Causal Inference Using Educational Observational Data: Statistical Bias Reduction Methods and Multilevel Data Extensions

Jose M. Hernandez

A Dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Robert Abbott, Chair
Elizabeth Sanders
Joe Lott

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

**Causal Inference Using Educational Observational Data:
Statistical Bias Reduction Methods and Multilevel Data Extensions**

Jose M. Hernandez

Chair of the Supervisory Committee:

Professor Robert Abbott, Ph.D.
College of Education

This study utilizes a data driven simulation design, which deviates from the traditional model-based approaches most commonly adopted in quasi-experimental Monte Carlo (MC) simulation studies, to answer two main questions. First, this study explores the finite sample properties of the most utilized quasi-experimental methods that control for observable selection bias in the field of education and compares them to traditional regression methods. Second, this study lends an insight into the effects of ignoring the multilevel structure of data commonly found in the field when using quasi-experimental methods. Specifically, treatment effects were estimated using (1) Ordinary Least Squares (OLS) multiple linear regression (treatment effects, adjusted for mean differences on confounders), (2) Propensity Score Matching (PSM) using nearest neighbor 1:n with replacement, (3) Propensity Score Matching using Inverse Probability Weighting (IPW) of the propensity score, and (4) Propensity Score Matching using Sub-classification (Subclassification). There were five main factors that were varied to simulate the data, all of which were fully crossed, as follows: Four sample sizes (600, 1000, 2000, and 5000)**;** three association levels among simulated variables (low, moderate, high); two treatment exposure levels (25% and 50%)**;** four treatment effect sizes using Cohen's *d* (none, low, moderate, and high)**;** and five levels of ICCs (0, .10, .20, .30, and .40). These 480 conditions were each analyzed with four methods of analysis, for a total of 1920 conditions. Additionally, using data from the Educational Longitudinal Study of 2002 (ELS:2002), an applied study demonstration of the different estimation methods in question was performed and compared to the simulation results. Findings indicate that under certain conditions all methods compared perform the same and have similar estimates of treatment effects. Additionally, when the clustering of the data is ignored bias is introduced for smaller sample size conditions.

Table of Contents

LIST OF FIGURES iii

LIST OF TABLES iv

Chapter I: Overview of quasi-experimental methods in education research: Setting the foundation and identifying future directions 1

Randomized Experiments 2
Quasi-Experiments 3
Assumption of Common Support/Overlap 6
Organizing Framework 7
Prevalence of Quasi-Experiments in Education Research 9
Use of Propensity Score Matching Methods in Policy Research 10
Research on PSM Methods 11

Chapter II: Quasi-experimental methodology: Bias reduction methods for observable characteristics 14

Regression and Estimation of Treatment Effects 14
Matching Methods Review 18
Propensity Score Methods (PSM) 23
Weighting Using the Inverse Propensity Score 29
Sub-Classification using the Propensity Score 29
Optimal Subclassification/Full Matching 30
Issues and gaps 31
Research Questions 34

Chapter III: Monte Carlo Simulation Methods 36

Monte Carlo Simulation Design 36
Data Generation (Simulation) 41
Analysis Approach and Evaluation Criteria 42
Secondary Analyses of Simulation Results 44

Chapter IV: Monte Carlo Simulation Results 45

Descriptive Statistics 45

Null Effects of the Single Level Model Results 46

All Condition Results for the Single Level Data Model 47

Varying ICC Values 52

Chapter V: Applied Analysis Demonstration with Multilevel Modeling 54

Methods 55

Results 57

Summary 57

Chapter VI: Discussion 58

Extensions 61

References 62

Appendix A: Additional Simulation Results 68

Appendix B: R Sample Simulation Code 71

LIST OF FIGURES

Figure 1. Experimental Design Organizing Framework 8
Figure 2. Unbiased Estimation of Effects 16
Figure 3. The Relationship of P and e when X is not Modeled but Exists 17
Figure 4. Causal Graph of The Regression Treatment Estimator 18
Figure 5. Bias Reduction Methods for Observed Characteristics 34
Figure 6. Subclassification Methods Relative to the other PSM and OLS Methods Shows the Most Bias 49
Figure 7. Correlation Level "r" Main Effects Across Three Levels (0.10, 0.30 , 0.50) 49
Figure 8. Estimation Method Bias as a Function of Method and Higher Correlations 50
Figure 9. 95% CI Coverage by Correlation Magnitude 52
Figure 10. Bias Increase as a Function of Sample size n and ICC 53

LIST OF TABLES

Table 1. Quasi-experiments in Educational Research, 2012-2014 9
Table 2. Matching Example Using Years of Math in HS as a Covariate 20
Table 3. Students Matched on Exact Values of X 20
Table 4. Exact Matching Using Two Covariates 21
Table 5. ANOVA Results for Single Level Data: Bias Outcome 48
Table 7. Empirical Coverage Rates by Method 51
Table 8. NELS 2002 Variables for use in the Applied Study 55
Table 9. Correlation of Variables form NELS 2002 Applied Study 55
Table 10. Monte Carlo Simulation Results Relevant to Applied Study 56
Table 11. Results of Different Treatment Effect Estimation Methods 57

Acknowledgements

This work is not possible without the tremendous support that I've received from my loved ones, I am forever indebted to all of you. Thanks to my adviser, Robert Abbott, he has supported me, challenged me, and trained me to be an independent researcher. Thank you for giving me the liberty to pursue and develop my own research agenda. I aspire to emulate your leadership and statistical knowledge in the future. I can confidently say that there was not one meeting that we ever had where I didn't leave without a new drop of knowledge. Many thanks go to my amazing committee – Elizabeth Sanders, Joe Lott, Thomas Halverson and Adrian Dobra, for their tremendous support and guidance. I am greatly appreciative of Elizabeth Sanders, the dedication and mentorship that she provides to students is unparalleled. Thanks to Mike Knapp for always challenging me to stay balanced in my academic pursuits. Your ability to provide guidance and mentorship was invaluable. Lastly, I'd like to thank my CREST and CS&SS families, my development as a scholar greatly benefited from being an active member of those groups.

Dedication

This dissertation is dedicated to my mother and father: Maria Del Rosario and Manuel Hernandez-Estrada. I hope that I can live long enough to one-day repay the tremendous sacrifices that you've made in order to secure the success of your children.

Overview of quasi-experimental methods in education research: Setting the foundation and identifying future directions.

The field of education is often concerned with answering causal questions about the effects that new or reformed policies and programs, or "treatments", on student academic and behavioral outcomes. One issue, however, is that many educational studies are not randomized experiments. Research designs that do not employ randomization are called "observational," "correlational," or "quasi-experimental." Given that these types of non-randomized studies may appear often in educational research (prevalence discussed in the forthcoming sections), there has been a push for the development and adoption of rigorous quasi-experimental methodology for estimating causal effects (Reynolds & DesJardins, 2009b). The development of these methods, with an emphasis on the unique challenges that arise in the field of education (data, samples, etc.) is crucial in the future development of quantitative research methodology in education. Randomized experiments are still thought of as superior methods over quasi-experiments for use on evaluations but there has been a growing demand for the development of "strong" quasi-experimental methodology within the field of education, albeit with little consensus (Hill, 2008; Ryan & Cousins, 2009; Steiner, Cook, Shadish, & Clark, 2010).

Researchers who are interested in making causal claims using observation data are left with the task of finding ways to control statistically for the different types of confounding that might be present. Some educational research benefits from the abundance of large scale data to test educational intervention effects on student outcomes (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). Yet, despite their positive qualities, large educational databases do not result from random assignment condition of subjects to specific treatments or program interventions.

Randomized Experiments

The advantage of using randomized experiments to test causal hypothesis rely on the following assumptions:

I. In randomized experiments, the assignment to the causal condition (treatment or control group) is not systematic, but by chance. That is, treatment assignment is made by an "objectively defined random mechanism" and not by a subjective decision making process by the people conducting the study or the participants of the study (Rubin, 1978).

II. Relationships between covariates, outcomes, and interventions are known. Another advantage of a random mechanism for the placement of individuals in one of the two causal conditions (i.e. flipping a fair coin where all individuals have a 50% probability of being in one of the two conditions) is that other factors (both observed and unobserved) that could potentially have an effect on the outcome (i.e. confounders, or biases) are balanced between the two causal conditions (Gilbert, 1974; Weinstein, 1974).

III. Potential alternative explanations for the difference other than it being associated with the intervention can be dismissed. The randomization creates two groups that are similar on average (probabilistically) to each other, making comparisons between the two yield differences that are due to controlled factors.

Formally, the assignment condition for any individual is said to be independent of their potential outcome under either condition (Y^1,Y^0) (Morgan & Winship, 2007; Rubin, 1974) as:

$$(Y_i^0,Y_i^1) \perp\!\!\!\perp T \qquad (1)$$

This is also known as un-confoundedness (Rosenbaum & Rubin, 1983a) and ignorable treatment assignment (Morgan & Winship, 2007). The probability of any outcome Y for any individual is the same regardless of the causal condition to which they are assigned. Randomization and

consequently this independence assumption guarantee that both treatment conditions are well balanced in terms of potential biases that could arise from chance. As a result, causal inference under randomization is a direct comparison of a difference between individuals in the treatment condition versus the individuals in the control condition.

For example, consider a study in which we wish to explore the effects of participation in the Math Engineering Science Achievement (MESA) program. MESA programs are implemented in high schools for historically disadvantaged youth, and are designed to help build a pathway to college and careers in science, technology, engineering, and mathematics (STEM). It is anticipated that students who participate in MESA may experience increased interest in STEM-related content in high school as well STEM-related college majors and/or post-high school careers. If students are randomly assigned to either participate in this program or not, then a direct comparisons between the mean of those who received treatment and those who did not provide strong causal evidence about the impact of such program. Again, their assignment to participate or not is completely independent of the intended program intervention.

On the other hand, what if participation in the MESA program was voluntary? Or what if it was deemed unethical to withhold treatment? Other reasons that might restrict randomization could be logistical in nature or involve funding limitations. For example, sometimes a new program needs to be piloted first in order to secure future funding, and is administered at some specified time of the day to some specific subset of the population. When randomization is not feasible due to logistic, financial, or ethical reasons, we turn instead to quasi-experimental designs to provide evidence for treatment impacts.

Quasi-Experiments

Quasi-experiments, popularized by Campbell and Stanley (1963), can also be used to test

causal hypotheses using non-randomized observational data. There are however, some important caveats to consider when drawing inferences from quasi-experiments. For example, a central problem lies in the situation in which a number of subjects are exposed to an intervention/treatment (as in a randomized experiment) but there is no well-defined (methodologically or systematic) control group with which to make the appropriate comparison (Dehejia & Wahba, 2002). Therefore, any comparison between a treatment group and a non-equivalent control group is assumed to be biased due to three sources: (1) self-selection into the treatment, (2) researcher-selection bias into the treatment, and (3) the treatment was not intended to be randomly assigned, as is the case with interventions that are driven by a new implemented blanket policy across a whole population.

Individuals often self-select into the treatment condition or participation is non-randomly predetermined by a provider. As a result, there is no direct comparison group that is "identical" or "equivalent" to the treatment group (equivalence refers to equality of groups prior to treatment participation). For example, a MESA program implemented at the high school level might show most if not all of these sample selection bias issues. One way this can happen is when a program official controls participation, or some participants voluntarily participate. Another way self-selection can occur is when the sample of participants might over-represent a certain population who may be pre-disposed to believe the treatment is worthwhile for their particular circumstance (i.e., students from high poverty backgrounds). Additionally, researchers may purposefully, or unwittingly, over-recruit certain portions of the population to participate in the treatment that are not representative of the population as a whole. This results in the possibility that there exist systematic differences between treatment and control groups, *other than the treatment*, making the direct causal inference of group differences inappropriate due to the confounding.

Quasi-experiments are always associated with some type of confounding condition. A confound is defined as a variable that explains a relationship between a potential causal variable on an outcome, such as math ability or family socioeconomic status, that are both related to participation on a math intervention program and math achievement outcomes (Shadish, Cook, & Campbell, 2001). Researchers must account for these confounds in one of the two following ways:

- Control confounding conditions by a priori research design.
- Controlling for them statistically post-hoc.

Quasi-experimental based statistical methods have become a popular means for drawing causal inference based on probabilistically estimated treatment effects. These methods involve the pairing of individuals that have been exposed to a well-defined treatment condition to individuals that were not exposed to a treatment condition using a set of observable characteristics (also known as "observables," "covariates," "predictors," "matching variables," and "explanatory variables") in the data (Cochran & Rubin, 1973; R. Dehejia & Wahba, 2002). The logic is as follows: if individuals who were exposed to a treatment condition can be sufficiently represented by a set of observables (covariates in the data), an unbiased estimate of the treatment effect is possible even without the randomization mechanism (Angrist & Pischke, 2009; Morgan & Winship, 2007; Rubin, 1974).

Observational studies, in lieu of the conditions/assumptions that allow for direct comparison between treatment and control individuals in true experiments, use observational data (covariate/s) to define a set of comparable control individuals. The conditional independence assumption is analogous to the randomization assumption (1) with an added condition. In the case that we are using non-randomized data from an observational study,

conditional dependence can be satisfied if we can control for observed confounders, defined by X_i in Equation 2 below (Angrist & Pischke, 2009; Morgan & Winship, 2007; Rosenbaum & Rubin, 1983a, 1984).

$$(Y_i^0, Y_i^1) \perp\!\!\!\perp T \mid X_i \quad (2)$$

Equation 2, which is called the "ignorable treatment assumption" given above is read: the potential outcome for an individual in either the treatment (Y_i^1) or the control (Y_i^0), is independent of the treatment condition if we can control for a set of observable covariates. Using the specifications of the previous examples, X_i represents any covariate for the ith student, such as socioeconomic status (SES), and T represents the treatment condition, such as participation in the MESA program. Both Y_i^0 and Y_i^1 correspond to the outcome associated with the ith student's participation (Y_i^0) or not (Y_i^1) in the MESA program. In this example, if the values of the SES for each student in the treatment are assumed to be comparable to values of the covariate(s) of the individuals in the non-treatment (control) group, and further, if we observe the same set of covariates (such as SES) for both groups, then the conditional probability of any outcome $Y(Y_i^0, Y_i^1)$ for any individual will be the same (equivalent), regardless of the condition they were assigned to (Rosenbaum & Rubin, 1983a; Rubin, 1974).

Assumption of Common Support/Overlap

An intrinsic assumption of Equation 2 in terms of proper estimation of causal effects using quasi-experiments is called "common support," which is the sufficient overlap between the treatment group and (potential) control group created from the covariate(s) used in the analysis to ensure group equivalence. This is actually the implied meaning of the assumption that we must "observe the same set of covariates" in order to infer causal treatment effects (Equation 2). Using the MESA study example, all students would need to have similar distributions of SES in both

treatment and control groups. This ensures that there are sufficient "matches" between the treatment and control group individuals to be able to make a valid group comparison (Rosenbaum & Rubin, 1983a). Formally, common support takes on the following form:

$$0 < P(T=1 \mid X) < 1. \quad (3)$$

In Equation 3, the probability of receiving the treatment must lie between zero and one, if the probability of receiving the treatment is exactly zero or exactly one (perfectly discriminate between selection) there will be no comparable subjects at those values (Hirano, Imbens, & Ridder, 2003).

If the ignorable treatment assumption (2) and the common support/overlap assumption (3) hold, then the inference about a treatment-control group difference becomes analogous to that of the completely randomized case, and we can express the expectation of the potential outcomes for any individual following (Rubin, 1974, 1977):

$$E[Y_i^{T(0,1)} | T_i=0 | Xi\ ] = E[Y_i^{T(0,1)} | T_i=1 | Xi\ ] \quad (4)$$

The expectation given in Equation 4 of the potential outcome for any individual given that the individual is in the control group is equal to the expectation of the potential outcome for any individual given that that individual is in the treatment group, given that we have conditioned on the observed covariates. The central idea behind (4) is that, by conditioning on the covariate(s), in terms of the expectation of the potential outcomes in both treatment and control groups will be balanced. This allows us to continue an analysis in an observational study as if it was a true randomized study (J. S. Sekhon, 2009).

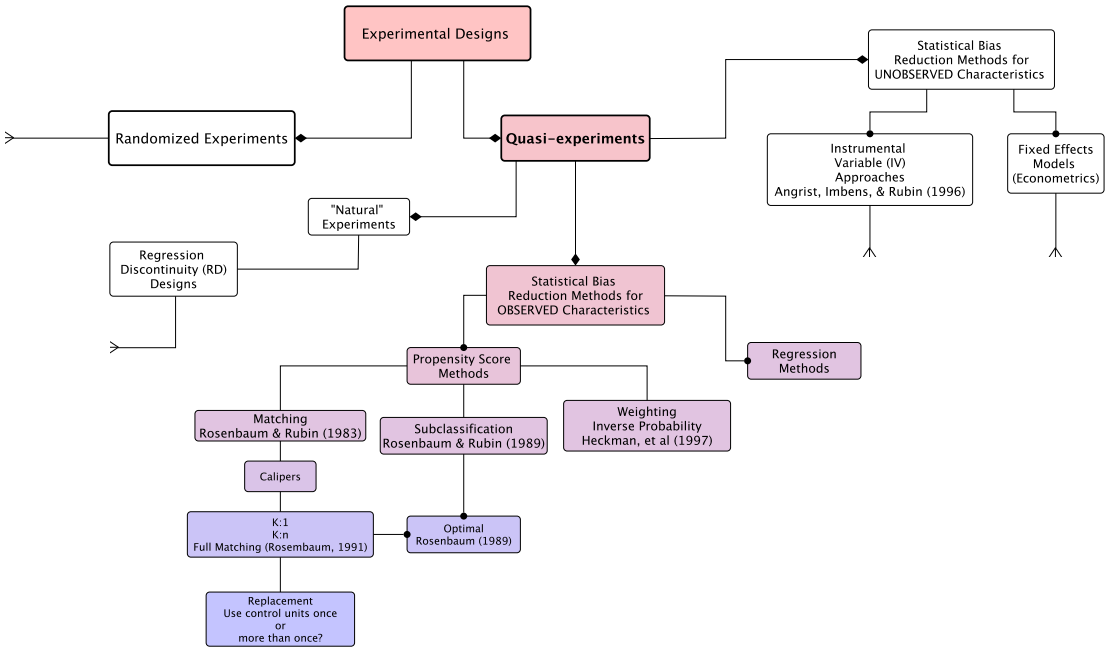
Organizing Framework

For organizational purposes, the present study makes three distinctions among types of quasi-experiments:

- Natural experiments,
- Statistical bias reduction methods for observed (measured) characteristics, and
- Statistical bias reduction methods for unobserved characteristics.

Figure 1 displays the pathways of these three types of quasi-experiments.

