning on the propensity score include stratification, inverse weighting, matching and regression adjustment (see, e.g., Stuart, 2010 for a review). In this paper, we focus on propensity score matching because it stratifies the units into many groups, within which the individual treatment effect may be approximated. We use optimal full matching (Rosenbaum,

2002) to create strata because (a) it performs well when compared with other propensity score conditioning strategies (Kaplan & Chen, 2012; Steiner & Cook, 2013) and (b) it typically results in many relatively small strata. Having units grouped in many small strata is helpful for preserving important variability in imputed potential outcomes that would be lost if strata were too large.

Tree-based Methods

Regression Trees

A *regression tree* is an algorithmic tool invented by Breiman, Friedman, Olshen, and Stone (1984) that models the relationship between an outcome variable, Y, and predictors, $X_1, \ldots, X_p$, by iteratively splitting the units into subgroups based on predictor values. Every split creates two subgroups, called *nodes*, and any node that is unsplit is called a *terminal node*. Within each terminal node, the tree-predicted value for each member of the node is simply the mean outcome score for all units in the node. Thus, for unit i in terminal node t, where N_t represents the set of units in t, the tree-predicted value for unit i is $\hat{Y}_i = \frac{1}{|N_t|} \sum_{i \in N_t} Y_i$. At each iteration, the deviance, $\sum_i \left(\hat{Y}_i - Y_i\right)^2$, is determined before splitting. Then, every possible split on every variable is considered, and the split that results in the largest decrease in deviance is selected.

If left unchecked, regression trees would continue to split until each terminal node contained only one point, yielding a perfect fit to the data. To prevent overfitting of this sort, which would result in poor prediction accuracy, a stopping rule is imposed. The usual approach for determining the stopping rule is based on adding a penalty to the deviance based on the number of terminal nodes in the tree. This approach, referred to as *cost-complexity pruning*, is implemented in the `rpart` package (Therneau, Atkinson, & Ripley, 2015) in R (R Core Team, 2018), which we use to fit regression trees. The implementation of cost-complexity pruning requires an additional tuning parameter, which is typically selected through cross-validation.

Because regression trees model a response surface with discrete jumps, they can only roughly approximate a smooth surface, which results in bias. While it is possible to lower the minimum number of units per terminal node to get a better fit, this leads to overfitting and results in higher variance. More complex tree-based methods overcome this shortfall by pooling

results based on many trees. Bootstrap aggregation (bagging), random forests, and boosted trees all benefit from the nonparametric flexibility of regression trees, but combine them in unique ways to improve the predictive performance.

The main advantage to using a single regression tree over other, more complex methods such as those mentioned above, is that the output from a single tree is easily interpretable because each split results in a partition on some variable. Thus, for meaningfully segmenting a sample based on variable characteristics (which is our goal in identifying heterogeneous subgroups), single trees are a useful, if somewhat crude, tool.

Random Forests

Random forests were invented by Breiman (2001). The goal of random forests is to turn what would otherwise be a rather weak learner (a single tree) into a strong learner by repeatedly fitting individual regression trees and combining them in an intelligent way. Random forests are based on aggregated results from fitting regression trees to B bootstrap samples of the data. A regression tree is grown for each bootstrap sample by recursively splitting until each terminal node is as small as possible without being smaller than a prespecified minimum node size. Importantly, each split uses a random sample of only m predictors, where $m << p$, the total number of predictors. The resultant set of trees constitute the "random forest". The out-of-bag (OOB) cases for replication b are the cases that were not part of the bth bootstrap sample. Regression predictions are calculated as the average of the predicted values across the B trees.

For our purposes, random forests are useful because they provide a quantitative summary of each variable's importance in predicting the outcome. This metric proves extremely useful in winnowing down a large set of covariates (such as the 34 in our case study) to a more manageable number of important variables on which we can search for heterogeneous treatment effects. We use a permutation-based approach for variable selection with random forest variable importance described in Keller & Zhang, 2018, March.

Identification of Subgroups with Heterogeneous Treatment Effects

Effect heterogeneity is characterized by interactions between one or more covariates and the treatment variable. Regression trees handle interactions in their naïve implementation, so it is not

surprising that tree-based methods have played a prominent role in the literature.

Randomized Trials

In particular, a number of tree-based methods have been proposed to identify treatment effect heterogeneity in randomized experiments. Su, Tsai, Wang, Nickerson, and Li (2009) used regression trees with a customized loss function designed to measure the strength of treatment by subgroup interactions. Foster and colleagues (Foster, Taylor, & Ruberg, 2011; Foster, Taylor, Kaciroti, & Nan, 2015) used random forests to identify covariate regions for binary and continuous outcomes. Imai and Strauss (2011) fit regression trees followed by Bayesian models to estimate group-specific treatment effects.

Approaches for identifying effect heterogeneity in randomized experiments based on methods other than trees have been proposed as well. Imai and Ratkovic (2013) introduced an algorithm based on regularized support vector machines. Ding, Feller, and Miratrix (2016) developed a randomization approach to evaluate effect variations that are unexplained by observed covariates. In a 2017 issue of the *Journal of Research on Educational Effectiveness* devoted to variation in treatment effects in randomized trials, Bloom, Raudenbush, Weiss, and Porter (2017) used mixed effects models to study cross-site variation in program efficacy and Harding, Morris, and Hill (2017) demonstrated the use of principal stratification to identify local average treatment effects.