Figure 1. Experimental Design Organizing Framework



Broadly, natural experiments are studies in which treatment(s) is/are a naturally occurring phenomenon to which individuals cannot be randomly assigned. A common example of this in education would be children who are lower academic achievers compared with children who are higher performing. The second type of quasi-experiment, bias reduction when observed characteristics are known, refers to studies in which treatments are assigned in a non-randomized fashion, but for which covariates were measured and can be used to statistically control for confounding. The third type, which involves bias reduction methods for unobserved characteristics, includes studies in which covariates were not measured and therefore cannot be used to statistically control for confounding.

The main focus of the present work unpacks the statistical bias reduction methods for "observed" characteristics in Figure 1 (middle pathways). These methods are intended to be utilized when a researcher has access to measurable characteristics that can be used to describe a population of interest. The theoretical foundations of these observable bias reduction methods come from the seminal work of Cochran and Rubin (1973) and Rubin (1973,a,b;1977, 1976, 1979). These methods will be unpacked in Chapter 2.

Prevalence of Quasi-Experiments in Education Research

To assess the prevalence of quasi-experiments in current educational research, a systematic review consisting of four of the most prevalent education research journals was conducted for the three previous years (2012, 2013, and 2014). These journals included *American Research Education Journal* (AERJ), *Educational Evaluation and Policy Analysis* (EEPA), *Research in Higher Education* (RHE), and *Journal of Research on Educational Effectiveness* (JREE). The results of the review, shown in Table 1, revealed that 55 (19%) of the original research studies published employed quasi-experiments of some form, and that 20 (36%) of the quasi-experiments used propensity score matching (PSM) methods for data analysis.

Table 1. Quasi-experiments in Educational Research, 2012-2014

Journal	Totals	Quasi Experiment	PSM	Instrumental Variable	Regression Discontinuity	Fixed Effects	Other
AERJ	81	14	4	0	0	0	10
EEPA	76	25	9	2	11	2	1
RHE	91	7	6	1	0	0	0
JREE	46	9	1	2	6	0	0
Grand Total	294	55	20	5	17	2	11

In fact, the use of matching-based techniques like PSM methods is relatively new to educational research yet it is largely under studied (Hill, 2008; Steiner et al., 2010). Reynolds & DesJardins (2009a) provided an extensive review of matching methods and their application to

education research; however they only focused on PSM methods. Hong and Raudenbush (2006) used propensity score stratification in a multilevel model framework to assess the effects of a kindergarten retention policy on student success in mathematics in the later schooling years. Herzog (2008) explored the effects of financial aid on retention of college students using a stratified propensity score model to match aided versus unaided college freshman. Reliance on the propensity score in education is of no surprise, given the large data sets available for analysis, use of PSM does enable researchers to use a large number of covariates to balance treatment and control groups.

Most recently, researchers are utilizing large scale data (primarily available through NCES) to look at various effects using propensity score methods, including, effects of high school course taking on secondary and post-secondary educational attainment (Long et al., 2012); the influence of dual enrollment programs on academic achievement (An, 2013); the earning benefits of majoring in science, technology, engineering, and mathematics (STEM) degrees on high achieving minority students (Melguizo and Wolniak, 2011); and the impact of undergraduate research programs (Eagan et al., 2013).

Use of Propensity Score Matching Methods in Policy Research

The application of matching based methods within the context of policy research is dominated by studies examining the effects of state or federal programs on students' performance outcomes (Hahs-Vaughn & Onwuegbuzie, 2006; G. Hong & Raudenbush, 2005; G. Hong & Yu, 2008). For example in 2006, Mathematica Policy Research, Inc. used a propensity score analysis to conduct a national evaluation of Talent Search programs in several states (US Department of Education, 2006). In 2009, the Department of Education, Office of Planning, Evaluation and Policy Development, conducted a national study on the Student Support Services

(SSS) program, using (among other advanced methodological tools) matching students using the propensity score (Pell Institute, 2009). More recently, propensity score methods have been applied to studies looking at different curricular programs implemented at the state level (Bhatt and Koedel, 2012); the efficacy of private sector support providers for failing schools under No Child Left Behind (Heinrich and Nisar, 2013); Assessment of the impacts of city ran principal training programs on student achievement (Corcoran et al., 2012).

Research on PSM Methods

The previous approaches to estimating treatment effects using observational data has been widely adopted in education research and educational policy evaluation (Reynolds & DesJardins, 2009b). As educational researchers and policy evaluators, we are often interested in why certain programs work well in some context and not in others. This is often the case when we look at intervention programs that are implemented at a number of different sites with varying site characteristics (i.e. high schools, college campuses). From a policy perspective, the decision to fund certain programs based on their causal impact rests on (1) the proper evaluation of these programs, (2) a distinction between aggregate impacts versus local context impacts (Howard, Raudenbush, & Weiss, 2014). If there exists variation in the effects of specific intervention between different sites, understanding when programs work, for whom they work, and how and why they work when they do, rests on properly estimating the treatment effects.

Recent work on program impact variation and multi-site estimation of program effects using a multi-level modeling framework is strictly focused on multi-site Randomized Control Trials (RTC's). For example, Zhou et al., (2013) propose a Bayesian hierarchical model to analyze multi-site experimental fMRI data. This approach models the hierarchical structure of the data (subjects nested within each experimental site) and allows for the modeling of the

variation between each site. As a result, the estimation of both the within site variation and the between site variation is more precise. In the case where there exists high variation between sites, their method provides a more precise estimation of treatment effects across all sites compared to an approach that ignores the hierarchical structure of the data (Weiss, Bloom, & Brock, 2013; Zhou et al., 2013). Similarly, Dehejia (2003) illustrated the benefits in using a hierarchical modeling of the data in assessing the impact of multiple sites that were exposed to a treatment condition (RCT's). Additionally, they find that modeling site specific characteristics, using Multi-level Modeling (MLM) allows for the potential to more accurately predict treatment effects on sites that were not given a treatment but that are similar in terms of site specific characteristics (R. H. Dehejia, 2003).

In the realm of quasi experiments, and MLM, recent work by Thoemmes & West (2011), Arpino & Mealli (2011), G. Hong & Raudenbush, (2005; 2006), and Kim & Seltzer (2007) provide a spectrum of possible modeling decisions in the context of both multilevel modeling and various matching algorithms for the purpose of making causal inferences. Hong and Raudenbush (2005; 2006) and Kim and Seltzer (2007) explored a situation in which there exists clustering in terms of characteristics that affect the treatment assignment condition at the level-2. These authors make the case for using a random intercept and random slope multilevel modeling specification to account for the level-1 and level-2 treatment condition interaction. Much like Rosenbaum (1986), who wanted to account for the between school variation in both treatment condition assignment and treatment effect, Hong and Raudenbush (2005; 2006) and Kim and Seltzer (2007) do this by specifying a stratified propensity score multilevel model to perform the matches. Their approach, allows for direct comparison between different cluster-level factors (low-retention schools vs high retention schools) and their treatment effects.

This recent work on MLM and matching based methods has allowed applied researchers to begin the foundational work on application of these methods in applied educational journals. For example, Vaughan, Lalonde, & Jenkins-Guarnieri (2014), investigate the effect of participation in first year seminar courses on student academic achievement (i.e. first semester GPA). Methodologically, they utilize a hierarchical propensity score matching technique to control for the self-selection of students into the courses and to additionally model the potential variation due to instruction and potential dependence between subjects.

Similarly, Guanglei Hong, Corter, Hong, & Pelletier (2012) and Kelcey(2011), utilize a multilevel propensity score model that was initially developed by Guanglei Hong & Raudenbush (2006), to assess the differential effects of ability grouping in kindergarten students on their academic achievement and the effects of teachers reading knowledge on student reading achievement respectively. Both these studies are concerned with modeling potential variation of effects that are due to school-level characteristics, and both specify complex multilevel models (i.e. random-intercept and random-slope) that require specific methodological considerations.

Quasi-experimental methodology: Bias reduction methods for observable characteristics

Currently, numerous approaches exist for comparing observations in treatment and control groups, some very distinct, but many have various overlapping components. This chapter unpacks the use of methods used to reduce the bias due to observable characteristics in the data between treatment and control groups. This chapter will also provide a comprehensive look at the entire spectrum of these techniques, and will provide guidelines for appropriate applications.

Regression and Estimation of Treatment Effects

Regression is perhaps the most utilized form of data analysis across the social sciences and education. Regression methods have benefited from the rigorous advancement and assortment of data modeling options developed by statisticians in the previous decades. Regression within the context of causal inference and treatment effect estimation however, is either seldom discussed within textbooks or is often faced with a level of contention (Morgan & Winship, 2007). The apprehension is not completely without warrant; regression estimates are biased when there is an omission of critical explanatory variables in the regression formulation. As a result, most statistical courses/text books focus on simple regression estimation procedures as a data reduction tool for descriptive purposes. Regression estimators however, as will be shown in the following section lend themselves to: (1) regression adjustment as a procedure of treatment effect estimation, and (2) regression lends itself to the formal notation of causal inference in the form of the potential outcomes framework (Cochran & Rubin, 1973; Morgan & Winship, 2007; Splawa-Neyman, 1990). Although there are several forms of treatment effects within the regression context, this chapter will focus primarily on Ordinary Least Squares (OLS) regression as an estimator of treatment effects.

Regression and Bias. Consider the case where you have a measured outcome Y, and an explanatory measured variable P. To add substantive meaning to this, consider the MESA example where the measured outcome Y corresponds to a students math outcome and P corresponds to a participation indicator, where students are coded depending on whether or not they participated in the MESA program. A regression in the form of:

$$Y = \mu_0 + \delta P + e \,, \tag{5}$$

where, μ_0 is the intercept, δ is the beta coefficient corresponding to the program participation, and e is a random variable that represents all other causes of y that have not been explicitly accounted for. If participation of students in MESA measured by P is a random process then equation 5 is an unbiased estimator of average treatment effects (Morgan & Winship, 2007; Rubin, 1974).

Figure 2. Unbiased Estimation of Effects

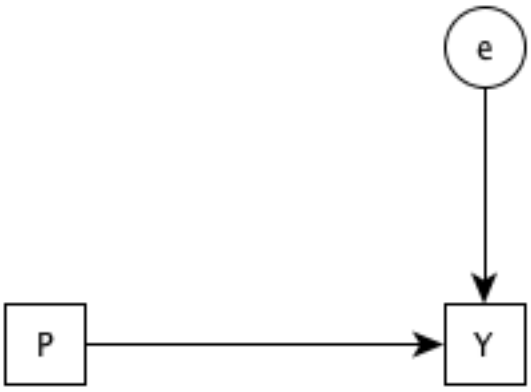
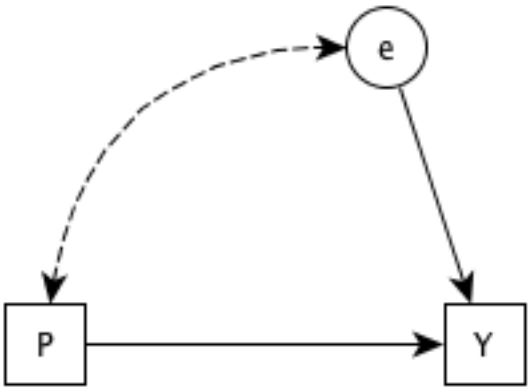


Figure 2 represents the fully randomized condition which can also be expressed as in equation 5, where treatment assignment P has no relationships with other observed or unobserved explanatory variables.

If, however, there exists a variable X that is also associated with either Y or P, for example X can be a variable indicating math self confidence, and it is omitted from equation 5, and that same equation is used to estimate the effect (without X) then the estimation of the main effect δ will be biased. To understand why, referring to the error term 'e' can help illustrate this. The introduction of variable X underlines the idea that P is associated (correlated) with other explanatory variables, and when this relationship is not explicitly modeled, as shown in figure 3, the error term 'e' directly accounts for this relationship.

Figure 3. The Relationship of P and e when X is not Modeled but Exists



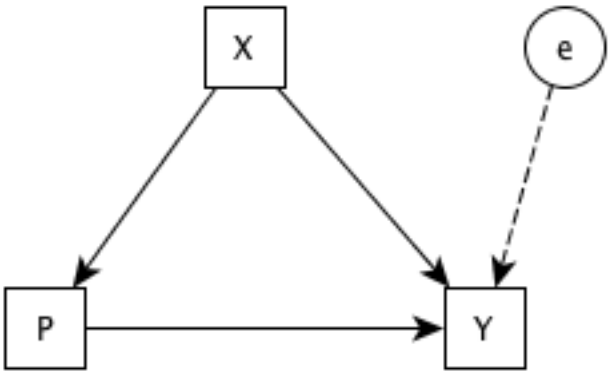
The problem however is that e is not a measured variable and as a result introduces bias in the form of this relationship between the error term (referred to as omitted variable bias) into the estimation of Y.

Regression: Towards causal inference. What if a researcher knew that there was other causes of Y and had indeed a measure for these? Would including them solve the causality problem? Formally represented in equation 6, notice the additional measured variable X.

$$Y = \mu_0 + \delta P + \beta X + e, \tag{6}$$

where β represent the regression coefficient of X in this model. The answer is not completely definite, as can be expected when dealing with statistical concepts, but in some cases including X in the estimation of Y will yield consistently accurate estimates of the treatment effect. If you introduce many X's associated with P and Y, the relationships between P and e in figure 3 will approximate zero.

Figure 4. Causal Graph of The Regression Treatment Estimator



As Figure 4 illustrates, the known and measured relationship of X will properly represent the association between these variables. The caveat is that there can exist an unmeasured variable that is not included with even a large number of X's. If the following three conditions can be met (Morgan & Winship, 2007):

1. P is uncorrelated with e.
2. The effect of P does not vary as a function of X, and
3. X is parameterized properly (linear, categorical, etc.)

Then the regression estimator with measured variables X can estimate causal effects, efficiently and accurately.

Matching Methods Review

Matching methods in practice consist of (1) defining a distance measure to be used to assess the "closeness" of treatment and control subjects, and (2) selecting a matching algorithm to perform matches between treatment and control subjects dependent on the measure of "closeness" (Rosenbaum & Rubin, 1983a, 1984; Rubin, 1974; Stuart & Green, 2008; Stuart, 2010). The two most common distance metrics used in sociology, statistics and economics for multivariate matching are the Mahalanobis distance (Cochran & Rubin, 1973; Rubin, 1980) and

the propensity score (Rosenbaum & Rubin, 1983b).

Single Variable Case. Matching methods have been discussed formally since the late 1930's by (Chapin, 1938, 1947; Fisher, 1926, 1935; Greenwood, 1945). In the early 1970's, work by Cochran & Rubin (1973) and Rubin (1973 a, b) provided a theoretical foundation for matching methods.

Early work by (Cochran, 1968) focused on matching using two methods for adjusting treatment and control groups: (1) exact sub-classification and (2) exact regression (covariance) adjustment. Using sub-classification stratifies both treatment and control groups on the pretreatment confounding variable X_i (Cochran & Rubin, 1973). The treatment effects are then estimated with-in each stratum (sub-class) and pooled to get a global estimate of the treatment effects. For example, consider the example where there is a math tutoring intervention designed to improve student math test scores. Additionally consider a confounding variable of math confidence that is measured by a five point likert scale. Subclassification would require the researcher to estimate in each subclass of math confidence (1,2,3,4,5) the difference between the proportion of students (in both treatment and control groups) of their math test score outcome, specific to each subclass. These sub-class specific effects would be pooled to form a global estimate of the effect of participating in the tutoring program.

In regression (Covariance) adjustment on a single confounding variable x_i a linear model is fit to the regression of y on x_i, it is then used to create an adjusted estimate of treatment effects (Cochran & Rubin, 1973):

$$Effect = \bar{y}_1 - \bar{y}_2 - \hat{\beta}(\bar{x}_1 - \bar{x}_2) \quad (7)$$

Where, $\bar{y}_1, \bar{x}_1$ are the y and x means in the treatment group and $\bar{y}_2, \bar{x}_2$, are the y and x means in the control group; beta is an estimate of the slope, on x.

Both these methods used a form of exact matching in which subjects were matched based on exact values on x. In the univariate case, this is straight forward and in many ways is the most ideal (Imai, King, & Stuart, 2008). Once again consider the Math tutoring program and its effect on math performance outcomes (such as test scores or course grades), represented in table 2. The matching variable of interest is the students' number of math courses completed in high school.

Table 2. Matching Example Using Years of Math in HS as a Covariate

Individual (i)	Treatment Assignment (T)	Math in HS (x)
$Student_1$	1	1
$Student_2$	0	3
$Student_3$	0	2
$Student_4$	1	4
$Student_5$	0	2
$Student_6$	1	3
$Student_7$	0	3
$Student_8$	1	2
$Student_9$	1	3
$Student_{10}$	0	1

T=0 is control group and T=1 treatment group

If we were to match participants from the treatment group to subjects in the control group that have the "exact" values on X. Table 3 represents the resulting matching that would occur.

Table 3. Students Matched on Exact Values of X

Individual (i)	Treatment Assignment (T)	Math in HS (x)	Control Match
$Student_1$	1	1	$Student_{10}$
$Student_4$	1	4	
$Student_6$	1	3	$Student_7$
$Student_8$	1	2	$Student_3$
$Student_9$	1	3	$Student_2$
$Student_2$	0	3	
$Student_3$	0	2	
$Student_5$	0	2	
$Student_7$	0	3	
$Student_{10}$	0	1	

note: T = 0 is control group and T = 1 treatment group

Student comparisons will only occur between matched student subjects on the variable "Math in

HS". However, notice on Table 3 that subject 4 has no match in the control group; in the final analysis this student would be discarded. Exact matching in this case is perhaps the most straightforward and non-parametric approach (Sekhon, 2009). Exact matching is commonly represented by this distance measure (Stuart, 2010):

$$D_{ij} = \begin{cases} 0, & if\ X_i = X_j \\ \infty, & if\ X_i \neq X_j, \end{cases} \quad (8)$$

When researchers are matching on a single variable, the term "similar" or "exact" to that variable is clearly defined. However as the number of covariates increases, the idea of similar or exact is not clearly defined and exact matching becomes inefficient (Cochran & Rubin, 1973; P. Rosenbaum & Rubin, 1983a; D. Rubin, 1976; Stuart, 2010).

Multivariate Extensions. Consider the case with more than one covariate. As multiple covariates were considered, finding "exact" matches for individuals becomes a bit more challenging. For example, refer to Table 4, can you find exact matches for the student participants using two variables?

Table 4. Exact Matching Using Two Covariates

Individual (i)	Treatment Assignment (T)	Math in HS (X_1)	Parent education (X_2)
$Student_2$	0	3	9
$Student_3$	0	2	0
$Student_5$	0	0	2
$Student_7$	0	3	3
$Student_{10}$	0	1	5
$Student_1$	1	1	6
$Student_4$	1	4	4
$Student_6$	1	3	1
$Student_8$	1	2	7
$Student_9$	1	3	6

This dimensionality problem, makes it very difficult and sometimes impossible to find exact matches with similar values on all covariates for matched individuals (Zhao, 2004). Exact matching is often impossible to implement if exact matches do not exist in the population/sub-

groups in the data (Sekhon, 2009). A study conducted by Chapin (1947) found that only 23 pairs of matched participants were possible using six covariates with initial pools of 671 treated and 523 control participants. That's a significant and problematic loss of valuable data.

The theoretical foundation for multivariate settings was first unpacked by Rubin (1976) and Cochran & Rubin (1973). These methods would aim to summarize a any set of $p < 1$ variables into a single measure that would adequately contain their relationships to both the outcome and the treatment indicator (Rubin & Thomas, 1996; Rubin & Thomas, 2000). These methods reduce bias in all covariate directions (i.e., makes the covariate means closer) by the same amount, ensuring that if close matches are obtained in some direction, then the matching is also reducing bias in all other directions.

Mahalanobis Distance. The Mahalanobis distance addresses the complexity of dimensionality directly by scaling multiple covariates into a distance metric (Rubin 1980). The Mahalanobis distance is a scalar quantity that measures the multivariate distance between two individuals corresponding to the treatment and control group using a set of covariates. Using standard statistical notation the Mahalanobis Distance (MD) between two individuals corresponding to a treatment and control conditional on a vector of X covariates is (Rubin, 1980):

$$MD(X_{iT},X_{iC}) = \{(X_{iT} - X_{iC})^{T} S^{-1} (X_{iT} - X_{iC}) \}^{1/2}, \quad (9)$$

where X_{iT} and X_{iC} are values of X for subjects from G_1(treatment) and G_2(control) and S is the sample covariance matrix of X. A matched sample using MD will have N pairs of the difference between G_1-G_2 subjects that are matched conditional on the set of covariates X.

Genetic Matching as a distance. Genetic Matching is an iterative algorithm that searches amongst a range of distance metrics to find an optimizing post-matching covariate

balance (Diamond & Sekhon, 2012; Sekhon, 2008; Sekhon, 2007). Formally the generalized Mahalanobis distance takes on the following form (Diamond & Sekhon, 2012; Sekhon, 2009):

$$\textit{Generalized}\ \mathrm{MD}(X_{iT}, X_{iC}, W) = \{(X_{iT} - X_{iC})^{T} (S^{-1/2})^{T}\ WS^{-1/2}(X_{iT} - X_{iC})\}^{1/2} \qquad (10)$$

Genetic matching (GMD) generalizes a form of the Mahalanobis distance, and adds the additional weight parameter W that corresponds to each variables "relative" importance in achieving optimal balance (Diamond & Sekhon, 2012, p.6). The parameter W corresponds to a k x k positive definite weight matrix that weights each variable to maximize balance. Genetic Matching combines both propensity score matching and Mahalanobis distance to optimize the balance of the comparison groups (Diamond & Sekhon, 2012; Sekhon, 2009; Sekhon, 2008; Sekhon, 2007). For further technical details see Diamond and Sekhon (2012). For example, the decision can be made to match on the propensity score in combination with the set of covariates. The X in equation (10) is replaced by Z, which is a matrix that contains the Propensity Score in combination with the covariates X_i. The algorithm will then give a corresponding weight to the Propensity Score and to the covariates. In theory, the algorithm could converge to giving a 1 to the propensity score and a zero weight to the covariates, and matching using genetic matching will simply be equivalent to matching on just the propensity score. The alternative scenario (giving a zero weight to the PS) is the equivalent of matching on just the Mahalanobis distance. In the general sense however, the algorithm will most likely neither minimize only on using the MD or only the PS, and will iteratively converge on the best balanced solution (Diamond & Sekhon, 2012).

Propensity Score Methods (PSM)

Propensity score methods (PSM) involve using the probability of assignment to treatment conditional on a set of covariates as a way to compose a comparable group to compare to the

treatment group (Rosenbaum & Rubin, 1983a). Matching using the propensity score refers to using a selected number of covariates to balance between those students who received the treatment condition and a control group using an estimated propensity score. The vector X_i can contain a large number of variables with various higher order dimensions; the propensity score reduces all of these dimensions into one scalar quantity(Guo & Fraser, 2010). Each respondent on the treatment group is then matched with the closest non-treatment group subject based on their predictive probability of being in the treatment (Pearl, 2009; Rosenbaum & Rubin, 1984, 1985; Sekhon, 2009). The propensity score is formally described as (Rosenbaum &Rubin, 1983):

$$e(x_i) = \Pr(T = 1 \mid X_i = x_i) \tag{11}$$

where propensity score $e(x_i)$ is the predictive probability that a given student will receive the treatment (T=1), conditional on a set of covariates. Rosenbaum and Rubin (1983, 1985) suggest using the logit of the predicted probability as the propensity score:

$$\text{Exp}(x_i) = \text{logit}\,[e(x_i)/1 - e(x_i)]. \tag{12}$$

The propensity score has been shown to provide the best balancing score, especially in scenarios where you have a large number of covariates (X. Gu & Rosenbaum, 1993). The predictive probability difference between the predictive probabilities of interest is as follows:

$$P(X_{iT}, X_{iC}) = \left| e_{iT} - e_{iC} \right| \tag{13}$$

Matching on the estimated linear propensity score has been shown to reduce the most bias due to linear trends in the selected covariates (Rosenbaum & Rubin, 1983a; Rubin & Thomas, 1996). Matching on the predicted probabilities as discussed in the previous section might give you misleading results as the predicted probabilities are bounded between zero and one and it is possible that compression exists along the tails (Rosenbaum & Rubin, 1985). As such, matching

using the linear predictor of the PS $\mu` = X_i\hat{\beta}$ will not compress the PS near its tails (J. S. Sekhon, 2009). The linear predictor is also more often than not closer to being normally distributed . The difference probability of interest is formally (Stuart, 2010):

$$LP(X_{iT},X_{iC}) = |\text{logit}(e_{iT}) - \text{logit}(e_{iC})|. \quad (14)$$

Algorithms. Once a matching method has been selected, the next step is to find an efficient way to match control units to the treatment units from the sample. The algorithms available for the most part have two important functions: (1) find the best possible individual matches (weight) from the populations obtained by the matching method, and (2) provide a basis for selecting the number of individuals that will be used for the matching from a possible set of control groups.

Nearest Neighbor K:1 Matching. The most common matching procedure and inexpensive to implement is "nearest available" pair matching, known as k:1 nearest neighbor matching or "greedy" matching in the literature across disciplines (Rubin, 1973a, 1973b) In practice, treated units are ordered and the first subject in that group, is paired with the nearest subject in the control group. The second subject in the treatment group is paired with remaining subjects in the control group. If k is greater than 1 (more than one control group for unit every treatment unit) after the algorithm completes one full iteration, the first treatment unit will be assigned a control unit from the remaining subjects in the control group (Rubin, 1973a; Gu & Rosenbaum, 1993). In general, nearest neighbor matching does not minimize the individual distance between treated and control subjects, however it does minimize an "overall" distance between subjects (Gu & Rosenbaum, 1993b).

Additionally, if multiple control subjects are "closest" to one treatment subject, the matching of individuals becomes a bit dubious, and further considerations must be made. In

K:1 nearest neighbor matching, the easiest form to implement is selecting a 1:1 ratio of control units to treatment units and finding and assigning matches that minimize their distance. The general from of nearest neighbor ratio matching is K:1, and if the data yield a large number of potential control units (control units > treatments units) the researcher can increase the number to a fixed number of "k" controls or a variable number of "K" controls (Ming & Rosenbaum, 2001; Rubin & Thomas, 2000; Stuart, 2010). Rubin and Thomas (1996,2000) suggest using no more than 5 control units per treatment subject, as each additional match beyond the initial match will increase the distance between probabilities (2nd match will be closer to the 1st than the 3rd, etc.) and thus will result in an increase of bias (Stuart, 2010). Furthermore the researcher has to assess whether there exists some treated subjects that have many close matches while other might have very few, and a make a decision taking the bias trade-off into consideration.

Furthermore, nearest neighbor matching assigns a subject in the treatment group from an unmatched closest subject in control group at each iteration. As a result the quality of the matches are defined by the order that the treatment subjects are matched (Rubin 1976a; Staurt, 2010). Rubin (1976a) considers three orderings of subjects: (1) random, subjects in both treatment and control are randomly ordered, (2) low to high, which means that the subject not yet matched with the lowest score (MD, PS, etc.) is matched next, and (3) high to low, which means that the subject not yet matched with the highest score (MD, PS, etc) is matched next. This of course does not come without complication. The quality of the matches will always be dependent on the order the researcher selects.

Calipers. It could be the case that the sample being used in a matching method does not contain "good" matches based on a set of covariates. For instance, the treatment group might be completely different than the control group, yet you will still get distance measures (MD) and/or

a Propensity Score. A common practice to avoid erroneous matches is to use a caliper (or similar metric) to put a definitional guideline to what constitutes a "poor" or "good" match (Stuart, 2010; Rosenbaum & Rubin, 1985a). A caliper sets the maximum (or minimum distance (MD or PS) that is allowed for each individual matched pair. A caliper takes on the following form (Heckman, Ichimura, Smith, & Todd, 1998):

$$C(X_j) = \{X_j \mid \left\| X_i - X_j \right\| \} < \varepsilon \tag{15}$$

Equation 15 translates in the following: For a subject in the control group matched to a subject in treatment group such that the normed distance $||X_i - X_j||$ is less than the specified caliper ε. When using calipers, it is possible however that no matched individual exists, that is there might be no such X_j for any observations in i (Gu & Rosenbaum, 1993). For caliper specifications and technical recommendations, see Cochran (1968) and Cochran and Rubin (1973). Rubin and Thomas (2000) combine Mahalanobis matching with propensity score calipers.

Replacement. An important step in the process of matching control units to a set of treatment units involves deciding whether matching will be done "with replacement" or "without replacement". Matching "with replacement" involves a scenario where a subject from the control group can be used as a match for a subject in the treatment group more than once. Matching with replacement involves a bias variance tradeoff not too different then when considering the number of individuals to match in the k:1 scenario in the previous section. In short, replacement of control subjects improves the average quality of the matching while also reducing the bias (Caliendo & Kopeinig, 2008). This comes in handy when you have significantly different distributions of causal groups. For example, consider the situation where subjects in the treatment group have high propensity scores, if we were to match control subjects only once and there was only a very limited pool of potential "good" matches of control units

with also high propensity scores, we would end up with some individuals with poor matches. However, if control units are allowed to be used more than once, then we can circumvent this issue. Additionally, there is an increase in variance from using fewer control group subjects while the decrease in bias is due to the use of "good" matches being used (Smith & Todd, 2005; Caliendo & Kopeinig, 2008).

Some issues associated with replacement include, using very few control observations to estimate an effect. Researchers should monitor the amount of subjects used for comparisons. Additionally, the interpretations of the inferences and the effects being estimated become a bit more complex when control units are being used more than once. In short control units that are being used more than once are no longer independent and frequency weights should be considered (Stuart, 2010).

Optimal Matching. A matching procedure that avoids the ordering effects of nearest neighbor matching on the quality of matches is Optimal Matching. Optimal matching a priori, takes into account all the potential matches from the overall set before making an individual match (P. R. Rosenbaum, 1995). It does this by implementing a network flow type algorithm in which matches for a subject in one group are made by taking into consideration the total unmatched subjects in the other group, and finding a combination of matches that minimize the distance (pertaining to the matching method) between individual matches (Fulkerson & Dantzig, 1955; X. Gu & Rosenbaum, 1993). As the algorithm iterates, if the case arises that an already matched subject from the control group is a better fit (smaller distance measure) for a current treatment subject, then the original match will be broken and the previously matched subject from treatment will now have a better fitting match with the current control subject (Bertsekas, 1991; X. Gu & Rosenbaum, 1993). As a result, Optimal-matching produces the best possible

individual matches, and does not in particular, focus on the overall sample means of the matching covariates (Gu & Rosenbaum, 1993; Rubin & Thomas, 2000).

Weighting Using the Inverse Propensity Score

An alternative approach to matching involves adjusting the confounding in the variables by estimating a propensity score to construct weights for individual subject observation using the inverse of nonparametric estimated propensity score (Imbens, Hirano & Ridder, 2003). Weighting involves adjusting the treated and control groups to that of a full sample, this method is analogous to Horvitz & Thompson (1952) who proposed a method for creating survey sampling weights for inference at the population level (Lunceford & Davidian, 2004; P. Rosenbaum, 1987; Stuart, 2010). Furthermore, weighting using the inverse probability of the propensity score has been shown to be an unbiased estimator of Average Treatment Effects (ATE) (Czajka & Hirabayashi, 1992; Rosenbaum, 1987).

Weighting adjustments are most commonly used only on the propensity score, however the guidelines on how to treat extreme weights, which happens when the PS is close to 0 or 1 or how to set weight maximums or minimums are not very developed (Stuart, 2010). Furthermore, the correct specifications of weights rely on the assumption that the propensity score model will be properly specified and estimated by the researcher.

Sub-Classification using the Propensity Score

Subclassification separates both treatment and control groups on the pretreatment confounding variables and forms groups or bins of subjects that are similar in propensity score values (Cochran & Rubin, 1973; Cochran, 1968). The treatment effects are then estimated within each stratum/group/bin (sub-class) and pooled to get a global estimate of the treatment effect. Cochran (1965; 1968) outlined the guidelines and justification for matching using

subclassification in the univariate case. Cochran (1968) presents two examples in which he compares the death rates of men with different smoking habits after subjects are subclassified (stratified) using the covariate of age. He also compares the frequency of type of sexual behavior among men after subclassifying subjects using the covariate of socio-economic status. In his paper, Cochran (1968) found that five subclasses on average remove 90% of the bias due to the covariate being used for the subclassifications.

However, similar to the dimensionality issue with the distance metrics, as you increase the number of covariates, the number of possible subclasses increases exponentially (Cochran, 1965). For example, consider the case were you have a number of covariates V each with 2 different categories (i.e. one covariate in V can be gender and indicator of whether or not someone graduated from HS). With V number of covariates with 2 categories there are 2^V possible subclasses, and as V gets larger there is a possibility that there will be some subclasses that only contain subjects from either a treatment or control group but not both (Cochran, 1965; Rosenbaum & Rubin, 1984).

Rosenbaum and Rubin (1984) take advantage of the balancing and scaling properties of the propensity score and show that stratifying on the propensity score using five subclasses, removes 90% of the bias due to the covariates. That is, subclasses formed from the scaled propensity score (probability between 0-1 for all covariates) also balances all V covariates (Rosenbaum & Rubin, 1984).

Optimal Subclassification/Full Matching

In subclassification the researcher is responsible for selecting the number of subclasses that will be used in the matching. Optimal-subclassification, or as known more commonly, Full Matching, selects the number of classes algorithmically (using network flow theory) (Gu &

Rosenbaum, 1993; Rosenbaum, 1991). Full Matching is in essence a combination of Optimal Matching (as described in the previous section), K:1 matching, and Subclassification discussed in the previous section. Full Matching creates subclasses that contain either (a) one treated unit and one control unit, (b) one treated unit and two or more control units, or (c) two or more treated units and one control unit (Guo & Rosenbaum, 1993). Treated and control subjects and their distances (i.e. PS, MD, GMD) are grouped into subclasses that minimize the average distance between subclasses.

The methods presented in the previous section are not contending methods but are potentially (given the right data and researcher goals) complementary. Propensity Score methods in practice are intended to be applied in an iterative process across the different scenarios while taking into consideration the general guidelines that do exists in the methodological literature (as outlined in this thesis), while keeping in mind the goal to achieve a comparable treatment and control units. This requires the applied researcher to have a combination of (1) an extensive background in the statistical and applied foundational underpinnings of different matching based methods, (2) breadth and depth of knowledge about the unique challenges in the implementation of matching based methods that can be field specific (i.e. data structures in education).

Issues and gaps

Most studies conducted for the evaluation of local, state, or federal programs, which rely on quasi-experimental methods rarely indicate any form of justification or an insight to the design process of the quasi-experiment. As a result, replication of such studies is somewhat untenable. The methods discussed in the previous sections, involve making very specific methodological decisions that depend on the data that is being analyzed. For example,

Rosenbaum (1986) using data from the High School and Beyond study (HSB) attempted to estimate the effects of dropping out of high school on student achievement test scores. In this study the "treatment" condition was whether or not a student dropped out of school during their sophomore year. Using a propensity score matching method, dropouts were compared to students who persisted with similar propensity scores. Since there were 1,105 high schools represented in the data, an important modeling decision had to be made. Would they create an overall pool of dropouts and a comparison group from all 1105 high schools? Or, would they control for school factors by matching students within individual schools? The later approach was used for the following reasons: (1) contextual factors that might influence the treatment conditions are constant for all students within each school (average per-pupil expenditure), (2) matching within each school controls for geographic variables (Urban vs. suburban vs. rural), and (3) eliminates the between school component of variation. Rosenbaum (1986) states: *If no adjustment was made for the high school that a student attends, then the analysis would tend to compare students who dropped out of economically disadvantaged schools with students who remained in somewhat wealthier schools, since dropping out is more common in disadvantaged schools*.