Observational Studies

Despite the aforementioned work in the context of randomized experiments, fewer studies have aimed to identify effect heterogeneity in observational study settings. Xie, Brand, and Jann (2012) proposed several methods to detect treatment by propensity score interactions, but did not consider heterogeneity based on individual covariates. Hsu, Small, and Rosenbaum (2013) and Hsu, Zubizarreta, Small, and Rosenbaum (2015) developed a framework for sensitivity analysis in the presence of effect heterogeneity by identifying subgroups of matched-pairs formed using exact matching and leveraged the method for identification of heterogeneity. Exact matching is ideal because it requires no functional form or parametric assumptions; however, with many covariates or a handful of continuously measured covariates, exact matching is not feasible because the

number of possible combinations of covariate values grows exponentially with the addition of each new covariate.

Wager and Athey (2018) derived asymptotic results for random forests and applied them to forests based on trees that are tuned for treatment effect estimation. A strength of their causal forest algorithm, which directly estimates CATEs, is the recourse to asymptotically-valid confidence intervals for treatment effect estimates. Like other methods that have been used to estimate individual treatment effects, such as Bayesian additive regression trees (BART; Hill, 2011), causal forests do not explicitly address identification of heterogeneous subgroups. The output must be further processed to identify groups for which the treatment is differentially effective.

The Proposed Method

In this paper, we propose and evaluate an approach for heterogeneous subgroup identification with observational data that combines propensity score matching and regression trees. First, optimal full propensity score matching is used to estimate individual treatment effects within matched strata. Under optimal full matching, each stratum will necessarily have at least one treated and at least one comparison unit. If ignorability is satisfied with the set of conditioning variables and the propensity scores are estimated well, then units will be balanced within matched subgroups (recall, under ignorability, $X \perp\!\!\!\perp T|e(X)$). Thus, we propose impute missing potential outcomes within strata as follows. Let N_{T_s} and N_{C_s} be the set of indexes for treated units and comparison units, respectively, in stratum s, and let $|N_{T_s}|$ and $|N_{C_s}|$ be their respective cardinalities. For each $i \in N_{T_s}$, define the imputed value for the unobserved potential outcome under assignment to the comparison group as

$$\hat{Y}_i^0 = \frac{1}{|N_{C_s}|} \sum_{j \in N_{C_s}} Y_j^0.$$

That is, the unobserved potential outcome for each treated unit in stratum s is estimated by the average of the observed outcomes for the comparison units in the same stratum. The individual treatment effect (ITE) for each unit i is estimated as $Y_i^1 - \hat{Y}_i^0$. The imputed potential outcomes for the comparison units in stratum s are defined analogously.

The goal of heterogeneous subgroup identification, by definition, is to discover subsets of the covariate space for which the average effects of the intervention differ. A straightforward way to operationalize this is to examine which covariates, if any, are related to the ITEs. Regression trees are particularly well-suited for this task because they partition the covariate space based on the strength of covariate-outcome relationships. Important partitions identified by regression trees provide evidence for heterogeneous subgroups.

Criticisms have characterized the practice of subgroup identification as suffering from a lack of planned statistical approaches that are both valid and sufficiently powered (e.g., Assman, Pocock, Enos, & Kasten, 2000). In particular, this criticism is salient with respect to regression trees because, while the use of cross-validation to select tuning parameters optimizes out of sample predictive performance, it does not address validity concerns, per se. To that end, we propose the systematic use of permutation-based Type I error rate studies to ensure valid inferences. The Type I error rate for subgroup detection with a particular cost-complexity tuning parameter value, *cp*, may be estimated as follows.

1. Create R copies of the data set, each with identical covariates and each with permuted yoked outcome/treatment data. By permuting yoked outcome/treatment data, average treatment effects remain unperturbed, while all covariate relationships with outcome and treatment are destroyed.

2. For each of the R data sets, estimate propensity scores, create strata via optimal full matching, impute missing potential outcomes, and run a regression tree with cost-complexity parameter value fixed to *cp*.

3. Each replication for which the *cp*-pruned regression tree splits on any variable is a Type I error. The total number of Type I errors divided by R gives the approximate level of the test.

Binary search may be used to iteratively close in on a *cp* value that approximates a desired nominal Type I error rate. Although permutation testing has been proposed as a primary method for detecting treatment effect heterogeneity in randomized experiments (see, e.g., Foster, Nan, Shen, Kaciroti, & Taylor, 2016), our use of permutation here is solely to generate null distributions for testing *cp* tuning parameter values.

The proposed method for heterogeneous subgroup identification (summarized in five steps):

1. *Estimate propensity scores.* A vector of estimated propensity scores, $\hat{e}(\mathbf{X})$, is created. A logit model is often used, but nonparametric methods, such as random forests or boosted CART, align well with our approach because they automatically handle interactions.

2. *Estimate individual treatment effects.* Units are grouped into strata via optimal full propensity score matching and ITEs are estimated, as described above.

3. *Type I error rate permutation study.* The cost-complexity tuning parameter value for the regression tree is selected based on permutation-based Type I error rate studies. Optionally, causal variable selection may be applied to reduce the dimension of the covariate space.

4. *Identify subgroups.* A single regression tree using the cost-complexity tuning value identified in Step 3 is fit to the estimated ITEs with covariates as predictors. Any subgroup splits identified by the regression tree are interpreted as evidence for effect heterogeneity on the variable or variables involved in the splits.

5. *Estimate subgroup-specific effects.* Perform many-to-many optimal matching again, subject to the additional constraint that matching takes place within heterogeneous subgroups identified in Step 4, and estimate overall average and subgroup-specific treatment effects.

We designed two Monte Carlo simulation studies to examine the efficacy of the proposed method. The studies are intended to explore (a) the ability of the permutation approach to control Type I error rate as intended, (b) the subsequent power to detect heterogeneous subgroups when they exist, and (c) the accuracy of estimation of overall and subgroup-specific treatment effects.