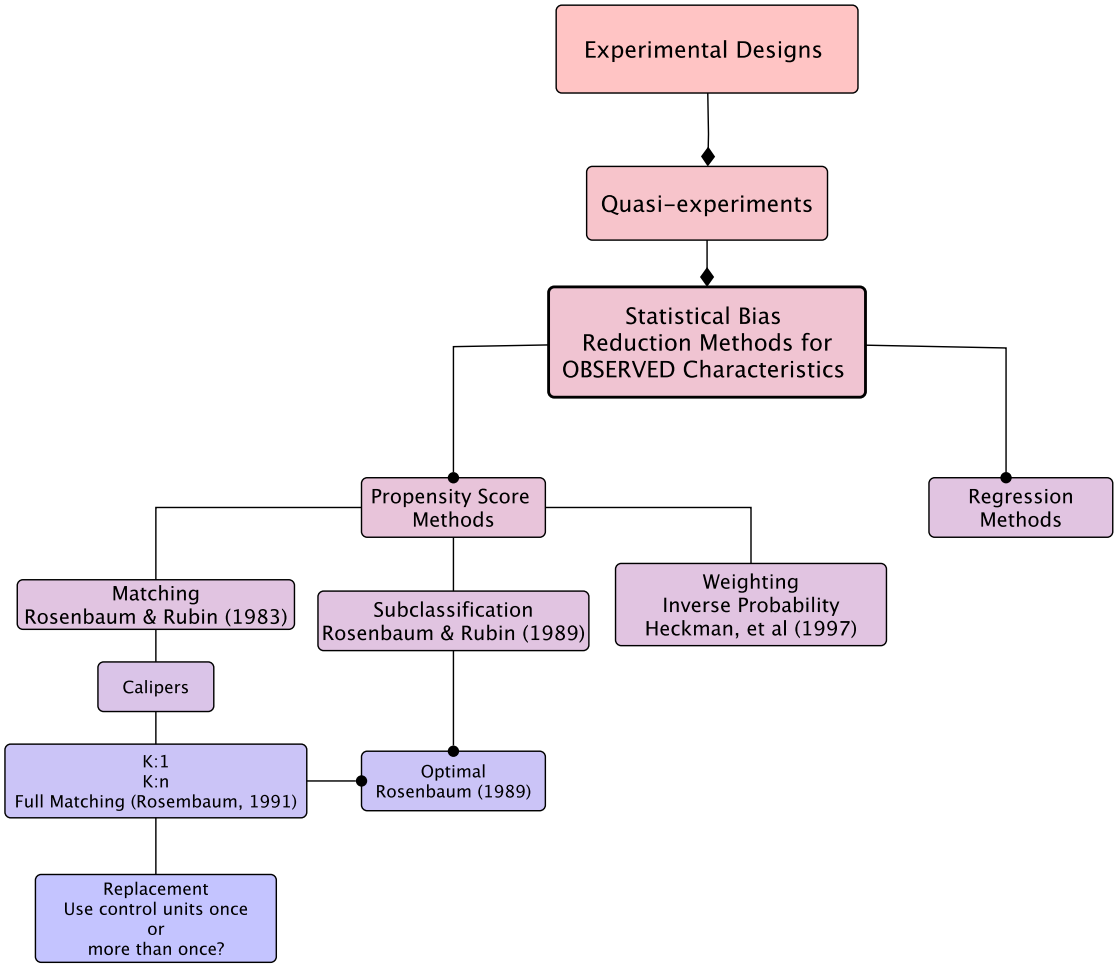
By ignoring the between-school variation and only incorporating within-school characteristics and focusing on the variation between individual subjects (students), proper estimates of treatment effects were plausible.

Given the many options available to researchers, which approach should be used? Figure 5 illustrates in summary, the possible options that are available to researchers as options for controlling confounding due to observed variables.

Additional complexities. Although work on the extension of these different methods to

multilevel modeling in an ongoing process, there doesn't seem to be any type of consensus or guidelines on the application of these merged methodologies within education research. Most of the applied work within education, (that was described in the previous chapter), focuses on unpacking the methodological specifics of the causal effect estimator and glance over the intricacies associated with making decisions about modeling multilevel data structures. This is somewhat concerning, considering that the decision to model random intercepts and random slopes in the multilevel modeling context is a both a data driven decision (does the variation exists?) and a model specification exercise (different models have to be specified and compared in terms of fit, etc.) (Gelman & Hill, 2007).

Figure 5. Bias Reduction Methods for Observed Characteristics



Research Questions

This study will try to inform some of these design decisions a researcher has to make when attempting to conduct a study involving commonly used bias reduction methods for observational studies. Furthermore, this study will explore the conditions for making the decision to utilize a multilevel modeling framework in terms of how much variability is captured between different clusters in the analysis. More specifically, this dissertation will answer the following questions:

1. Under what conditions do ordinary least squares-estimated (OLS) multiple regression techniques for estimating causal effects perform as well as propensity score method (PSM) techniques in terms of bias and coverage for estimating treatment effects?
2. Which PSM technique is preferred among the most common techniques found in the literature, and does it depend on certain conditions like sample size and variable correlation magnitudes?
3. What are the implications of ignoring the sample properties of clustered or nested observational data found in education in the context of estimation of Average Treatment Effects (ATE)?

Monte Carlo Simulation Methods

Monte Carlo Simulation Design

The present study employs a Monte Carlo (MC) simulation experiment for examining the relative bias and coverage in estimating Average Treatment Effects (ATE) cross five different treatment effect bias reduction techniques. Specifically, treatment effects were estimated using (1) Ordinary Least Squares (OLS) multiple linear regression (treatment effects, adjusted for mean differences on confounders), (2) Propensity Score Matching (PSM) using nearest neighbor 1:n with replacement, (3) Propensity Score Matching using Inverse Probability Weighting (IPW) of the propensity score, and (4) Propensity Score Matching using Sub-classification (Subclassification). There were five main factors that were varied to simulate the data, all of which were fully crossed, as follows.

- Four sample sizes (600, 1000, 2000, and 5000)
- Three association levels among simulated variables (low, moderate, high).
- Two treatment exposure levels (25% and 50%)
- Four treatment effect sizes using Cohen's *d* (none, low, moderate, and high)
- Five levels of ICCs (0, .10, .20, .30, and .40)

These 480 conditions were each analyzed with four methods of analysis, for a total of 1920 conditions. Sample size, association, and treatment exposure conditions were selected for ease of comparison with prior research on PSM models as well as natural conditions that are found in applied research. The exception was effect sizes: for this, the no-treatment condition was purposefully used as the "null" condition with which to evaluate baseline bias and coverage (analogous to determining Type I error rates) prior to the introduction of true treatment effects

(analogous to determining Power rates). Below I describe the rationale for these selections in more depth.

Sample Sizes. Sample sizes were selected based on both applied and theoretical literature. An extensive literature review of applied educational research literature ensures that the conditions specified mirror those found in the real world. From the many research articles reviewed (see Chapter 1 for review specifics), 19 used a type of quasi-experimental design similar to those explored in this study (i.e., matching, weighting, subclassification). Of those 19, 15% reported overall sample sizes ranging from 600 to 1,000 total cases; 21% reported sample sizes ranging from 1,000 to 2,000; 42% reported sample sizes ranging from 2000 to 5,000; and, lastly, 21% reported sample sizes ranging from 5,000 to 32,000 cases.

Sample sizes used in the prior simulation literature also varied considerably. Frölich (2004) uses sample sizes of 100 and 400 cases. Lee, Lessler, and Stuart (2009), employed sample sizes of, 500, 1,000, and 2,000; Hade and Lu (2014), used sample sizes of 500 and 2,500; Arpino and Mealli (2011), used sample sizes ranging from 500 to 4,000; and Thoemmes and West (2011), explore sample sizes up to 100,000. To capture the range of sizes observed but still limited for the sake of brevity, the present study is based on total sample sizes of 600, 1,000, 2,000, and 5,000 cases. To capture the multilevel structure of the data, the different sample sizes were divided up into even clusters of 30, 40, 50 and 100, respectively. These cluster conditions were selected because they represent a range of cluster sizes (from small to large) based on what is commonly found in educational data (Maas & Hox, 2005).

Correlation Structure among Variables (3 levels). Using three predefined correlation structures, inter-variable relationships among the Xs were fixed at three different levels:

1) Low relationship, $r = 0.10$

2) Moderate relationship, $r = 0.30$

3) High relationship, $r = 0.50$

These correlations were then translated into regression coefficient weights (for generating data) using a Cholesky decomposition of the pre-specified correlation matrix (Pourahmadi, 2007). This method is analogous to Austin's (2011) simulation procedure using a Multivariate Normal distribution to generate Ys. The decision to assign a correlational structure to the data was driven by two ideas:

1) Researchers can look at the correlational structure of the data prior to any analysis and make decisions about the appropriate approach to analysis within the context of quasi-experimental designs.
2) High or low relationships (in terms of correlations) between covariates and treatment identification variables can have different impacts on estimated propensity scores and treatment effects (Keisuke Hirano & Imbens, 2001; Donald B. Rubin & Thomas, 1996).

An alternative approach to assigning relationships between covariates and the treatment identification variable, involves the explicit modeling of the treatment identification variable conditional on a known set of covariates. This method is popular amongst the most recent PSM simulation studies (Arpino & Mealli, 2011; Austin & Small, 2014; Alexis Diamond & Sekhon, 2012; Lee et al., 2009; Zhong Zhao, 2008). The explicit modeling of the probability of treatment exposure is effective within the simulation context, however it has little relevance in practice because the “true” functional form of the probability of treatment exposure is rarely known in observational studies (Hade & Lu, 2014; Donald B. Rubin & Thomas, 1996).

Treatment Exposure Level (2 levels). Treatment exposure level was not commonly reported in the reviewed literature. This is in fact very problematic in the context of PSM, where the analysis consists on the subset of exposed units in the data. Because it is not openly reported, it is sometimes difficult to establish the sub-sample of cases that have been exposed to a treatment in an observational dataset.

This study will use exposure levels of 25% (that is a ratio of 1:4 of treated to control units) and 50% (a ratio of 1:1), which reflect ranges of exposure levels found in the simulation literature within the context of quasi-experiments and PSM (Frölich, 2004; Hade & Lu, 2014; Donald B. Rubin, 1979; Zhong Zhao, 2004). The condition when only 25% (1:4) of the population is exposed to a treatment effect is the most pertinent to what is found in practice.

Effect Sizes (4 levels). Cohen characterizes effect sizes in the social sciences as follows: an effect of .20 in magnitude is considered small; +.50 as moderate; +.80 as large (Cohen, 1988). In education research however, effect sizes of +.80 and even +.50 are rarely seen in practice. For example, many education intervention studies that investigate effects on academic performance using standardized measures of math and reading test scores are rarely above .30 (Cohen, 1988).

This study will explore four different effect sizes, that in relative terms reflect small to large effects found in practice, these include: Cohen's $d = 0$ (no effect); 0.20 (small effect), 0.50 (moderate), and 0.80 (high). Given that in this study the effect was to be expressed as the beta coefficient for participation in the treatment effect β_{tx} in equation 17, the effect size of treatment effect on y and the betas for the additional covariates were estimated using the following relationship (Cohen, 1988):

$$\beta_{Y\,tx} = r\frac{\sigma_Y}{\sigma_{tx}} \tag{16}$$

Where r is the correlation coefficient corresponding to the different effect size magnitudes of low, medium, high. Cohen (1988), characterizes an effect size of $d = 0.20$ (low) as a correlation coefficient of $r = .10$; $d = 0.50$ (moderate) as a correlation coefficient of $r = 0.30$; $d = 0.80$ (high) as a correlation coefficient of $r = 0.50$. These correlations were use and converted to beta coefficients used in the data generation model.

Intraclass Correlations (5 levels). The intraclass correlation coefficient (ICC) value is an indicator of within-cluster dependency. More specifically, the ICC is the fraction of the total variation in the data that is accounted for by the between group variance as is evident in equation 18 (Gelman & Hill, 2007; Searle, Casella, & McCulloch, 1992; Snijders & Bosker, 1999). In practice ICC levels are typically small/moderate and range from .05 to .30 in educational research (Raudenbush & Bryk, 2002). However it has been found that even small levels of ICC values can have an impact on the estimated residuals (Musca et al., 2011).

The preferred ICC values were obtained by varying σ_{α}^{2} in the simulations and using the relationships expressed in equation 18, following the simulation designs by Maas & Hox (2005), Moineddin, Matheson, & Glazier (2007), Thoemmes & West (2011), and Vallejo, Fernández, Cuesta, & Livacic-Rojas (2015).

The simulation values of ICC considered for this study are 0, .10, .20, .30, and .40, which are plausible values for educational clustered data found in practice (Niehaus, Campbell, & Inkelas, 2014). When there is no association between the subjects in a cluster, (ICC = 0) and is analogous to single level data. An ICC of .40 can be considered an exceptional case in most studies, however the simulation literature tends to explore ICC values of up to .60 (Moineddin et al., 2007; Paccagnella, 2011; Thoemmes & West, 2011).

Data Generation (Simulation)

For each condition, 10,000 datasets of six variables were generated for use in five analyses. First, four variables (X_1, X_2, X_3, X_4) are randomly drawn from a unit normal distribution. Next, an exposure condition variable was then simulated (*t*) to match the proportion of the population that was simulated to be exposed to the treatment condition. Lastly, Y was simulated following the equation given below:

$$y_i = a_{j[i]} + \beta_{tx} t + \beta_x x_i^T + e_i \text{, for i = 1, …, n,} \quad (17)$$

$$e_i \widetilde{iid}\ N\left(0, \sigma_y^2\right),$$

$$a_j \widetilde{iid}\ N(\mu_\alpha, \sigma_\alpha^2), \text{ for j= 1,…,J,}$$

where β_{tx}, is the estimated effect when comparing treatment effects participants (*t*=1) to the control (*t*=0). The variability of random effects (σ_α^2) and residual error (σ_y^2) relate directly to the ICC in the following form:

$$ICC = \frac{(\sigma_\alpha^2)}{(\sigma_\alpha^2 + \sigma_y^2)} \quad (18)$$

The correlation between independent variables were determined by multiplying the simulated data frame with the Cholesky decomposition of the correlation matrix "R" (3 levels), with:

- D = Simulated data frame containing X_1-X_4 and tx.
- R = 5x5 pre-specified correlation matrix.
- Correlated data = [Cholesky(R)] * D, and

then outcome y_i was created using equation (17) above. Regression coefficients for X_1-X_4, were fixed and corresponded to the different correlation levels used in this simulation study. Correlations were converted to beta coefficients following the predetermined relationships between X_1-X_4 and β_{tx} (Cohen, 1988; Pehazur, 1982). For each dataset simulated, four

analyses were conducted (as previously described), and for each of the five models, results were saved for secondary data analyses. Note that the assumption of common overlap was held constant for the present study throughout all conditions, and a frequentist approach to all estimation methods was employed.

A notable alteration to the multilevel data simulation approach used by Maas & Hox (2005) and Moineddin, Matheson, & Glazier (2007), involves assigning a pre-specified correlational structure to the simulated variables. This, in addition to the manipulated components of the outcome model in equation 17 used to obtain the desired ICC values, could have potential implications for the variance components of the overall simulated data. This study will focus on the bias associated with the different estimation methods of treatment effects, however future studies will focus on the variance components related to this specific data simulation design. The Monte Carlo results presented in Chapter 4 are unique to the multilevel data simulation strategy adopted in this study.

Analysis Approach and Evaluation Criteria

For each of the 10,000 datasets across each of the 480 conditions, the Average Treatment Effect (ATE) and corresponding standard error were estimated using the following four analysis methods:

1. Ordinary Least Squares (OLS) multiple regression (treatment effect estimated after adjusting for mean differences between groups on four confounders).
2. Propensity Score Matching using nearest neighbor 1:n (PSM) with replacement
3. Propensity Score Matching with Inverse Probability Weighting (IPW) using the Propensity Score

4. Propensity Score Matching using Sub-classification matching (Sub-classification), using the optimal recommendation of five subclasses (X. S. Gu & Rosenbaum, 1993a; P. R. Rosenbaum & Rubin, 1984)

To evaluate the accuracy and precision of the simulations the following quantities were saved and examined in secondary analyses.

Bias. Both raw bias, defined as the difference between the true treatment effect (θ) defined for the population and the estimated treatment effect ($\hat{\theta}$), averaged across 10,000 replications, and Relative Bias, which measures the average tendency that the simulated treatment effect ($\hat{\theta}$) is below or above the true treatment effect (θ) in terms of a percent, were used to assess whether a given estimation method is over or underestimated the treatment effect.

Mean Squared Error (MSE). Represents the squared deviation of the estimated treatment effect($\hat{\theta}$), from the true effect (θ), average across 10,000 replications. The MSE is a combination of bias and variance and is a measure of the overall variability of the repetitions.

Coverage. The coverage probability represents the amount of times the true treatment effect is contained within the confidence interval of the estimated treatment effect throughout the 10,000 replications. The confidence interval is computed using the estimate of the treatment effect and estimated standard error for each repetition of the estimated parameters. An estimator that is performing well will contain the true treatment effect closer to the nominal value of 95%. Confidence interval inspection can lend insight into the many sources of bias associated with the different factors being tested as well as bias present in the standard errors of the different estimation procedures.

Secondary Analyses of Simulation Results

The saved simulation results described above were summarized into means across the 10,000 replications for each of the conditions, for each of the different estimation methods. For secondary analyses, two dependent variables were used: relative bias, which take the average tendency that the simulated treatment effect is below or above the true treatment effect and allows for better comparison across methods, and coverage, which allows us to understand the extent to which the confidence interval is accurate. Analyses of variance with main effects and 2-way interactions were examined to understand the effects of the MC simulation conditions on relative bias and coverage. Significant main effects with large effect sizes where then followed up with post-hoc pairwise contrasts.

Monte Carlo Simulation Results

The results of this study are presented in the following sections. The Monte Carlo simulations were first evaluated in terms of convergence. Overall, 100% of all models converged successfully. In order to fully summarize the simulation results the different conditions that were fixed in this simulation study were examined via a factorial Analysis of Variance (ANOVA) with selected two-way interactions for both bias and coverage of simulated values. In order to properly address and interpret potential significant effects within the different ANOVAs, the effect size, ω^2, was used. ω^2 is estimated as:

$$w^2 = \frac{SS_{Between} - (J-1)MS_{Within}}{SS_{Total} + MS_{Within}} \tag{19}$$

This effect size was chosen over other commonly used effect size estimates due to its properties associated with less bias. According to Cohen (1988), $\omega^2 = 0.01$, 0.06, and 0.14 are interpreted as small, medium, and large effect sizes.

Descriptive Statistics

Simulation results were summarized in the form of means across the 10,000 replications. An initial descriptive analysis provided insight to the potential presence of bias in the different estimation methods used as well as discernable patterns across conditions. As mentioned in the previous chapter, for each estimation method, means were computed for the treatment effect estimate, Raw bias, Relative bias, Mean Squared Error (MSE), and 95% CI Coverage.

As shown in the tables (12 - 14) provided in the appendix for reader interest, raw bias estimates indicate some deviation from the true effect for most methods. These summary tables attempt to disaggregate the results by all their conditions. The tables also show results for a method that assumes random assignment ("Naïve method"). Overall, there is some evidence that

the subclassification estimator is associated with greater values of both raw and relative bias. When you focus on the correlation levels, higher correlation levels seem to increase the bias for all methods. Due to the large number of conditions, mean data from all conditions were evaluated with (a) main effects only analyses of variance (ANOVAs) for null conditions (to establish baseline bias and coverage), and (b) main effects and 2-way interactions among the simulation conditions and the four methods of analysis across all conditions. Notably, because of the precision of estimation for each condition (due to use of 10,000 replicates each), only the ANOVA results with substantial effect sizes were considered for further inspection.

Null Effects of the Single Level Model Results

For data from the null conditions (in which there was no true difference between groups), the results of a five-factor ANOVA with main effects only on relative bias showed significant main effects for method of analysis, $p < 0.001$, $\omega^2 = 0.46$; sample size, $p < 0.001$, $\omega^2 = 0.01$; and correlation level, $p < 0.001$, $\omega^2 = 0.19$. The sample size main effect was ignored due to the negligible effect size ($\omega^2 < 0.01$). The method main effects and the correlations main effects were further evaluated using Tukey's HSD. Tukey's HSD resulted in significant differences between the subclassification method and other analysis methods, OLS, PSM, and IPW, ps < 0.001. The mean difference between the subclassification and the other estimation methods ranged from 0.010 to 0.015, indicating a very small difference in the levels of bias in the null condition of estimation method (there were no differences in bias across the other methods of anlaysis). Similarly, the correlation main effect was further evaluated using Tukey's HSD multiple comparison procedure. These results showed significant differences between the correlation level of 0.50 and the other correlation levels, 0.10 and 0.30, ps < 0.001. The mean

difference between 0.50 and the other levels ranged from 0.006 to 0.010, indicating again a very small difference in the levels of bias due only to correlation magnitudes.