Simulation Study 1

In the first simulation study, we generated data with independent covariates and effect heterogeneity based on only one covariate. Data generation for Simulation Study 1 is described next.

1. Covariates $x_1, x_2, \ldots, x_{10}$ were independently generated from the following distributions, which were selected to reflect the natural variety of predictor types encountered in practice

(i.e., continuous, dichotomous and count). Generated samples were of size $n = 1000$.

$$
\begin{aligned}
x_1 &\sim N(0,1); & x_6 &\sim Bernoulli(0.3) \\
x_2 &\sim Poisson(2); & x_7 &\sim N(-1,3) \\
x_3 &\sim Bernoulli(0.5); & x_8 &\sim N(2,2) \\
x_4 &\sim N(0,2); & x_9 &\sim N(1,0.8) \\
x_5 &\sim Bernoulli(0.6); & x_{10} &\sim N(2,1)
\end{aligned}
$$

2. Let $X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}, x_{i9}, x_{i10})'$ be the vector of covariate values for unit i. The propensity score was modeled as follows,

$$e_i(X) = \frac{\exp(\alpha + \beta X_i)}{1 + \exp(\alpha + \beta X_i)}, \tag{1}$$

where the generating coefficients are $\alpha = -1$ and $\beta = (0.6, 0.1, -0.3, -0.4, 0.2, -0.3, -0.2, 0.2, 0.5, 0.3)$. These coefficients range in magnitude from small to moderate, are both positive and negative, and yield sufficient propensity score overlap.

3. Values of the treatment assignment vector, T, were determined by comparing the propensity score $e_i(X)$ to a random uniform variable on [0,1], U_i, such that $T_i = 1$ if $U_i \leq e_i(X)$ and $T_i = 0$ otherwise.

4. Potential outcomes were generated as follows,

$$Y_i^1 = \rho X_i + \gamma + \lambda I(x_{i1} > 0.5) + \epsilon_{i1} \tag{2}$$

$$Y_i^0 = \rho X_i + \epsilon_{i0}, \tag{3}$$

where $\rho = (0.2, 0.1, 0.2, -0.1, -0.2, 0.2, -0.2, 0.1, 0.2, 0.1)$, $\epsilon_1, \epsilon_0 \sim N(0,1)$ and γ, the treatment main effect, was set equal to 1.0. The indicator was defined such that $I(x_{i1} > 0.5) = 1$ if $x_{i1} > 0.5$ and $I(x_{i1}) = 0$ if $x_{i1} \leq 0.5$. That is, two heterogeneous subgroups are defined by covariate x_1 with the cutoff value at 0.5. The observed outcome Y_i

was determined by the potential outcomes as follows: $Y_i = Y_i^1 T_i + Y_i^0(1 - T_i)$.

5. Data $= \{(Y_i, X_i, T_i), i = 1, \ldots, n\}$.

6. Replicate the above steps for $R = 1000$ replications.

The coefficient of the treatment by x_1 interaction, denoted by λ in Equation 2, controls the strength of the effect heterogeneity. To study the sensitivity of the method to correctly detect heterogeneous subgroups, λ was set to 1, 2, or 3. For those values, semi-partial correlations between the interaction term and the the outcome variable were, respectively, .12, .21, and .27, which correspond loosely to Cohen's (1988) guidelines of .1 and .3 for small and moderate effect sizes for Pearson correlation.

The true value of the overall average treatment effect may be determined as follows:

$$\begin{aligned}
\text{ATE} &= E[Y_i^1 - Y_i^0] \\
&= E[\rho X_i + \gamma + \lambda I(x_{i1}) + \epsilon_{i1}] - E[\rho X_i + \epsilon_{i0}] \\
&= \gamma + \lambda E[I(x_{i1})] \\
&= \gamma + \lambda(1 - \Phi(0.5)) \\
&= 1 + \lambda(0.31),
\end{aligned}$$

where the second-to-last equality follows from the fact that $x_1 \sim N(0, 1)$. Thus, ATE = 1.31, 1.62, and 1.93, for $\lambda = 1$, 2, and 3, respectively. Bias, simulation standard deviation, and mean squared error were used to measure the adequacy of the ATE estimators over the R simulation replications as follows.

$$\begin{aligned}
\text{Bias} &= \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\text{ATE}}_r - \text{ATE}(\lambda)\right) \\
\text{SD} &= \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\text{ATE}}_r - \bar{\hat{\text{ATE}}}_r\right)^2} \\
\text{MSE} &= \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\text{ATE}}_r - \text{ATE}(\lambda)\right)^2,
\end{aligned}$$

where $\text{ATE}(\lambda) = 1 + \lambda(0.31)$.

The values of subgroup-specific CATEs also depend on the λ values. Recall that the treatment main effect, γ, was set equal to 1 and the data-generating cutoff value for x_1 was set to be 0.5, so the two subgroup-specific CATEs are calculated as follows:

$$\text{CATE}(\lambda)_{\{x_1 \leq 0.5\}} = E[Y^1 - Y^0 | x_1 \leq 0.5] = 1 \tag{4}$$

$$\text{CATE}(\lambda)_{\{x_1 > 0.5\}} = E[Y^1 - Y^0 | x_1 > 0.5] = 1 + \lambda. \tag{5}$$

Thus, for $\lambda = 1$, 2, and 3, $\text{CATE}_{\{x_1 > 0.5\}} = 2$, 3 and 4, respectively. For the CATE estimators, bias and mean squared error are defined as follows.

$$\text{Bias} = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\text{CATE}}_r - \text{CATE}(\lambda) \right)$$

$$\text{MSE} = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\text{CATE}}_r - \text{CATE}(\lambda) \right)^2,$$

where $\text{CATE}(\lambda)$ is determined as described in Equations (4) and (5).

Results of Simulation Study 1

Cost-complexity tuning parameter values .01, .02 and .03 yielded Type I error rates of .82, .18 and .03, respectively. After iterative search, we settled on a cost-complexity value of .028, which was associated with Type I error rate of approximately .05. Thus, we set the cost-complexity parameter at .028 for subsequent analyses. Results are presented in Table 1, which displays the proportion of replications for which the method identified subgroups. When the treatment by covariate interaction was nil (i.e., $\lambda = 0$), the Type I error rate was .039, close to the nominal level of .05, as expected. When the treatment by covariate interaction was small (i.e., $\lambda = 1$), the correct covariate, x_1, was correctly identified as the only predictor of subgroup heterogeneity in 621 out of 1000 replications. For moderate and large magnitudes (i.e., $\lambda = 2$ and 3), x_1 was correctly identified as the only predictor of subgroup heterogeneity in all replications.

For $\lambda = 1$, the average cut score on x_1 for the 621 replications for which the correct

covariate was identified was 0.52, with a moderately large standard deviation of 0.18; recall the data-generating cut score was 0.50. For the moderate heterogeneity condition with $\lambda = 2$, the average cut score identified was 0.51, with a standard deviation of 0.5. For the large heterogeneity condition with $\lambda = 3$, the average cut score was also 0.51, with a standard deviation of 0.03.

Table 1
Subgroup Identification Performance across 1000 Replications in Simulation Study 1

Case	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
x_1 only	0	621	1000	1000
Split but not on x_1	39	12	0	0
No subgroups identified	961	367	0	0

Note. λ is the heterogeneity coefficient; see Equation 2.

Table 2 summarizes the accuracy of estimators of ATEs and subgroup-specific CATEs through bias, simulation standard deviation, and mean squared error. For each estimator, there is a row of the table for the "Naive" version and the "Step (5)" version. The "Naive" versions of estimators are based directly on the regression tree output from Step (4), whereas the "Step (5)" estimators are based on stratified matching within subgroups, as specified in Step (5). It is also important to note that when the magnitude of the treatment by x_1 interaction was small (i.e., $\lambda = 1$), our approach correctly identified heterogeneity about 62% of the time. Thus, estimates for $\lambda = 1$ in Table 2 are based on 621 replications.

The overall ATE was estimated without bias across all three values of λ and for both the naive and Step (5) estimators. As for the conditional effects, the naive estimators carried relatively large and biases which increased as the strength of the treatment effect modification increased. In contrast, the Step (5) estimators, which matched propensity scores within subgroups, were close to unbiased.

Simulation Study 2

Data generation for Study 2 was similar to Study 1 with two exceptions: (a) Study 2 included three correlated covariates and (b) heterogeneous subgroups were defined using two covariates in Study 2. In particular, for Study 2, x_1, x_7 and x_9 were generated from a multivariate

Table 2
Accuracy and Precision of Subgroup Treatment Effect Estimates across 1000 Replications in Simulation Study 1

	$\lambda = 1$			$\lambda = 2$			$\lambda = 3$		
Estimator	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
Cut-off value on x_1	0.02	0.18	0.03	0.01	0.05	0.00	0.01	0.03	0.00
ATE (Naive)	0.01	0.11	0.01	0.01	0.12	0.01	0.01	0.13	0.02
ATE (Step (5))	0.01	0.10	0.01	0.00	0.11	0.01	0.00	0.12	0.01
$\text{CATE}_{\{x_1 \leq 0.5\}}$ (Naive)	0.10	0.11	0.02	0.25	0.13	0.08	0.36	0.14	0.15
$\text{CATE}_{\{x_1 \leq 0.5\}}$ (Step (5))	-0.02	0.17	0.02	0.01	0.17	0.02	0.01	0.17	0.02
$\text{CATE}_{\{x_1 > 0.5\}}$ (Naive)	-0.16	0.13	0.05	-0.51	0.14	0.29	-0.77	0.13	0.62
$\text{CATE}_{\{x_1 > 0.5\}}$ (Step (5))	0.10	0.21	0.05	0.03	0.21	0.04	0.02	0.21	0.04

λ is the heterogeneity coefficient described in Equation 2; note that estimates are based on 621 replications for λ = 1 and 1000 replications for λ = 2 and 3, respectively; "Naive" estimates are obtained from the regression tree output after Step (4); "Step (5)" estimates are obtained after stratified matching within subgroups.

normal distribution with mean and variance-covariance matrix as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.3 & 0.3 \\ 0.3 & 1.0 & 0.3 \\ 0.3 & 0.3 & 1.0 \end{bmatrix}.$$

The other seven covariates $x_2, x_3, x_4, x_5, x_6, x_8$ and x_{10} were independently generated as in Study 1. Propensity score generating coefficients for Study 2 were set to $\alpha = -1$ and $\beta = (0.6, 0.2, -0.3, -0.4, 0.2, -0.3, -0.2, 0.2, -0.5, 0.3)$ (see Equation 1), and potential outcomes were generated as follows:

$$Y_i^1 = \rho X_i + \gamma + \lambda I(x_{i1} > 0.5) I(x_{i3} = 1) + \epsilon_{i1} \tag{6}$$

$$Y_i^0 = \rho X_i + \epsilon_{i0}, \tag{7}$$

where $\rho = (0.2, 0.1, 0.2, -0.1, -0.2, 0.2, -0.2, 0.1, 0.2, 0.1)$, $\epsilon_{i1}, \epsilon_{i0} \sim N(0, 1)$ and γ, the treatment main effect, was set equal to 1. Three heterogeneous subgroups are defined by the indicator functions. The first subgroup corresponds with $x_{i1} \leq 0.5$ and $x_{i3} = 0$; the second subgroup consists of cases for which either $x_{i1} \leq 0.5$ or $x_{i3} = 0$, but not both, and the third subgroup is when $x_{i1} > 0.5$ and $x_{i3} = 1$. The observed outcome Y_i was determined by the potential outcomes as follows: $Y_i = Y_i^1 T_i + Y_i^0 (1 - T_i)$.