For empirical coverage rates, the results were similar to those for bias: there were main effects for method, $p < 0.001$, $\omega^2 = 0.30$, and correlation level, $p < 0.001$, $\omega^2 = 0.02$, but not sample size. Tukey's HSD post-hoc results showed again significant differences between the subclassification method and other methods, OLS, PSM, and IPW, ps < 0.001. The mean difference in coverage rates between the subclassification and the other estimation methods ranged from 0.01 to 0.03. Tukey's HSD results also showed significant differences between the correlation level of 0.50 and the two other correlation levels, 0.10, and 0.30, ps < 0.001. The mean difference in coverage between 0.50 and the other levels ranged from 0.007 to 0.020. Although these differences are small, they should be kept in mind as baseline bias and coverage rates.

All Condition Results for the Single Level Data Model

Bias. As shown in Table 5, the results of the ANOVA on relative bias testing main effects of study conditions as well as their two-way interactions with methods of analysis, showed significant main effects for analysis method, effect size level, sample size level, and correlation level. There were also two significant interactions between analysis method and effect size, and analysis method and correlation level. Again, due to the precision associated with using 10,000 replications per condition, not every statistically significant effect will be "practically significant". Thus, only non-negligible effect sizes were be examined for follow-up comparisons. The results showed that both the effect size and sample size main effects on bias were non-substantial, as was the analysis method by effect size interaction, but that analysis

method and correlation level had large effects on bias ($\omega^2 = 0.50$ and $\omega^2 = 0.15$, respectively), as was the interaction between the two main effects.

Table 5. ANOVA Results for Single Level Data: Bias Outcome

ANOVAS for All Main Effects and Method Interactions

Dependent Variable	Bias in Average Treatment Effect Estimation	
Source	F	w^2
Analysis Method	1429.94 ***	0.50
Effect Size (ES)	20.64 ***	0.00
Sample size (N)	39.23 ***	0.01
Correlation	648.42 ***	0.15
Treatment:Control (T:C)	0.62	0.00
Method * ES	10.60 ***	0.00
Method * N	1.46	0.00
Method * Corr.	464.99 ***	0.32
Method * T:C	0.22	0.01

note: T:C= Treatment and Control group size; ***p<.001; **p<.01; *p<.05

Specifically, bias associated with analysis method indicated that some methods were associated with greater bias in the estimation of treatment effects, and that ignoring the correlation levels of the measured variables could potentially bias overall results.

Plots of mean bias by analysis method and correlation level given in Figures 6 and 7 illustrate the effects of the subclassification method and high correlation levels on bias, respectively. Tukey's HSD follow-up tests revealed that the subclassification method shows the largest statistically significant difference $p < 0.001$, in terms of bias, between all methods (.025) or about 2.5%. Although a 2.5% increase in bias seems large relative to the other estimation methods, however it is not large in magnitude. Similarly, Tukey's HSD follow-up tests on the main effect of correlation levels revealed higher bias associated with higher correlation levels,

the largest difference is between low correlation and high correlations with a bias difference of about .02 (2%).

Figure 6. Subclassification Methods Relative to the other PSM and OLS Methods Shows the Most Bias

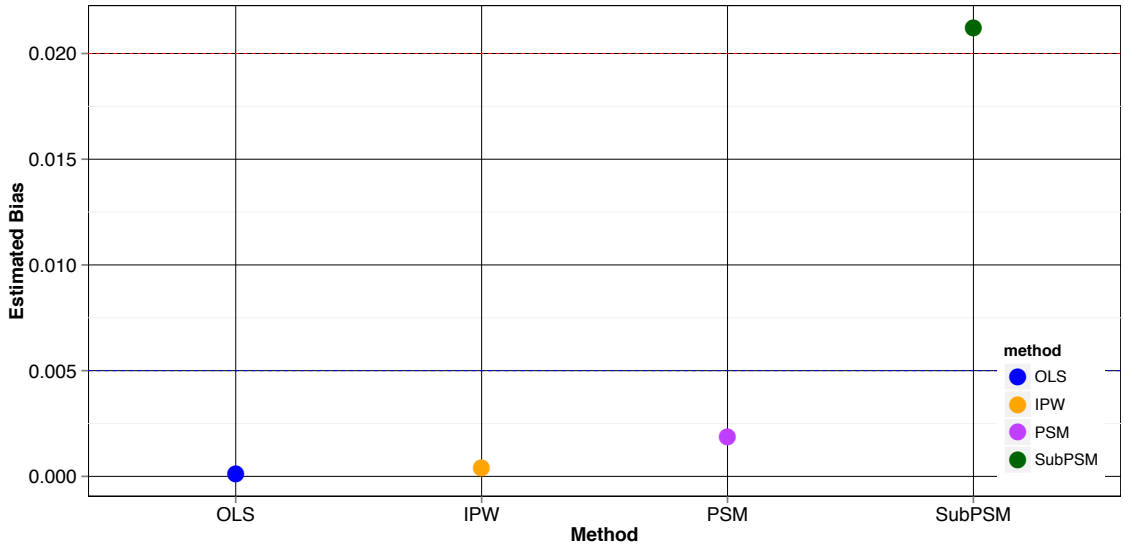
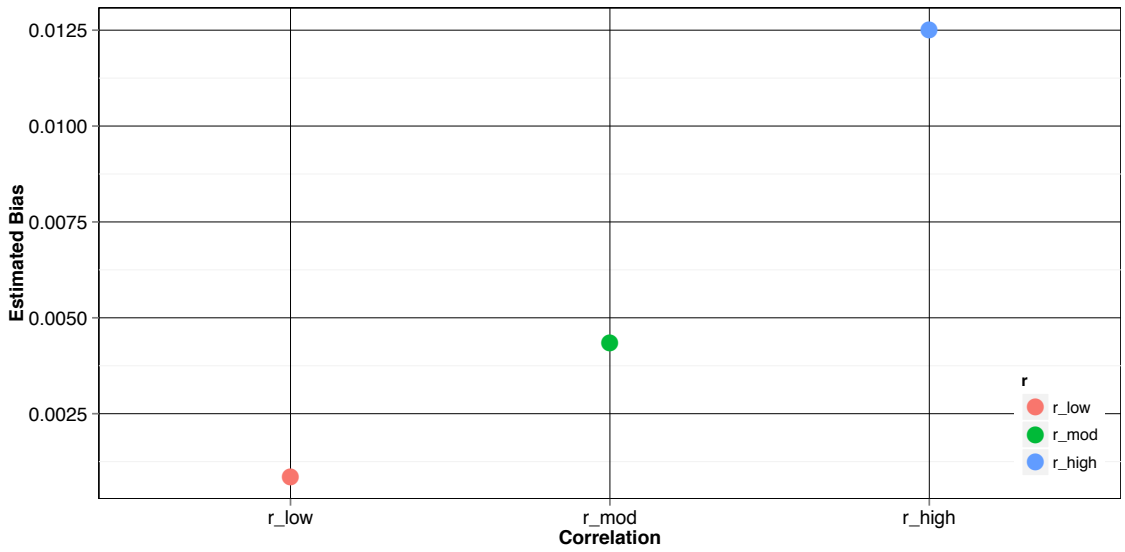


Figure 7. Correlation Level "r" Main Effects Across Three Levels (0.10, 0.30 , 0.50)

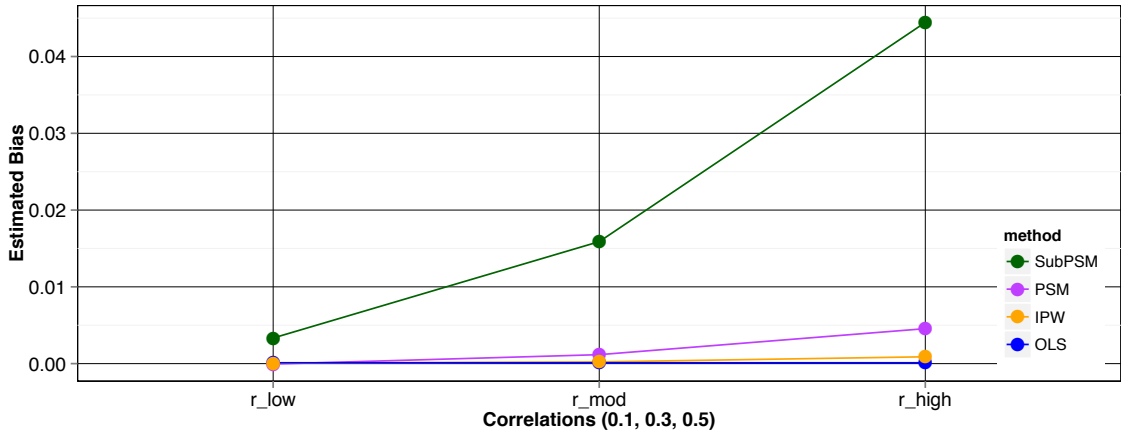


Finally, as can be seen in Figure 8, the interaction effect between analysis method and

correlation magnitude on bias shows that the effect of the magnitude in bias on the method increases as a function of increase magnitudes of correlation for the subclassification method in particular. Additionally, propensity score matching seems to follow a similar trend, however the magnitude of the bias is negligible (bias < 1%). In truth, the matching method along with OLS and the IPW are associated with the lowest bias across all conditions.

All other effects observed in Table 5, although statistically significant, had negligible effect sizes, as a result they are not interpreted any further.

Figure 8. Estimation Method Bias as a Function of Method and Higher Correlations



Coverage. Coverage rates were assessed for each estimation method's ability to recover the true treatment effect given the estimated mean and standard error. The previous analysis (on the estimated bias rates) showed that there were significant and substantial method and correlation effects, as well as their interaction. As shown in Table 6, the ANOVA on the 95% coverage showed nearly identical results; however, in addition to substantial method and correlation effects, there was also a substantial interaction between method type and the ratio of treatment:control group sizes. Table 7, along with Tukey's HSD follow-up analysis on the method main effect, indicated that the subclassification estimator had lower CI coverage rates relative to the other methods, by about 2%.

Figure 9 illustrates that, for higher correlation magnitudes, the subclassification estimator is associated with far worse CI coverage, at 83% instead of the nominal 95%.

Table 6. ANOVA and Method Interactions for Coverage Rates

ANOVAS for All Main Effects and Method Interactions

Dependent Variable	Coverage Rate for Average Treatment Effect Estimation		
Source	F		w^2
Analysis Method	219.39	***	0.50
Effect Size (ES)	17.03	***	0.02
Sample size (N)	4.17	*	0.00
Correlation	10.51	***	0.01
Treatment:Control (T:C)	8.60	**	0.01
Method * ES	3.46	**	0.01
Method * N	4.23	***	0.01
Method * Corr.	76.48	***	0.34
Method * T:C	26.34	***	0.06

note: T:C= Treatment and Control group size; ***p<.001; **p<.01; *p<.05

Table 7. Empirical Coverage Rates by Method

Empirical Coverage Rates

Method	Estimate	SD
Traditional Regression (OLS)	0.95	0.002
Propensity Score Matching (PSM)	0.95	0.004
Inverse Probability Weights (IPW)	0.96	0.009
Subclassification (SubPSM)	0.93	0.01

note: Nominal coverage rate = 95%

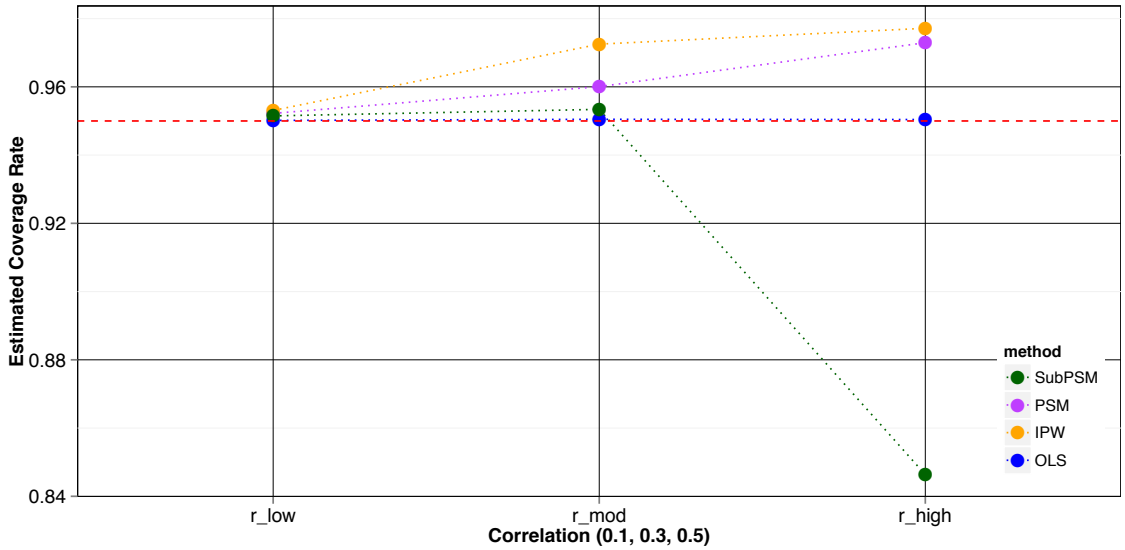


Figure 9. 95% CI Coverage by Correlation Magnitude

Varying ICC Values

Varying levels of ICC was evaluated for bias. The results of the ANOVA for the intraclass correlation (ICC) level main effects and their two-way interactions across the main factors showed significant main effects for ICC, $p < 0.001$, $\omega^2 = 0.00$. Due to the negligible effect size, the ICC main effect was not further explored. There was, however, a significant ICC and levels of sample size interaction, $p < 0.001$, $\omega^2 = 0.10$. As shown in Figure 10, bias increased as a function of ICC for sample sizes of 600 and 1,000; with most bias due to ICC eliminated in larger sample sizes (2,000 and 5,000).

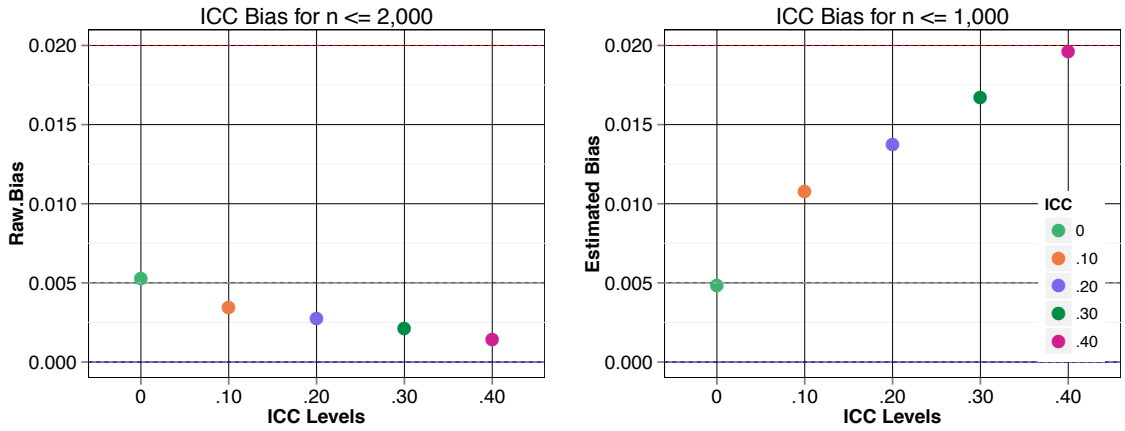


Figure 10. Bias Increase as a Function of Sample size n and ICC

All other ICC interaction effects were both not statistically significant and had an effect size that was nil to negligible. Those will not be addressed or interpreted further.

Applied Analysis Demonstration with Multilevel Modeling

To illustrate the use of quasi-experiments in education research, the following study illustrates the process for identifying a treatment condition that has not been randomly assigned and potential confounding variables that can be used for the estimation of treatment effects. This study aims at estimating the causal effects of participation in a college preparation program (Gear Up/Upward Bound) and how the program influences a student's postsecondary aspirations in the form of the number of four-year colleges a student choses to apply to.

The data for this applied study were obtained from the Educational Longitudinal Study of 2002 (ELS:2002), through the National Center for Education Statistics (NCES). ELS is a nationally representative sample of over 15,000 respondents in over 700 schools. The sample for this study included students in the first two respondent waves in the 10^{th} and 12^{th} grades, and is made up of 8,104 students representing 694 schools. A total of four variables were chosen because they have been shown to influence a student's decision to pursue a college education or to participate in a program designed to help students apply to college. Descriptive statistics are presented in the Table 8. The participation variable indicates that about 10% of the respondents participated in a College Preparation program. The key outcome of interest is the number of applications submitted by the student in the 12^{th} grade.

Table 9 presents the correlations between confounding variables, outcomes, and treatment exposure variables. It should be noted that the participation indicator has small to moderate correlation between confounding variables, most importantly it has a small correlation with the outcome of interest ($r=0.03$).

Table 8. NELS 2002 Variables for use in the Applied Study

Variables	Description	M	SD
Dependent Variable			
Number of Apps	Total number of applications submitted to 4-year colleges in the 12th grade	2.8	2.16
Participation Indicator			
Participation in Program	Student participated in a college preparation program during HS (yes=1)	0.1	0.21
Parent Ed	Parent education indicator ranges from 1 (did not finish HS) to 8(Completed PhD, MD, or advanced degree)	4.8	2.04
Parent College Discuss	A measure that captures the frequency in which students and parents discuss college plans 1 (never) to 4 (often)	1.5	0.48
Taken ACT SAT	Student took the ACT or SAT (yes=1)	0.85	0.28
Student Aspirations	Students educational aspirations from 1 (High School graduation) to 6 (Obtain a PhD, MD, or advanced degree)	4.4	0.99

Note: Educational Longitudinal Study of 2002 (ELS:2002). Sample size is 8,104 respondents.

Table 9. Correlation of Variables form NELS 2002 Applied Study

	Parent Ed	Number of Apps	Parent Discussion	Taken ACT SAT	Participation in Program	Student Aspirations
Parent Ed	1	0.25	0.16	0.19	0.05	0.3
Number of Apps	--	1	0.28	0.31	0.03	0.39
Parent Discussion	--	--	1	0.35	0.2	0.36
Taken ACT or SAT	--	--	--	1	0.02	0.4
Participation in Program	--	--	--	--	1	-0.01
Student Aspirations	--	--	--	--	--	1

Methods

To estimate treatment effects, the same four estimation methods used in the simulation study were utilized for the present dataset, as follows.