The coefficient of the treatment by x_1 by x_3 interaction, denoted by λ in Equation 6, governs the strength of the heterogeneity across subgroups. As in Study 1, λ was set to 1, 2, or 3. For those values, semi-partial correlations between the interaction term and the the outcome variable were, respectively, .12, .22, and .30.

The true value of the overall average treatment effect may be determined as follows:

$$\begin{aligned}
\text{ATE} &= E[Y_i^1 - Y_i^0] \\
&= E[\rho X_i + \gamma + \lambda I(x_{i1} > 0.5) I(x_{i3} = 1) + \epsilon_{i1}] - E[\rho X_i + \epsilon_{i0}] \\
&= \gamma + \lambda E[I(x_{i1} > 0.5) I(x_{i3} = 1)] \\
&= \gamma + \lambda E[x_{i1} > 0.5] E[x_{i3} = 1] \\
&= \gamma + \lambda (1 - \Phi(0.5))(0.5) \\
&= 1 + \lambda (0.31)(0.5) \\
&= 1 + \lambda (0.155),
\end{aligned}$$

Thus, for $\lambda = 1$, 2, and 3, the ATE is 1.15, 1.31, and 1.46, respectively. Bias, simulation standard deviation, and mean squared error are used to measure the adequacy of the ATE estimators over the R simulation replications, $r \in \{1, \ldots, R\}$, as follows.

$$\begin{aligned}
\text{Bias} &= \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\text{ATE}}_r - \text{ATE}(\lambda) \right) \\
\text{SD} &= \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{\text{ATE}}_r - \bar{\hat{\text{ATE}}}_r \right)^2} \\
\text{MSE} &= \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\text{ATE}}_r - \text{ATE}(\lambda) \right)^2,
\end{aligned}$$

where $\text{ATE}(\lambda) = 1 + \lambda(0.155)$.

Because the heterogeneity is based on both x_1 and x_3, there will be three true subgroups, no matter whether the first split is on x_1 or x_3. We denote them as left, middle and right groups, respectively. When the first split is on x_1 and the second split is on x_3, the three

subgroup-specific CATEs may be calculated as follows:

$$\text{CATE}(\lambda)_{\{x_1 \leq 0.5;\text{left}\}} = E[Y^1 - Y^0 | x_1 \leq 0.5] = 1 \tag{8}$$

$$\text{CATE}(\lambda)_{\{x_1 > 0.5 \ \& \ x_3 = 0;\text{middle}\}} = E[Y^1 - Y^0 | x_1 > 0.5 \ \& \ x_3 = 0] = 1 \tag{9}$$

$$\text{CATE}(\lambda)_{\{x_1 > 0.5 \ \& \ x_3 = 1;\text{right}\}} = E[Y^1 - Y^0 | x_1 > 0.5 \ \& \ x_3 = 1] = 1 + \lambda. \tag{10}$$

When the first split is on x_3 and the second split is on x_1, the three subgroup-specific CATEs may be calculated as follows:

$$\text{CATE}(\lambda)_{\{x_3 = 0;\text{left}\}} = E[Y^1 - Y^0 | x_3 = 0] = 1 \tag{11}$$

$$\text{CATE}(\lambda)_{\{x_3 = 1 \ \& \ x_1 \leq 0.5;\text{middle}\}} = E[Y^1 - Y^0 | x_3 = 1 \ \& \ x_1 \leq 0.5] = 1 \tag{12}$$

$$\text{CATE}(\lambda)_{\{x_3 = 1 \ \& \ x_1 > 0.5;\text{right}\}} = E[Y^1 - Y^0 | x_3 = 1 \ \& \ x_1 > 0.5] = 1 + \lambda. \tag{13}$$

Thus, for $\lambda = 1$, 2, and 3, $\text{CATE}_{\{x_1 > 0.5 \ \& \ x_3 = 1\}} = 2$, 3 and 4, respectively. For the CATE estimators, bias and mean squared error are defined as follows.

$$\text{Bias} = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\text{CATE}}_r - \text{CATE}(\lambda) \right)$$

$$\text{MSE} = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\text{CATE}}_r - \text{CATE}(\lambda) \right)^2$$

where $\text{CATE}(\lambda)$ is determined as described in Equations (8) to (13).

Results of Simulation Study 2

We examined the relationship between cost-complexity tuning parameter values and Type I error rate through permutation with $\lambda = 0$. Cost-complexity values of .01, .02 and .03 yielded Type I error rates of .70, .09 and .02, respectively. After iterative search we settled on a cost-complexity value of .023, which corresponded with a Type I error rate of about .04.

Results on subgroup identification performance are presented in Table 3, which displays the proportion of replications for which the method identified subgroups at $\lambda = 1, 2, 3$. When the treatment by covariate interaction was nil (i.e., $\lambda = 0$), the Type I error rate was .051, close to the desired nominal level of about .04. When the treatment by covariate interaction was relatively

small (i.e., $\lambda = 1$), x_1 and x_3, were correctly identified as the only two predictors of subgroup heterogeneity in only 10 out of 1000 replications. For moderate magnitudes of λ (i.e., $\lambda = 2$ and 3), the number of correct decisions improved to 658 and 965 out of 1000 replications, respectively. For $\lambda = 1$, the average cut score on x_1 for the 10 replications for which the correct covariate was identified was 0.65, with a standard deviation of 0.35. For $\lambda = 2$, the average cut score of x_1 for the 658 replications was 0.55, with a standard deviation of 0.17. For the large heterogeneity condition with $\lambda = 3$, the average cut score of x_1 was also 0.55, with a standard deviation of 0.13.

Table 3
Subgroup Identification Performance across 1000 Replications in Simulation Study 2

Case	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
x_1 and x_3	0	10	658	965
1^{st} split on x_1 but 2^{nd} split not on x_3	0	77	129	27
1^{st} split on x_3 but 2^{nd} split not on x_1	3	2	4	2
Split but not on x_1 or x_3	48	37	32	4
No subgroups identified	949	874	177	2

λ is the heterogeneity coefficient described in Equation 6.

Table 4 summarizes the accuracy of estimators of ATEs and subgroup-specific CATEs through bias, simulation standard deviation, and mean squared error. As before, the naive versions of estimators are based directly on the regression tree output from Step (4); whereas, Step (5) estimators are based on stratified matching within subgroups. It is also important to note that estimates for $\lambda = 1$ in Tables 3 and 4 are based on only 10 replications, so they are much less stable than for the other two values of λ, which are based on 658 and 965 replications, respectively.

The results related to accuracy of estimation for Study 2 largely mirror those of Study 1. As in Study 1, the overall ATE was estimated without bias for both the Naive and Step (5) estimators. Furthermore, as in Study 1, the CATEs were estimated with large biases via the Naive approach that were largely corrected with the Step (5) modification. The results of Study 2 diverge from those in Study 1 with respect to the sensitivity to detect heterogeneous subgroups. In particular, although effect sizes for effect heterogeneity were similar across the two studies at the three levels of λ, the addition of another predictor governing the subgroups led to a substantial loss in sensitivity for the small effect size condition in Study 2.

Table 4
Accuracy and Precision of Subgroup Treatment Effect Estimates across 1000 Replications in Simulation Study 2

	$\lambda = 1$			$\lambda = 2$			$\lambda = 3$		
Estimator	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
Cut-off value on x_1	0.15	0.35	0.13	0.05	0.17	0.03	0.05	0.13	0.02
ATE (Naive)	-0.01	0.08	0.01	0.00	0.10	0.01	0.00	0.11	0.01
ATE (Step (5))	-0.03	0.10	0.01	-0.00	0.10	0.01	-0.00	0.10	0.01
$\text{CATE}_{\{left\}}$ (Naive)	-0.04	0.07	0.01	0.12	0.11	0.03	0.20	0.12	0.05
$\text{CATE}_{\{left\}}$ (Step (5))	-0.11	0.09	0.02	-0.02	0.12	0.01	0.00	0.13	0.02
$\text{CATE}_{\{middle\}}$ (Naive)	0.07	0.12	0.02	0.18	0.15	0.05	0.23	0.16	0.08
$\text{CATE}_{\{middle\}}$ (Step (5))	0.01	0.25	0.06	0.07	0.25	0.07	0.05	0.26	0.07
$\text{CATE}_{\{right\}}$ (Naive)	-0.00	0.06	0.00	-0.62	0.21	0.43	-1.08	0.23	1.21
$\text{CATE}_{\{right\}}$ (Step (5))	0.35	0.39	0.26	0.05	0.28	0.08	-0.00	0.28	0.08

λ is the heterogeneity coefficient described in Equation (7); note that estimates are based on only 10 replications for λ = 1, 658 replications for λ = 2, and 965 replications for λ = 3; "Naive" estimates are obtained from the regression tree output after Step (4); "Step (5)" estimates are obtained after stratified matching within subgroups.

ECLS-K Case Study

The Early Childhood Longitudinal Study-Kindergarten cohort (ECLS-K; NCES, 2001), is a national longitudinal study focused on child development and early school experiences. The treatment effects of interest for this case study are related to the impacts of receiving special education services versus not receiving special education services on mathematics scores measured during fifth grade (year 2004). The data are motivated by Morgan, Frisco, Farkas, and Hibel (2010) and were described in Keller and Tipton (2016). Morgan et al. (2010) identified 34 covariates, measured at either kindergarten or first grade, based on theoretical considerations or prior empirical research results, distributed across six domains: demographic (three variables), academic (nine variables), school composition (five variables), family context (six variables), health (four variables), and parent rating of child (seven variables). The covariates are described in greater detail, including ECLS-K code names, in Keller and Tipton (2016). After eliminating cases with missing data, the final analytic sample included 7,362 children, among whom 429 had received special education services and 6,933 had not.

Using the five-step procedure, we first estimated propensity scores using a logit model with interactions[1]. For the second step, we used optimal full matching on the estimated propensity

[1]The usual recommended steps for the specification of the propensity score, including iterative respecification of the propensity score model to achieve optimal balance, and an examination of overlap on the logit of the propensity score, are important here, though we do not describe them in detail because our focus is on the detection of heterogeneous

score logit with a caliper of 0.2 pooled SDs to obtain 371 matched strata; matching was carried out with package **optmatch** (Hansen & Klopfer, 2006). The number of children within each matched stratum ranged from 2 to 501 with a median of 6 and mean of 19. Eighty percent of matched strata had fewer than 20 children and ninety percent of matched strata had fewer than 39 children in each stratum. For each child, unobserved potential outcomes were imputed by taking the mean outcome score for all children of opposite exposure status in the same stratum. Finally, individual treatment effects were estimated as differences, either $Y_i^1 - \hat{Y}_i^0$ or $\hat{Y}_i^1 - Y_i^0$, as appropriate.