- Ordinary Least Squares (OLS) regression using the four control variables.
- Propensity Score 1-1 matching with replacement (PSM).
- Inverse Probability Weights using the Propensity Score (IPW)
- Sub-classification using the PS (Subclassification) using five subclasses.

Of interest is how this applied analysis fits in with the various simulation scenarios conducted in the simulation study. Prior to implementing the different treatment effect estimation methods,

the ICC was estimated by running an unconditional model of the treatment effect outcome. The ICC value was estimated at .05, indicating that 5% of the variation can be accounted for by cluster level differences. Furthermore, the effect size of the treatment condition on the effect outcome was estimated at .04, indicating a very small effect and large sample size (n=8,104). Table 10 illustrates the simulation scenarios that most likely match this applied analysis. If we had a large sample and all the other simulation scenarios were closely matched with this study, we would expect that all methods with the exception of the naïve estimator, would provide similar treatment effect estimates. Furthermore, due to the low relationships between the confounders, treatment, and the outcomes, even the naïve estimator would show reduced bias rates relative to the other simulation scenarios (bias = .01), however its standard errors would still be somewhat biased (coverage = 90%).

Table 10. Monte Carlo Simulation Results Relevant to Applied Study

Monte Carlo results for n=5,000, r=.10, ES=0.10, ICC=0

Method	Estimated Effect	Bias	se	MSE	Coverage
OLS	0.071	0.000	0.052	0.003	0.95
PSM	0.071	-0.001	0.060	0.004	0.95
IPW	0.071	0.000	0.052	0.003	0.95
SubPSM	0.074	0.003	0.052	0.003	0.95

Note: Each cell is the average across 10,000 replications

Results

Table 11. Results of Different Treatment Effect Estimation Methods

Method	Main Program Effect	SE	
Naïve	0.10	0.11	
OLS regression	0.21	0.10	*
Propensity score matching	0.17	0.12	
Inverse probability weights	0.15	0.11	
Subclassification PSM	0.20	0.11	
ICC	0.05		
Effect Size*	0.04		

Note: PSM sample consisted of 809 respondents; * p > .05; Effect Size* = treatment on outcome.

The results are presented in Table 11. The naïve estimator produced the lowest treatment effect estimate at (0.10). The OLS and subclassification estimator produced similar results (0.21 and 0.20) respectively. The propensity score matching estimator (PSM) and the inverse-probability weighting (IPW) estimator produced similar results (0.17 and 0.15).

Summary

In order to fully utilize the simulation results and translate them into an applied study, the different conditions manipulated in the simulation must match (almost exactly) what you will find in practice. Although the simulation studies presented here have value in informing the applied field, it is important to note how this applied study doesn't exactly match the simulation conditions. For example, as seen in Table 9, correlation levels vary between observed variables, which is different from the fixed correlation levels tested in the simulation. Additionally, the treatment exposure level and ICC values found in the applied study (10% and 5%) respectively were not tested directly in the simulation. All these discrepancies can have an impact on the different estimators, and comparisons must be made carefully.

Discussion

The goal of the Monte Carlo simulation study was three-fold. First, the finite sample properties of the most utilized quasi-experimental methods that control for observable selection bias in the field of education were examined and compared to traditional regression methods. Second, an insight into the effects of ignoring the multilevel structure of data commonly found in the field of education was explored. Lastly, the approach of this study was to utilize a "data-driven" design similar to Hade and Lu (2013) within the simulation process to better address the applied field as a whole. This last goal, which deviates from the traditional model-based (the functional form of treatment assignment is assumed to be "known") approaches most commonly adopted in quasi-experimental simulation studies, is an important contribution to the field of education.

The first phase of this study provided insight to the conditions for which propensity score methods provide similar results to traditional regression methods. Prior meta-analysis research has shown similar results in treatment effect estimation between propensity score matching and traditional regression methods (Shah, Laupacis, Hux, & Austin, 2005). Most applied studies within education that conduct a type of PS method (matching, weighting, subclassification) and a traditional regression methods (OLS) as a form of sensitivity analyses for the purposes of comparison, show modest differences between treatment effect estimation results (An, 2013; Bhatt & Koedel, 2012; Long, Conger, & Iatarola, 2012; Melguizo & Wolniak, 2012). In one of the most recent applied studies conducted on the impact of charter schools and their effects on student academic achievement, their propensity score matching estimated effect when compared to traditional regression methods were similar in magnitude, direction, and significance (Furgeson, McCullough, Wolfendale, & Gill, 2014). The initial study showed significant

differences in bias associated within propensity score methods and traditional regression analyses. In short, when there are strong relationships between measured variables (i.e. correlations) and the sample size is small, there is slightly higher bias associated with propensity score matching when compared to OLS, and moderate to large biases associated with subclassification methods when compared to all other methods.

Additionally, prior research has shown that ignoring the multilevel structure of the data within quasi-experimental designs leads to bias in the estimation of treatment effects (Arpino & Mealli, 2011; Thoemmes & West, 2011). This study shows a modest increase in bias as a function of higher ICC values when the multilevel structure of the data is ignored at small sample sizes and the confounding variables do not vary by cluster. These results are consistent with Theommes and West (2011), who found that when you ignore the multilevel structure of the data in the estimation of treatment effects, there's a (.03) difference from the bias in the case when the ICC = 0, to the case where the ICC=.50, for their small sample condition. Additionally, the bias associated with increased ICC values diminishes for large sample conditions both in this study and in previous research findings. An important finding in the current study that adds to the existing simulation literature on MLM and propensity score methods is that it looks at different levels of values of ICC in addition to various possible correlational structures of the data and effect size estimates.

The following are important recommendations that can be ascertained from the results of this simulation study and be used by practitioners who have complex multilevel data and wish to utilize a quasi-experimental method for inference:

- A first step in dealing with educational data that might have clustering of subjects within different is identifying the level of variation associated with potential clustering. The

simulation results indicated that when you have low to moderate ICC levels (.10-.20) and there are no level-2 confounding variables of interest, observed bias is minimal.

- When you have variation at the cluster-level in the form of a random intercept, but no significant level-2 variation in the form of a random slope, the increase in bias from not modeling the level-2 variation for larger sample sizes and low to moderate ICC values is negligible for all methods.
- When data for use in the estimation of treatment effects has strong associations between control variables and treatment variables and the sample size is small, OLS along with weighting PS methods should be utilized over alternative methods.

The applied study was intended to first, substantively sort through the effects of a pre-college program on a student's postsecondary motivation to apply to college. Second, it was intended to illustrate the previous recommendations by looking at the decisions any researcher has to make in the design process of any study. Lastly, the applied study sheds light on some of the limitations of the simulation study that informs future directions for this type of research. For example, each of the five estimation methods was applied to the data for which we were interested in looking at the effects of a student pre-college program and their postsecondary opportunities. Overall, small to modest differences were observed between the different estimation methods. This applied study best fits the simulation scenario where observed variable relationships are either low or moderate. The overall ICC is value is small, and the sample size is large $N < 5{,}000$. Under these conditions, the treatment effect estimates aren't expected to be very different. However, an important limitation of the simulation scenarios is obvious, the correlational structure of the variables in any study will very seldom be exactly the same.

Extensions

Although this simulation is limited because it assumes the correlations among the covariates were the same, in practice relationships amongst variables will vary from low to high. Understanding how these varying relationships effect treatment affect estimation methods will be important in future research. An important extension of this research is to include level-2 confounders that contain both fixed and random slopes to fully address the scenarios found in applied research. Another important extension of this research should include situations when the treatment and control subjects vary by confounding variables. Propensity score matching methods might be better suited for studies where control and treatment units have significant overlap between confounding variables (Frölich, 2004; Hade & Lu, 2014). An important question that is echoed in these studies is how well a simple regression estimator would perform under these specific conditions.

References

A. Smith, J., & E. Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1-2), 305–353.

An, B. P. (2013). The influence of dual enrollment on academic performance and college readiness: Differences by Socioeconomic Status. *Research in Higher Education*, *54*(4), 407–432.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics : An empiricist's companion*. Princeton: Princeton University Press.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150–161.

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine*, *33*(24), 4306–4319.

Bertsekas, D. P. (1991). *Linear network optimization : Algorithms and codes*. Cambridge, Mass.: MIT Press.

Bhatt, R., & Koedel, C. (2012). Large-scale evaluations of curricular effectiveness: the case of elementary mathematics in Indiana. *Educational Evaluation and Policy Analysis*, *34*(4), 391–412.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Chapin, F. S. (1938). Design for social experiments. *American Sociological Review*, *3*(6).

Chapin, F. S. (1947). *Experimental designs in sociological research.* New York: Harper.

Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295–313.

Cochran, W., & Rubin, D. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, *35*(4), 417–446.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

Dehejia, R. H. (2003). Was there a Riverside miracle? A hierarchical framework for evaluating programs with grouped data. *Journal of Business & Economic Statistics*, *21*(1), 1–11.

Dehejia, R., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, *84*(1), 151–161.

Diamond, A., & Sekhon, J. (2014). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, Forthcoming.

Fisher, R.A.(1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503–513.

Fisher, R.A. (1935), *Design of Experiments*. Edinburgh: Oliver & Boyd.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics, 86*(1), 77–90.

Fulkerson, D. R., & Dantzig, G. B. (1955). Computation of maximal flows in networks. *NAV Naval Research Logistics Quarterly*, *2*(4), 277–283.
Furgeson, J., McCullough, M., Wolfendale, C., & Gill, B. (2014). *The Equity Project Charter School: Impacts on Student Achievement* (06635). Mathematica Policy Research.
Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge; New York: Cambridge University Press.
Gilbert, J. P. (1974). Randomization of human subjects. *New England Journal of Medicine*, *291*(24), 1305–1306.
Greenwood, E. (1945). *Experimental sociology : a study in method*. New York: King's Crown Press.
Gu, X., & Rosenbaum, P. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420.
Gu, X. S., & Rosenbaum, P. R. (1993a). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420.
Gu, X. S., & Rosenbaum, P. R. (1993b). Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2(4)*, 405–420.
Guo, S., & Fraser, M. (2010). *Propensity score analysis : Statistical methods and applications*. Thousand Oaks, Calif.: Sage Publications.
Hade, E. M., & Lu, B. (2014). Bias associated with using the estimated propensity score as a regression covariate. *Statistics in Medicine*, *33*(1), 74–87.
Hahs-Vaughn, D. L., & Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *Journal of Experimental Education*, *75*(1).
Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*(5), 1017–1098.
Herzog, S. (2008). Estimating the influence offinancial aid on student retention. *Fayetteville, AR: Education Working Paper Archive.*
Hill, J. (2008). Comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003"by Peter Austin, Statistics. *Statistics in Medicine*, 2055–2061.
Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.
Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology, 2(3-4)*, 259–278.
Hong, G., Corter, C., Hong, Y., & Pelletier, J. (2012). Differential effects of literacy instruction time and homogeneous ability grouping in kindergarten classrooms: who will benefit? who will suffer? *Educational Evaluation and Policy Analysis*, *34*(1), 69–88.
Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*(3), 205–224.
Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, *101*(475), 901–910.

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: an application of propensity score method to multivariate, multilevel data. *Developmental Psychology*, *44*(2), 407–421.

Horvitz, D. J., & Thompson, D. G. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association Journal of the American Statistical Association*, *47*(260), 663–685.

Howard, B., Raudenbush, S., & Weiss, M. (2014). *Using Multi-site Evaluations to Study Variation in Effects of Program Assignment*. MDRC.

Imai, K., King, G., & Stuart, E. a. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481–502.

Kelcey, B. (2011). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. *Educational Evaluation and Policy Analysis*, *33*(4), 458–482.

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)*.

King, G., Nielsen, R., & Coberley, C. (2011). Comparative effectiveness of matching methods for causal inference. *Unpublished manuscript.*

Lee, B. K., Lessler, J., & Stuart, E. A. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29, 337-346.

Long, M. C., Conger, D., & Iatarola, P. (2012). Effects of high school course-taking on secondary and postsecondary success. *American Educational Research Journal*, *49*(2), 285–322.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86.

Melguizo, T., & Wolniak, G. C. (2012). The earnings benefits of majoring in STEM fields among high achieving minority students. *Research in Higher Education*, *53*(4), 383–405.

Ming, K., & Rosenbaum, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, *10*(3), 455–463.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*(1), 34.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference : methods and principles for social research*. New York: Cambridge University Press.

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology*, *2(74)*.

Niehaus, E., Campbell, C. M., & Inkelas, K. K. (2014). HLM behind the curtain: unveiling decisions behind the use and interpretation of HLM in higher education research. *Research in Higher Education*, *55*(1), 101–122.

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: new simulation results. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *7*(3), 111–120.

Pearl, J. (2009). *Causality : models, reasoning, and inference*. Cambridge; New York: Cambridge University Press.

Pedhazur, E. J., & Kerlinger, F. N. (1982). *Multiple regression in behavioral research: explanation and prediction*. Holt, Rinehart, and Winston.
Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance–correlation parameters. *Biometrika*, *94*(4), 1006–1013.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models : Applications and data analysis methods*. Thousand Oaks: Sage Publications.
Reynolds, C. L., & DesJardins, S. (2009). The use of matching methods in higher education research: answering whether attendance at a 2-year institution results in differences in educational attainment. In J. Smart (Ed.), *Higher Education: Handbook of Theory and Research SE - 2* (24), 47–97. Springer Netherlands.
Rosenbaum, P. R. (1986). Dropping out of high school in the United States: an observational study. *Journal of Educational and Behavioral Statistics*, *11*(3), 207–224.
Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.
Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524.
Rosenbaum, P., & Rubin, D. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.
Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524.
Rosenbaum, P., & Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.
Rubin, D. (1976). Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics*, *32*(1), 109–120.
Rubin, D. (1980). Bias reduction using Mahalanobis-metric matching. *Biometrics*, *36*(2), 293–298.
Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, *29*(1), 159–183.
Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, *29*(1), 185–203.
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. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, *6*(1), 34–58.
Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*(366), 318.
Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 249–264.
Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, *52*(1), 249–64.
Rubin, D., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical ...*, *95*(450), 573–585.

Ryan, K. E., & Cousins, J. B. (2009). *The SAGE international handbook of educational evaluation*. Los Angeles: SAGE Publications.
Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating Causal Effects Using Experimental and Observational Designs (report from the Governing Board of the American Educational Research Association Grants Program)*. Washington, DC: American Educational Research Association.
Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.
Sekhon, J. (2007). Alternative balance metrics for bias reduction in matching methods for causal inference. *Survey Research Center, University of California.*
Sekhon, J. (2008). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, *42*(7).
Sekhon, J. S. (2009). Opiates for the matches: matching methods for causal inference. *Annual Review of Political Science*, *12*(1), 487–508.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, *58*(6), 550–559.
Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis : an introduction to basic and advanced multilevel modeling*. London; Thousand Oaks, Calif.: Sage Publications.
Splawa-Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472.
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), 250–67.
Stuart, E. a. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21.
Stuart, E. a, & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, *44*(2), 395–406.
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.
Vallejo, G., Fernández, P., Cuesta, M., & Livacic-Rojas, P. E. (2015). Effects of modeling the heterogeneity on inferences drawn from multilevel designs. *Multivariate Behavioral Research*, *50*(1), 75–90.
Vaughan, A. L., Lalonde, T. L., & Jenkins-Guarnieri, M. A. (2014). Assessing student achievement in large-scale educational programs using hierarchical propensity scores. *Research in Higher Education*, 1–17.
Weinstein, M. C. (1974). Allocation of subjects in medical experiments. *New England Journal of Medicine*, *291*(24), 1278–1285.
Weiss, M. J., Bloom, H. S., & Brock, T. (2013). *A conceptual framework for studying the sources of variation in program effects*. MDRC.
Zhao, Z. (2004). Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics*, *86*(2), 91–107.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters*, *98*(3), 309–319.
Zhou, B., Konstorum, A., Duong, T., Tieu, K. H., Wells, W. M., Brown, G. G., Stern, H. S., Shahbaba, B. (2013). A hierarchical modeling approach to data analysis and study design in a multi-site experimental fmri study. *Psychometrika*, *78*(2), 260–278.