Next, for step three, we ran a permutation-based Type I error rate study to select the value of the cost-complexity tuning parameter to use with regression trees. The key for permutation here is to randomly shuffle yoked outcome/treatment pairs, leaving covariate values fixed. This scheme ensures that the raw treatment effect is retained for each permutation, while any relationships between covariates and treatment effects are destroyed. By inducing a lack of systematic association between covariate values and treatment effects, we ensure that the permuted data align with the permutation null hypothesis of no effect. Then, permutation testing may be run with varying values of the regression tree cost-complexity tuning parameter to assess Type I error rates. Testing based on permutation is particularly useful because, under ignorability and a no interference assumption, it permits valid testing without reliance on parametric assumptions (Rubin, 1980; Keller, 2012). Five hundred permuted data sets were created for each cost-complexity value. Type I error rate was calculated as a proportion over the 500 randomizations for which our method resulted in regression trees with more than one node (i.e., at least one split), thereby spuriously indicating covariate-dependent treatment effect heterogeneity. We found the cost-complexity parameter value of 0.004 was associated with a Type I error rate of 0.01 for the mathematics outcome variable.

In the fourth step, regression trees were fit using cost-complexity parameter values of 0.004, with the vector of individual treatment effect estimates used as the outcome variable, to identify subgroups with heterogeneous treatment effects. Before fitting, we pre-processed the data with a random forest-based variable selection algorithm (Keller & Zhang, 2018, March) to reduce the dimension of the space over which we searched for heterogeneity; eight variables were retained.

subgroups. See, e.g., Keller & Tipton, 2016 for a summary of recommended steps in propensity score analysis.

After fitting, any splits identified by the regression tree were interpreted as evidence of heterogeneous subgroups. Finally, for step five, we reran optimal full propensity score matching with the added constraint that matches must be made within subgroups to estimate the overall ATE and the subgroup-specific CATEs.

Results of the Case Study

Setting the cost-complexity parameter at .004 led to a six-subgroup solution wherein one subgroup had only 248 units. To increase the likelihood of landing on a stable solution, we explored two approaches. First, we imposed a constraint on the regression tree so that splits on nodes with fewer than one thousand units would not be attempted. Second, we increased the value of the cost-complexity parameter (higher values are more conservative) incrementally until the tree solution became simpler; we settled on a value of .007. Both approaches led to the same four-group solution shown in Figure 1.

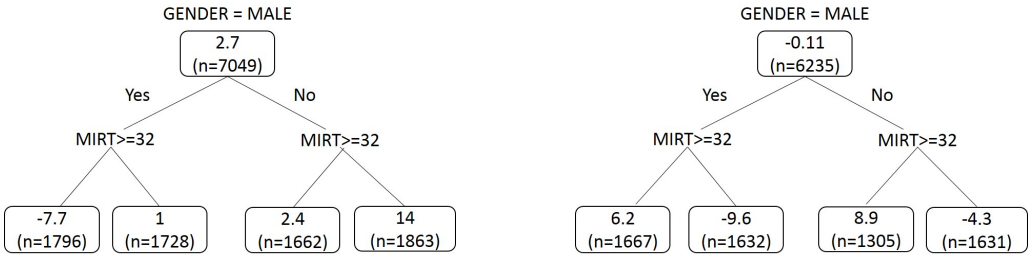

Figure 1. Heterogeneous subgroups for the 5th grade mathematics outcome variable with CATEs estimated by Naive (left panel) and Step (5) (right panel) methods; GENDER = student gender (male = 1, female = 0); MIRT = kindergarten mathematics score

To interpret the plot for the 5th grade mathematics outcome variable, shown in Figure 1 (left panel), note that the full sample of 7049 students was split into two subgroups, depending on whether the student's gender was male or female. The 3524 male students were split into two groups based on MIRT, thereby modeling an interaction between gender and kindergarten mathematics score; the 3525 female students were split into two groups based on MIRT as well. The right panel of Figure 1 shows the CATE estimates after exact matching within subgroups (i.e., Step (5)). Sample sizes are smaller after exact matching because we used a caliper of 0.2 on the pooled SD of the logit of the propensity score. See Table 5 for naive and Step (5) estimates of

the overall ATE and subgroup-specific CATEs.

Estimator	Naive		Step (5)	
	n	Est	n	Est
Overall ATE	7049	2.68	6235	-0.11
$\text{CATE}_{\{\text{MALE; MIRT}\geq 32\}}$	1796	-7.70	1667	6.24
$\text{CATE}_{\{\text{MALE; MIRT}<32\}}$	1728	1.02	1632	-9.59
$\text{CATE}_{\{\text{FEMALE; MIRT}\geq 32\}}$	1662	2.42	1305	8.91
$\text{CATE}_{\{\text{FEMALE; MIRT}<32\}}$	1863	14.46	1631	-4.32

Table 5

Overall and subgroup-specific estimates of the impact of exposure to special education services on students' mathematics achievement in fifth grade; Total sample size $n = 7362$ *is reduced to* $n = 7049$ *due to a caliper of 0.2 pooled SDs of the logit propensity score imposed on matches; Total sample size is further reduced for Step (5) due to the constraint that matches be made within subgroups*

By using a conservative cost-complexity parameter of 0.007, we can be confident that subgroup splits are likely not spurious. We found evidence for four heterogeneous subgroups based on the interaction of student gender and kindergarten mathematics score. Simulation study results underscore the fact that naive CATE estimates are untrustworthy. Step (5) results, which were much less biased in our simulations, are shown alongside the naive results in Table 5. Interpreting the Step (5) estimates, we found evidence for a main effect for kindergarten mathematics achievement such that exposure to special education is more likely to be helpful for children who scored above the median. Furthermore, there was a main effect for gender, wherein the effect of exposure to special education was more beneficial for girls than boys. Finally, gender and MIRT interact such that the impact of kindergarten mathematics score on the efficacy of special education was more dramatic for boys.