Appendix A: Additional Simulation Results

Table 12 Monte Carlo Results for various Average Treatment Effect Estimation Methods and different correlation levels between variables (n=600)

Treatment exposure = .25

		ES=0						ES=.1						ES=.3						ES=.5					
	Method	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage
	Naïve	-0.006	0.025	-79.683	0.097	0.010	0.9413	0.101	0.030	42.089	0.096	0.010	0.9377	0.313	0.035	12.540	0.096	0.010	0.9347	0.517	0.033	6.869	0.095	0.010	0.9361
	OLS	-0.030	0.002	-4.855	0.095	0.009	0.9484	0.073	0.002	2.134	0.096	0.009	0.9478	0.279	0.001	0.535	0.096	0.009	0.9478	0.486	0.001	0.297	0.098	0.010	0.9477
r=.1	PSM	-0.030	0.002	-5.145	0.112	0.013	0.9506	0.073	0.001	1.995	0.113	0.013	0.9479	0.278	0.000	0.144	0.115	0.013	0.9475	0.485	0.000	0.094	0.122	0.015	0.9496
	IPW	-0.030	0.002	-5.064	0.096	0.009	0.9503	0.073	0.002	2.205	0.096	0.009	0.9502	0.279	0.001	0.512	0.098	0.010	0.9482	0.485	0.001	0.214	0.103	0.011	0.947
	SubPSM	-0.028	0.004	-12.725	0.096	0.009	0.9493	0.076	0.004	6.284	0.096	0.009	0.9487	0.283	0.005	1.765	0.097	0.010	0.9429	0.489	0.005	0.962	0.100	0.010	0.9346
	Naïve	-0.057	0.177	-75.550	0.107	0.043	0.6134	0.066	0.177	-160.031	0.104	0.042	0.5993	0.301	0.165	121.036	0.100	0.037	0.6211	0.521	0.138	35.977	0.098	0.029	0.7013
	OLS	-0.232	0.001	-0.639	0.098	0.010	0.9473	-0.109	0.001	-1.343	0.097	0.009	0.9475	0.138	0.001	1.067	0.097	0.009	0.9479	0.385	0.001	0.368	0.099	0.010	0.9474
r=.3	PSM	-0.231	0.003	-1.205	0.123	0.015	0.9581	-0.107	0.003	-2.711	0.120	0.015	0.9526	0.141	0.005	3.397	0.119	0.014	0.9497	0.387	0.004	1.004	0.125	0.016	0.9496
	IPW	-0.232	0.002	-0.822	0.103	0.011	0.9696	-0.109	0.002	-1.699	0.101	0.010	0.9667	0.138	0.002	1.255	0.102	0.010	0.9605	0.384	0.001	0.356	0.107	0.011	0.953
	SubPSM	-0.214	0.020	-8.558	0.100	0.010	0.9587	-0.090	0.020	-18.141	0.100	0.010	0.9547	0.155	0.018	13.558	0.099	0.010	0.9475	0.399	0.016	4.125	0.102	0.011	0.9328
	Naïve	-0.166	0.501	-75.106	0.130	0.268	0.0285	-0.019	0.481	-96.177	0.123	0.246	0.0244	0.254	0.421	-252.690	0.111	0.190	0.0318	0.501	0.335	200.803	0.103	0.123	0.0993
	OLS	-0.665	0.001	-0.202	0.108	0.012	0.9468	-0.499	0.001	-0.278	0.104	0.011	0.9465	-0.165	0.001	-0.843	0.101	0.010	0.948	0.168	0.001	0.828	0.101	0.010	0.9483
r=.5	PSM	-0.649	0.018	-2.646	0.202	0.041	0.9739	-0.489	0.011	-2.186	0.162	0.026	0.9678	-0.158	0.009	-5.244	0.134	0.018	0.9522	0.173	0.007	4.085	0.133	0.018	0.9539
	IPW	-0.658	0.008	-1.229	0.188	0.035	0.9642	-0.495	0.005	-0.985	0.147	0.022	0.9718	-0.164	0.003	-1.897	0.122	0.015	0.971	0.169	0.003	1.506	0.119	0.014	0.9594
	SubPSM	-0.608	0.058	-8.755	0.141	0.023	0.903	-0.445	0.055	-10.955	0.119	0.017	0.9291	-0.120	0.047	-28.038	0.107	0.014	0.9354	0.204	0.038	22.549	0.106	0.013	0.9209

Treatment exposure = .50

		ES=0						ES=.1						ES=.3						ES=.5					
	Method	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage
	Naïve	-0.006	0.026	-81.331	0.083	0.008	0.9345	0.102	0.031	42.795	0.083	0.008	0.9303	0.313	0.035	12.709	0.082	0.008	0.9242	0.518	0.034	6.958	0.082	0.008	0.9265
	OLS	-0.030	0.002	-5.678	0.082	0.007	0.9477	0.073	0.002	2.476	0.082	0.007	0.9472	0.279	0.002	0.611	0.083	0.007	0.9477	0.486	0.002	0.333	0.086	0.007	0.9476
r=.1	PSM	-0.030	0.002	-5.961	0.095	0.009	0.9545	0.074	0.003	3.703	0.095	0.009	0.9517	0.280	0.002	0.721	0.097	0.009	0.9463	0.486	0.002	0.404	0.101	0.010	0.9494
	IPW	-0.030	0.002	-5.617	0.082	0.007	0.9518	0.073	0.002	2.450	0.082	0.007	0.9513	0.279	0.002	0.600	0.083	0.007	0.9498	0.486	0.002	0.311	0.086	0.007	0.9481
	SubPSM	-0.028	0.004	-12.377	0.083	0.007	0.9486	0.076	0.005	6.454	0.083	0.007	0.9473	0.283	0.005	1.829	0.085	0.007	0.9403	0.489	0.005	1.044	0.091	0.008	0.921
	Naïve	-0.056	0.178	-75.948	0.092	0.040	0.5108	0.067	0.178	-160.811	0.090	0.040	0.4927	0.313	0.035	12.709	0.082	0.008	0.5175	0.522	0.138	36.133	0.084	0.026	0.6213
	OLS	-0.232	0.002	-0.784	0.085	0.007	0.9502	-0.109	0.002	-1.626	0.084	0.007	0.949	0.279	0.002	0.611	0.083	0.007	0.9476	0.385	0.002	0.425	0.087	0.008	0.9469
r=.3	PSM	-0.231	0.003	-1.355	0.100	0.010	0.9609	-0.106	0.004	-3.554	0.099	0.010	0.9572	0.280	0.002	0.721	0.097	0.009	0.9493	0.386	0.003	0.867	0.102	0.010	0.952
	IPW	-0.232	0.002	-0.767	0.085	0.007	0.9738	-0.109	0.002	-1.592	0.085	0.007	0.9699	0.279	0.002	0.600	0.083	0.007	0.9621	0.385	0.002	0.407	0.088	0.008	0.9537
	SubPSM	-0.214	0.020	-8.431	0.089	0.008	0.9515	-0.091	0.020	-17.976	0.088	0.008	0.9476	0.283	0.005	1.829	0.085	0.007	0.9382	0.399	0.016	4.237	0.093	0.009	0.9168
	Naïve	-0.164	0.502	-75.332	0.112	0.265	0.0053	-0.018	0.482	-96.451	0.106	0.244	0.0046	0.256	0.422	-253.350	0.096	0.187	0.0077	0.502	0.336	201.306	0.088	0.120	0.0323
	OLS	-0.665	0.002	-0.281	0.097	0.009	0.9491	-0.498	0.002	-0.367	0.092	0.008	0.9491	-0.165	0.002	-1.049	0.089	0.008	0.9471	0.168	0.002	0.991	0.089	0.008	0.947
r=.5	PSM	-0.654	0.013	-1.901	0.149	0.022	0.9752	-0.491	0.009	-1.876	0.124	0.015	0.9692	-0.160	0.007	-4.053	0.108	0.012	0.9575	0.172	0.006	3.309	0.107	0.011	0.948
	IPW	-0.662	0.005	-0.720	0.132	0.018	0.985	-0.497	0.003	-0.620	0.107	0.011	0.9875	-0.164	0.002	-1.353	0.094	0.009	0.9798	0.169	0.002	1.191	0.094	0.009	0.9683
	SubPSM	-0.608	0.059	-8.787	0.156	0.028	0.8171	-0.447	0.053	-10.696	0.119	0.017	0.8866	-0.121	0.046	-27.655	0.100	0.012	0.9084	0.204	0.038	22.529	0.100	0.011	0.8931

Note: N=10,000 replicates per cell. Treatment exposure probabilities indicate the proportion of the sample that has been exposed to the treatment. ES=the effect size in terms of Cohen's d and the treatment effect applied to the sample.
Naïve=Regression of outcome y on treatment indicator; OLS=Least Squares regression; PSM=Propensity Score Matching estimation; IPW= Inverse Probability Weights using the propensity score; SubPSM=Sub classification using the propensity score.

Table 13 Monte Carlo Results for various Average Treatment Effect Estimation Methods and different correlation levels between variables (n=1000)

Treatment exposure = .25

		ES=0						ES=.1						ES=.3						ES=.5					
	Method	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage
	Naïve	-0.009	0.023	-71.389	0.074	0.006	0.95	0.099	0.027	38.375	0.073	0.006	0.9352	0.310	0.032	11.571	0.073	0.006	0.9291	0.515	0.031	6.305	0.073	0.006	0.9314
	OLS	-0.033	-0.001	4.137	0.073	0.005	0.9503	0.070	-0.001	-1.832	0.073	0.005	0.9507	0.276	-0.001	-0.468	0.073	0.005	0.9514	0.483	-0.001	-0.267	0.075	0.006	0.9515
r=.1	PSM	-0.033	-0.001	3.087	0.086	0.007	0.9529	0.071	-0.001	-0.953	0.087	0.008	0.9485	0.277	-0.001	-0.357	0.088	0.008	0.9527	0.483	-0.001	-0.132	0.093	0.009	0.9516
	IPW	-0.033	-0.001	3.958	0.073	0.005	0.9536	0.070	-0.001	-1.791	0.073	0.005	0.9527	0.276	-0.001	-0.481	0.074	0.006	0.9517	0.483	-0.001	-0.291	0.079	0.006	0.9523
	SubPSM	-0.031	0.001	-3.662	0.073	0.005	0.9535	0.073	0.002	2.336	0.073	0.005	0.9512	0.280	0.002	0.756	0.074	0.005	0.9468	0.486	0.002	0.372	0.076	0.006	0.9407
	Naïve	-0.059	0.174	-74.574	0.082	0.037	0.4317	0.064	0.174	-157.910	0.080	0.037	0.4135	0.299	0.163	119.230	0.077	0.032	0.4392	0.518	0.135	35.305	0.074	0.024	0.5621
	OLS	-0.235	-0.001	0.588	0.074	0.005	0.9513	-0.112	-0.001	1.234	0.074	0.005	0.9519	0.135	-0.001	-0.986	0.074	0.005	0.9531	0.382	-0.001	-0.347	0.075	0.006	0.9516
r=.3	PSM	-0.234	0.000	-0.072	0.092	0.008	0.9605	-0.110	0.001	-0.770	0.090	0.008	0.9575	0.137	0.000	0.318	0.090	0.008	0.9523	0.383	0.000	-0.115	0.096	0.009	0.9535
	IPW	-0.235	-0.001	0.365	0.079	0.006	0.9716	-0.111	-0.001	0.845	0.077	0.006	0.9703	0.135	-0.001	-0.777	0.077	0.006	0.9627	0.382	-0.001	-0.297	0.081	0.007	0.9589
	SubPSM	-0.217	0.017	-7.344	0.076	0.006	0.958	-0.093	0.017	-15.454	0.075	0.006	0.9548	0.152	0.016	11.634	0.075	0.006	0.9483	0.396	0.013	3.381	0.077	0.006	0.9388
	Naïve	-0.168	0.499	-74.826	0.100	0.259	0.0011	-0.021	0.479	-95.784	0.094	0.238	8.00E-04	0.252	0.419	-251.387	0.085	0.183	0.0014	0.499	0.332	199.373	0.078	0.117	0.0121
	OLS	-0.668	-0.001	0.224	0.082	0.007	0.9512	-0.501	-0.001	0.291	0.078	0.006	0.9513	-0.168	-0.001	0.845	0.076	0.006	0.9522	0.165	-0.001	-0.826	0.077	0.006	0.9523
r=.5	PSM	-0.658	0.009	-1.356	0.161	0.026	0.9704	-0.494	0.006	-1.270	0.125	0.016	0.9687	-0.163	0.003	-2.084	0.103	0.011	0.9587	0.169	0.002	1.275	0.102	0.010	0.9536
	IPW	-0.664	0.003	-0.462	0.147	0.022	0.9686	-0.499	0.001	-0.224	0.114	0.013	0.9756	-0.167	0.000	-0.010	0.093	0.009	0.9738	0.166	0.000	-0.174	0.091	0.008	0.9657
	SubPSM	-0.612	0.055	-8.258	0.105	0.014	0.9	-0.449	0.051	-10.167	0.089	0.010	0.929	-0.123	0.044	-26.291	0.081	0.008	0.9326	0.201	0.034	20.692	0.081	0.008	0.9195

Treatment exposure = .50

		ES=0						ES=.1						ES=.3						ES=.5					
	Method	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage
	Naïve	-0.009	0.023	-72.370	0.064	0.005	0.9354	0.099	0.028	38.819	0.064	0.005	0.9295	0.310	0.032	11.690	0.064	0.005	0.9194	0.515	0.031	6.376	0.063	0.005	0.922
	OLS	-0.033	-0.001	2.775	0.063	0.004	0.9518	0.071	-0.001	-1.216	0.064	0.004	0.952	0.277	-0.001	-0.303	0.064	0.004	0.952	0.483	-0.001	-0.166	0.066	0.004	0.951
r=.1	PSM	-0.033	-0.001	4.058	0.073	0.005	0.9515	0.071	-0.001	-1.046	0.073	0.005	0.9509	0.277	0.000	-0.111	0.073	0.005	0.9556	0.483	-0.001	-0.130	0.077	0.006	0.9492
	IPW	-0.033	-0.001	2.816	0.063	0.004	0.9548	0.071	-0.001	-1.243	0.064	0.004	0.9545	0.277	-0.001	-0.316	0.064	0.004	0.9537	0.483	-0.001	-0.180	0.067	0.004	0.9521
	SubPSM	-0.030	0.002	-5.050	0.064	0.004	0.9514	0.074	0.002	3.181	0.064	0.004	0.9513	0.280	0.003	0.968	0.066	0.004	0.945	0.487	0.003	0.548	0.070	0.005	0.9282
	Naïve	-0.059	0.175	-74.689	0.071	0.036	0.3085	0.064	0.175	-158.157	0.069	0.035	0.2916	0.299	0.163	119.440	0.067	0.031	0.3165	0.519	0.136	35.384	0.065	0.023	0.4468
	OLS	-0.235	-0.001	0.356	0.065	0.004	0.9517	-0.111	-0.001	0.746	0.065	0.004	0.9528	0.136	-0.001	-0.589	0.065	0.004	0.9532	0.382	-0.001	-0.202	0.066	0.004	0.9514
r=.3	PSM	-0.233	0.000	-0.165	0.077	0.006	0.9613	-0.110	0.000	-0.263	0.076	0.006	0.9555	0.136	-0.001	-0.370	0.076	0.006	0.9488	0.383	0.000	0.043	0.078	0.006	0.9503
	IPW	-0.235	-0.001	0.399	0.066	0.004	0.9742	-0.111	-0.001	0.832	0.065	0.004	0.9697	0.135	-0.001	-0.655	0.065	0.004	0.9643	0.382	-0.001	-0.225	0.068	0.005	0.9581
	SubPSM	-0.216	0.017	-7.402	0.069	0.005	0.9512	-0.093	0.017	-15.841	0.068	0.005	0.9494	0.153	0.016	12.044	0.068	0.005	0.9407	0.397	0.014	3.651	0.071	0.005	0.9193
	Naïve	-0.167	0.499	-74.878	0.087	0.257	0	-0.021	0.479	-95.847	0.082	0.236	0	0.253	0.419	-251.563	0.074	0.181	0	0.499	0.333	199.551	0.068	0.115	0.0021
	OLS	-0.667	-0.001	0.116	0.074	0.005	0.9515	-0.501	-0.001	0.153	0.070	0.005	0.9527	-0.167	-0.001	0.451	0.068	0.005	0.9521	0.166	-0.001	-0.437	0.068	0.005	0.9534
r=.5	PSM	-0.661	0.005	-0.824	0.114	0.013	0.9752	-0.496	0.004	-0.777	0.095	0.009	0.9691	-0.164	0.003	-1.509	0.083	0.007	0.955	0.169	0.002	1.199	0.082	0.007	0.9503
	IPW	-0.667	0.000	-0.019	0.103	0.011	0.9867	-0.501	-0.001	0.132	0.082	0.007	0.9893	-0.168	-0.001	0.514	0.072	0.005	0.9835	0.166	-0.001	-0.475	0.072	0.005	0.9712
	SubPSM	-0.613	0.054	-8.070	0.116	0.016	0.8183	-0.448	0.052	-10.340	0.090	0.011	0.8754	-0.122	0.044	-26.649	0.077	0.008	0.8975	0.202	0.035	21.114	0.077	0.007	0.8841

Note: N=10,000 replicates per cell. Treatment exposure probabilities indicate the proportion of the sample that has been exposed to the treatment. ES=the effect size in terms of Cohen's d and the treatment effect applied to the sample.
Naïve=Regression of outcome y on treatment indicator; OLS=Least Squares regression; PSM=Propensity Score Matching estimation; IPW= Inverse Probability Weights using the propensity score; SubPSM=Sub classification using the propensity score.

Table 14 Monte Carlo Results for various Average Treatment Effect Estimation Methods and different correlation levels between variables (n=2000)

Treatment exposure = .25

		ES=0						ES=.1						ES=.3						ES=.5					
	Method	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage
	Naïve	-0.008	0.023	-73.598	0.053	0.003	0.9285	0.100	0.028	39.375	0.052	0.004	0.903	0.311	0.033	11.839	0.052	0.004	0.903	0.515	0.031	6.464	0.052	0.004	0.9075
	OLS	-0.032	0.000	1.384	0.052	0.003	0.9495	0.071	0.000	-0.615	0.052	0.003	0.9485	0.277	0.000	-0.158	0.052	0.003	0.9485	0.484	0.000	-0.091	0.053	0.003	0.9474
r=.1	PSM	-0.032	-0.001	2.138	0.060	0.004	0.9512	0.071	-0.001	-0.804	0.060	0.004	0.9472	0.277	-0.001	-0.318	0.062	0.004	0.9472	0.484	0.000	0.036	0.066	0.004	0.9511
	IPW	-0.032	0.000	1.233	0.052	0.003	0.952	0.071	0.000	-0.585	0.052	0.003	0.9511	0.277	0.000	-0.172	0.053	0.003	0.9511	0.484	-0.001	-0.114	0.056	0.003	0.9509
	SubPSM	-0.030	0.002	-6.455	0.052	0.003	0.9517	0.074	0.003	3.595	0.052	0.003	0.9474	0.281	0.003	1.074	0.053	0.003	0.9474	0.487	0.003	0.592	0.054	0.003	0.9396
	Naïve	-0.059	0.175	-74.827	0.058	0.034	0.1435	0.065	0.175	-158.452	0.057	0.034	0.1311	0.300	0.163	119.691	0.055	0.030	0.1521	0.519	0.136	35.481	0.053	0.021	0.274
	OLS	-0.234	0.000	0.193	0.053	0.003	0.9512	-0.111	0.000	0.407	0.053	0.003	0.9513	0.136	0.000	-0.328	0.053	0.003	0.9494	0.383	0.000	-0.116	0.054	0.003	0.9488
r=.3	PSM	-0.234	0.000	0.008	0.066	0.004	0.9555	-0.110	0.000	-0.412	0.064	0.004	0.9537	0.136	0.000	0.085	0.063	0.004	0.9508	0.383	0.000	-0.021	0.068	0.005	0.9529
	IPW	-0.234	0.000	0.129	0.056	0.003	0.9713	-0.111	0.000	0.304	0.055	0.003	0.9671	0.136	0.000	-0.291	0.055	0.003	0.9605	0.383	0.000	-0.117	0.058	0.003	0.9544
	SubPSM	-0.216	0.018	-7.655	0.054	0.003	0.955	-0.093	0.018	-16.201	0.053	0.003	0.9518	0.153	0.017	12.305	0.053	0.003	0.9467	0.397	0.014	3.644	0.055	0.003	0.9358
	Naïve	0.500	0.333	199.759	0.056	0.114	0	-0.020	0.480	-95.926	0.067	0.235	0	0.253	0.420	-251.771	0.061	0.180	0	0.500	0.333	199.759	0.056	0.114	0
	OLS	0.166	0.000	-0.274	0.055	0.003	0.9532	-0.500	0.000	0.093	0.056	0.003	0.9523	-0.167	0.000	0.275	0.055	0.003	0.9513	0.166	0.000	-0.274	0.055	0.003	0.9514
r=.5	PSM	0.168	0.002	1.095	0.073	0.005	0.969	-0.496	0.004	-0.830	0.090	0.008	0.9637	-0.164	0.002	-1.424	0.074	0.005	0.9517	0.168	0.002	1.095	0.073	0.005	0.9505
	IPW	0.167	0.000	-0.093	0.065	0.004	0.9716	-0.500	0.000	-0.078	0.081	0.007	0.976	-0.167	0.000	0.047	0.066	0.004	0.9742	0.167	0.000	-0.093	0.065	0.004	0.9619
	SubPSM	0.202	0.035	21.202	0.057	0.004	0.8636	-0.448	0.052	-10.340	0.062	0.007	0.8958	-0.122	0.045	-26.787	0.057	0.005	0.9006	0.202	0.035	21.202	0.057	0.004	0.9007