Conclusion

In this paper we described an approach that identifies subgroups of units in an observational study for which a treatment is differentially effective based on the values of one or more covariates. Our approach is accessible because it uses the mechanics of propensity score matching and simple regression trees, both of which are familiar to applied researchers in education and behavior science, and easy to implement with existing software packages. We found through Monte Carlo simulation that the method correctly flagged covariates that interact with treatment

with high probability when their interaction with treatment was moderately strong or stronger. Furthermore, we found that by selecting the value of the cost-complexity tuning parameter for the regression tree fit through permutation testing, it was possible to control the proportion of false positives to a prespecified nominal level. As is the case with any simulation study, however, it is not appropriate to make strong generalizations to data scenarios that are very different.

In a case study using data from ECLS-K, we identified student gender and kindergarten mathematics score as the most important moderators of the effect of special education services on 5th grade mathematics achievement. In particular, we estimated the average effect of exposure to special education services to be negative (i.e., harmful) for those students with mathematics pretest scores lower than the median score of 32, and positive (i.e., helpful) for those with mathematics pretest scores at or above 32. With the inclusion of gender as well, results suggest that the moderating effect of kindergarten mathematics score is, itself, moderated by a student's gender such that exposure to special education was more helpful (and less harmful) for female students than for their male counterparts.

The results of the case study demonstrate how heterogeneous subgroup identification can provide meaningful and important results that are of immediate interest to stakeholders such as teachers, parents, and administrators. But how reliable are the results? First, we must underscore that the results will only be valid insofar as the assumptions required for identification and estimation of average causal effects through conditioning strategies, as laid out in Rosenbaum & Rubin, 1983, are upheld. In our case study, for example, we might be concerned that ignorability is not satisfied. Although a sensitivity analysis could be carried out to probe the robustness of estimates to unobserved confounding (Rosenbaum, 2002), we did not do so herein. Furthermore, the definition of the "treatment" is opaque. Nuances such as total duration, quality of services, staff training, etc., have not been considered and yet certainly varied from student to student and site to site. We might also be concerned that there may have been interference between students assigned to special education in the same schools; aside from accounting for several school-level covariates, our analyses do not handle school-level nesting.

Concerns about the assumptions required for identification notwithstanding, if we assume they are met, are the results on the identification of heterogeneous subgroups likely to be trustworthy using this method? Our simulation results are informative here because they suggest

that by using cost-complexity tuning parameter values selected through permutation testing, the Type I error rate for subgroup identification may be held to a nominal level. Thus, we may be confident that any covariates identified by the procedure as moderators are real, subject to the usual caveats regarding the possibility of Type I errors. With respect to estimation, although our simulation results suggest that subgroup-specific estimates based on Step (5) are likely to be accurate, a limitation of our method is that no standard errors or confidence intervals are provided for the CATEs. Future research aimed at proposing and evaluating resampling approaches such as the jackknife, bootstrap, and boosting to attain error bounds on estimated CATEs with this method would be useful.

References

Assman, S. F., Pocock, S. J., Enos, L. E., & Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, *355*, 1064–1069.

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

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks/Cole.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*.

Foster, J. C., Nan, B., Shen, L., Kaciroti, N., & Taylor, J. M. G. (2016). Permutation testing for treatment–covariate interactions and subgroup identification. *Statistics in Biosciences*, *8*, 77–98.

Foster, J. C., Taylor, J. M. G., Kaciroti, N., & Nan, B. (2015). Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. *Biostatistics*, *16*, 368–382.

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*, 2867–2880.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627.

Harding, J. F., Morris, P. A., & Hill, J. (2017). Understanding associations between low-income mothers' participation in education and parenting. *Journal for Research on Educational Effectiveness*, *10*, 704–731.

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

Hsu, J. Y., Small, D. S., & Rosenbaum, P. R. (2013). Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, *108*, 135–148.

Hsu, J. Y., Zubizarreta, J. R., Small, D. S., & Rosenbaum, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, *102*, 767–782.

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

Imai, K., & Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, *19*, 1-19.

Kaplan, D., & Chen, J. (2012). A two-step Bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, *77*, 581–609.

Keller, B. (2012). Detecting treatment effects with small samples: The power of some tests under the randomization model. *Psychometrika*, *77*, 324-338.

Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics*, *41*, 326-348.

Keller, B., & Zhang, T. (2018, March). *A New Method for Variable Selection with Random Forests in a Conditional Independene Framework*. Paper presented at the 2018 Society for Research on Educational Effectiveness Spring Conference, Washington, D. C.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*, 236-254.

NCES. (2001). *Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use data files user's manual* (Tech. Rep.). Technical Report No. NCES 2001-029. U.S. Department of Education.

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). NY: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized

studies. *Journal of Educational Psychology*, *66*, 688-701.

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

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. Little (Ed.), *Oxford handbook of quantitative methods, volume 1: Foundations.* Oxford, UK: Oxford University Press.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 213–236.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21.

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

Therneau, T., Atkinson, B., & Ripley, B. (2015). *rpart: Recursive partitioning and regression trees.* Retrieved from `http://CRAN.R-project.org/package=rpart` (R package version 4.1-9)

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*, 1228–1242.

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, *42*, 314-347.