Treatment exposure = .50

		ES=0						ES=.1						ES=.3						ES=.5					
	Method	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage	Mean	Raw Bias	Relative Bias	sd	MSE	Coverage
	Naïve	-0.009	0.023	-73.056	0.045	0.003	0.9184	0.099	0.028	39.149	0.045	0.003	0.9042	0.311	0.033	11.788	0.045	0.003	0.8865	0.515	0.031	6.439	0.045	0.003	0.8928
	OLS	-0.032	0.000	1.479	0.045	0.002	0.9533	0.071	0.000	-0.694	0.045	0.002	0.9526	0.277	-0.001	-0.198	0.045	0.002	0.9525	0.484	-0.001	-0.126	0.047	0.002	0.9526
r=.1	PSM	-0.032	-0.001	1.671	0.050	0.003	0.9526	0.071	0.000	-0.224	0.051	0.003	0.9517	0.277	-0.001	-0.247	0.051	0.003	0.9507	0.483	-0.001	-0.224	0.053	0.003	0.9524
	IPW	-0.032	0.000	1.461	0.045	0.002	0.956	0.071	0.000	-0.683	0.045	0.002	0.9557	0.277	-0.001	-0.194	0.045	0.002	0.9541	0.484	-0.001	-0.125	0.047	0.002	0.9533
	SubPSM	-0.030	0.002	-6.209	0.045	0.002	0.9544	0.074	0.002	3.426	0.045	0.002	0.9526	0.281	0.003	1.060	0.046	0.002	0.9457	0.487	0.003	0.559	0.049	0.002	0.9288
	Naïve	-0.059	0.175	-74.688	0.050	0.033	0.0659	0.064	0.175	-158.174	0.049	0.033	0.0545	0.299	0.163	119.498	0.047	0.029	0.0688	0.519	0.136	35.425	0.046	0.021	0.1612
	OLS	-0.234	0.000	0.163	0.046	0.002	0.9512	-0.111	0.000	0.382	0.046	0.002	0.9532	0.136	0.000	-0.363	0.046	0.002	0.9515	0.383	-0.001	-0.148	0.047	0.002	0.9531
r=.3	PSM	-0.233	0.000	-0.114	0.053	0.003	0.9641	-0.110	0.000	0.082	0.052	0.003	0.9555	0.136	0.000	-0.215	0.052	0.003	0.9506	0.383	0.000	0.053	0.055	0.003	0.9496
	IPW	-0.234	0.000	0.138	0.047	0.002	0.9746	-0.111	0.000	0.331	0.046	0.002	0.9714	0.136	0.000	-0.324	0.046	0.002	0.9648	0.383	-0.001	-0.135	0.048	0.002	0.9597
	SubPSM	-0.216	0.018	-7.635	0.048	0.003	0.946	-0.093	0.018	-16.071	0.048	0.003	0.9434	0.153	0.017	12.187	0.048	0.003	0.9354	0.397	0.014	3.617	0.050	0.003	0.9179
	Naïve	-0.167	0.499	-74.897	0.061	0.253	0	-0.021	0.479	-95.856	0.058	0.233	0	0.253	0.419	-251.562	0.053	0.179	0	0.499	0.333	199.581	0.048	0.113	0
	OLS	-0.667	0.000	0.014	0.053	0.003	0.9507	-0.500	0.000	0.048	0.050	0.003	0.9514	-0.167	0.000	0.243	0.048	0.002	0.9518	0.166	-0.001	-0.310	0.048	0.002	0.9523
r=.5	PSM	-0.662	0.004	-0.649	0.081	0.007	0.9739	-0.497	0.003	-0.599	0.066	0.004	0.9731	-0.166	0.001	-0.678	0.057	0.003	0.959	0.168	0.001	0.523	0.057	0.003	0.9503
	IPW	-0.666	0.001	-0.153	0.072	0.005	0.9869	-0.500	0.000	-0.071	0.058	0.003	0.9894	-0.167	0.000	0.055	0.051	0.003	0.9829	0.166	0.000	-0.176	0.051	0.003	0.972
	SubPSM	-0.611	0.056	-8.346	0.079	0.009	0.7761	-0.448	0.052	-10.316	0.063	0.007	0.833	-0.122	0.044	-26.580	0.054	0.005	0.8574	0.202	0.035	21.047	0.054	0.004	0.8577

Note: N=10,000 replicates per cell. Treatment exposure probabilities indicate the proportion of the sample that has been exposed to the treatment. ES=the effect size in terms of Cohen's d and the treatment effect applied to the sample.
Naïve=Regression of outcome y on treatment indicator; OLS=Least Squares regression; PSM=Propensity Score Matching estimation; IPW= Inverse Probability Weights using the propensity score; SubPSM=Sub classification using the propensity score.

Appendix B: R Sample Simulation Code

```
setwd("~/R/Sim_1000_25/r_mod_d3")
#install.packages('Matching')
#install.packages('survey')
#install.packages('MatchIt')
#install.packages('optmatch')

#library(Matching)
#library(survey)
#library(MatchIt)
#library(optmatch)

rm(list=ls())

source("~/R/sim_functions")

scenario.1 <- r.matrix(.3,.3,.3,.3,.3,.3,.3,.3,.3,.3)
##########
b0t <- 0.136363636
b0c <- 0
g <-  0.136363636
b1 <- 0.136363636
b2 <- 0.136363636
b3 <- 0.136363636
b4 <- 0.136363636

tau_list = c(0,2,3,4,5)
sigma.e=1

nT<-250
nC<-750
xC_bar<-0
xT_bar<-0
xC_sd<-1
xT_sd<-1

#sigma.y=sigma.g+sigma.e
nTreatClust=20
nControlClust=20
nChildPerClust=25

nClust <- nTreatClust + nControlClust
nObs <- nChildPerClust * nClust
```

```
S= 10000#number of replications

pb <- txtProgressBar(min = 0, max = S, style = 3)
#begin loop for tau_ICC
for (tau in tau_list){

  ICC<-matrix(NA, nrow=S,ncol=5)
  colnames(ICC)<-c("sigma2y","sigma2alpha","icc","Fvalue","pvalue")

  Mtch<-matrix(NA, nrow=S, ncol=12)
  colnames(Mtch)<-c("ate","sd.ate","ols","sd.ols","psm.ate","sd.psm","psm.att","sd.att","ipw",
                    "sd.ipw","sub.ate","sd.sub")

  #begin simulation of data sets

  for(s in 1:S) {
    set.seed(s)
    #generate variables
    x1_t <- rnorm(nT, mean=xT_bar, sd=xT_sd)
    x2_t <- rnorm(nT, mean=xT_bar, sd=xT_sd)
    x3_t <- rnorm(nT, mean=xT_bar, sd=xT_sd)
    x4_t <- rnorm(nT, mean=xT_bar, sd=xT_sd)

    x1_c <- rnorm(nC, mean=xC_bar, sd=xC_sd)
    x2_c <- rnorm(nC, mean=xC_bar, sd=xC_sd)
    x3_c <- rnorm(nC, mean=xC_bar, sd=xC_sd)
    x4_c <- rnorm(nC, mean=xC_bar, sd=xC_sd)

    t <-c(rep(1, nT), rep(0, nC))  # make trt indicator #

    x1 <-(c(x1_t,x1_c))
    x2 <-(c(x2_t,x2_c))
    x3 <-(c(x3_t,x3_c))
    x4 <-(c(x4_t,x4_c))
    #assing desired correlations
    d <- matrix(c(t,x1,x2,x3,x4), nrow=nObs, ncol=5)
    colnames(d)<-c("tx","x1","x2","x3","x4")
    #sim.d[,1] <- ifelse(sim.d[,1] > mean(sim.d[,1]), 1, 0)
    #x[,1] <- cut.v(x[,1], c(0, 1-p$treat, 1)) - 1
    x.cor <- d %*% chol(scenario.1)
    m <- matrix(NA,nrow=nObs,ncol=2)
    colnames(m) <- c("cluster","child")

    m[,1] <- rep(1:nClust,each=nChildPerClust)
    m[,2] <- 1:nrow(m)
    m<-data.frame(cbind(m,x.cor))
```

```
   names(m) <- c("cluster","child",
                 'tx','x1','x2','x3','x4')

   y_t <- b0t + b1*m$x1[m$tx==1] + b2*m$x2[m$tx==1] +b3*m$x3[m$tx==1] +
b4*m$x4[m$tx==1] + rnorm(nT,0,sigma.e)
   y_c <- b0c + b1*m$x1[m$tx==0] + b2*m$x2[m$tx==0] +b3*m$x3[m$tx==0] +
b4*m$x4[m$tx==0] + rnorm(nC,0,sigma.e)

   y <- (c(y_t,y_c))

   m$y <- y + rep(rnorm(nClust,0,tau),each=nChildPerClust)

   ################
   reg <- summary(lm(m$y~m$tx+m$x1+m$x2+m$x3+m$x4))
   Mtch[s,3] <- reg$coefficients[[2]]
   Mtch[s,4] <- reg$coefficients[[8]]

   I <- getICC(m$y,m$cluster)
   ICC[s,] <- c(I[[1]],I[[2]],I[[3]],I[[4]],I[[5]])

   ate.r <- summary(lm(m$y~m$tx))
   Mtch[s,1] <- ate.r$coefficients[[2]]
   Mtch[s,2] <- ate.r$coefficients[[4]]
   ################PSM model
   Y <- m$y
   Tr <- m$tx
   glm1 <- glm(tx ~ x1 + x2 + x3 + x4, family = binomial, data = m)

   rr1 <- Match(Y = Y, Tr = Tr, X = glm1$fitted, estimand="ATE")
   rr2 <- Match(Y = Y, Tr = Tr, X = glm1$fitted, estimand="ATT")

   Mtch[s,5] <- rr1$est
   Mtch[s,6] <- rr1$se
   Mtch[s,7] <- rr2$est
   Mtch[s,8] <- rr2$se

   #IPW model
   m$pihat.log <- glm1$fitted
   #Calculate Weights

   m$ipw.ate <- ifelse(m$tx==1, 1/m$pihat.log,1/(1-m$pihat.log))
   #ATE Outcome Analysis

   design.ate <- svydesign(ids= ~1, weights= ~ipw.ate, dat=m)
   mod.ipw.ate <- summary(svyglm(y ~ tx, design=design.ate))
```

```
  Mtch[s,9] <- mod.ipw.ate$coefficients[[2]]#slope
  Mtch[s,10] <- mod.ipw.ate$coefficients[[4]]#sd
  ##############################
  subcl <- matchit(tx ~ x1 + x2 + x3 +x4, data=m,
                   method='subclass',   subclass=5)
  #ATE Outcome Analysis
  sub.mtch<- summary(lm(y~ tx,weights=weights,data=match.data(subcl)))
  Mtch[s,11] <- sub.mtch$coefficients[[2]]#slope
  Mtch[s,12] <- sub.mtch$coefficients[[4]]#sd
  setTxtProgressBar(pb, s)

 }

 icc<- mean(ICC[,3])
 ##################################naive
 naive.ate <- sum(Mtch[,1]/S)
 raw.bias.ate <- sum(Mtch[,1]/S)-g
 relative.bias.ate <- (((sum(Mtch[,1]/S)-g)/g) * 100)
 sd.ate <- sqrt(((S-1)^-1)*sum((Mtch[,1]-(sum(Mtch[,1]/S)))^2)) # more precise sd
 MSE.ate <- (sum(Mtch[,1]/S)-g)^2 + (sqrt(((S-1)^-1)*sum((Mtch[,1]-
(sum(Mtch[,1]/S)))^2)))^2
 ###########ols
 ate.reg <- sum(Mtch[,3]/S)
 raw.bias <- sum(Mtch[,3]/S)-g
 relative.bias <- (((sum(Mtch[,3]/S)-g)/g) * 100)
 sd.reg <- sqrt(((S-1)^-1)*sum((Mtch[,3]-(sum(Mtch[,3]/S)))^2)) # more precise sd
 MSE.reg <- (sum(Mtch[,3]/S)-g)^2 + (sqrt(((S-1)^-1)*sum((Mtch[,3]-
(sum(Mtch[,3]/S)))^2)))^2
 ####################psm
 psm <- sum(Mtch[,5]/S)
 raw.bias.psm <- sum(Mtch[,5]/S)-g
 relative.bias.psm <- (((sum(Mtch[,5]/S)-g)/g) * 100)
 sd.psm <- sqrt(((S-1)^-1)*sum((Mtch[,5]-(sum(Mtch[,5]/S)))^2))
 MSE.psm <- (sum(Mtch[,5]/S)-g)^2 + (sqrt(((S-1)^-1)*sum((Mtch[,5]-
(sum(Mtch[,5]/S)))^2)))^2
 ########ipw
 ipw.ate <- sum(Mtch[,9]/S)
 raw.bias.ipw <- sum(Mtch[,9]/S)-g
 relative.bias.ipw <- (((sum(Mtch[,9]/S)-g)/g) * 100)
 sd.ipw <- sqrt(((S-1)^-1)*sum((Mtch[,9]-(sum(Mtch[,9]/S)))^2))
 MSE.ipw <- (sum(Mtch[,9]/S)-g)^2 + (sqrt(((S-1)^-1)*sum((Mtch[,9]-
(sum(Mtch[,9]/S)))^2)))^2
 #####subclass
 sub <- sum(Mtch[,11]/S)
 raw.bias.sub <- sum(Mtch[,11]/S)-g
 relative.bias.sub <- (((sum(Mtch[,11]/S)-g)/g) * 100)
```

```
  sd.sub <- sqrt(((S-1)^-1)*sum((Mtch[,11]-(sum(Mtch[,11]/S)))^2))
  MSE.sub <- (sum(Mtch[,11]/S)-g)^2 + (sqrt(((S-1)^-1)*sum((Mtch[,11]-
(sum(Mtch[,11]/S)))^2)))^2
  ###########psm.att
  psm.tt <- sum(Mtch[,7]/S)
  raw.bias.att <- sum(Mtch[,7]/S)-g
  relative.bias.att <- (((sum(Mtch[,7]/S)-g)/g) * 100)
  sd.att <- sqrt(((S-1)^-1)*sum((Mtch[,7]-(sum(Mtch[,7]/S)))^2))
  MSE.att <- (sum(Mtch[,7]/S)-g)^2 + (sqrt(((S-1)^-1)*sum((Mtch[,7]-(sum(Mtch[,7]/S)))^2)))^2

  result<-
data.frame(cbind(naive.ate,raw.bias.ate,relative.bias.ate,sd.ate,MSE.ate,ate.reg,raw.bias,relative.
bias,sd.reg,MSE.reg,
                   psm,raw.bias.psm,relative.bias.psm,sd.psm,MSE.psm,ipw.ate,

raw.bias.ipw,relative.bias.ipw,sd.ipw,MSE.ipw,sub,raw.bias.sub,relative.bias.sub,
                   sd.sub,MSE.sub,psm.tt,raw.bias.att,relative.bias.att,sd.att,MSE.att,icc))

  assign(paste0("result", tau), result)
  filename1 <- paste("result",tau,".csv", sep ="")
  write.csv( get(paste0("result",tau)) , file = filename1)

  assign(paste0("ICC", tau), ICC)
  filename2 <- paste("ICC",tau,".csv", sep ="")
  write.csv(get(paste0("ICC",tau)), file=filename2)

  assign(paste0("Mtch", tau), Mtch)
  filename3 <- paste("Mtch",tau,".csv", sep ="")
  write.csv(get(paste0("Mtch",tau)), file=filename3)
  rm(result)
  rm(ICC)
  rm(Mtch)
}
###
#Sample CI coverage rate sample code using saved MC datasets
##
rm(list=ls())

source("~/R/betas")
folder <-c("r_low_d0","r_low_d1","r_low_d3","r_low_d5",
          "r_mod_d0","r_mod_d1","r_mod_d3","r_mod_d5",
          "r_high_d0","r_high_d1","r_high_d3","r_high_d5")
folder<-as.factor(folder)
beta <- c(t_low_0,t_low_1,t_low_3,t_low_5,
          t_mod_0,t_mod_1,t_mod_3,t_mod_5,
          t_high_0,t_high_1,t_high_3,t_high_5)
```

```
es <- rep(1:4,3)
rc <- c(rep(1,4),rep(2,4),rep(3,4))
for (i in 1:length(folder)) {
  setwd(file.path("~/R/Sim_1000_25/", folder[[i]]))
  #for(s in 1:length(beta)){
    file <- "Mtch0.csv"
    data <- read.csv(file, header=TRUE)
    data <-round(data,5)

    ate.est <- data$ate #give me the first column corresponding to the random effects for each
class
    ate.se <- data$sd.ate #same as above only now for SE
    ate.lower <- ate.est - 1.96*ate.se #construct my lower confidence interval for the random
effects
    ate.upper <- ate.est + 1.96*ate.se #upper

    naive_0 <- mean(ate.lower < beta[[i]] & ate.upper > beta[[i]]  )

    ols.est <- data$ols #give me the first column corresponding to the random effects for each
class
    ols.se <- data$sd.ols #same as above only now for SE
    ols.lower <- ols.est - 1.96*ols.se #construct my lower confidence interval for the random
effects
    ols.upper <- ols.est + 1.96*ols.se #upper

    ols_0 <- mean(ols.lower < beta[[i]]  & ols.upper > beta[[i]]  )

    psm.est <- data$psm.ate #give me the first column corresponding to the random effects for
each class
    psm.se <- data$sd.psm #same as above only now for SE
    psm.lower <- psm.est - 1.96*psm.se #construct my lower confidence interval for the random
effects
    psm.upper <- psm.est + 1.96*psm.se #upper

    psm_0 <- mean(psm.lower < beta[[i]]  & psm.upper > beta[[i]]  )

    ipw.est <- data$ipw #give me the first column corresponding to the random effects for each
class
    ipw.se <- data$sd.ipw #same as above only now for SE
    ipw.lower <- ipw.est - 1.96*ipw.se #construct my lower confidence interval for the random
effects
    ipw.upper <- ipw.est + 1.96*ipw.se #upper

    ipw_0 <- mean(ipw.lower < beta[[i]]  & ipw.upper > beta[[i]]  )
```

```
   sub.est <- data$sub.ate #give me the first column corresponding to the random effects for each
class
   sub.se <- data$sd.sub #same as above only now for SE
   sub.lower <- sub.est - 1.96*sub.se #construct my lower confidence interval for the random
effects
   sub.upper <- sub.est + 1.96*sub.se #upper

   sub_0 <- mean(sub.lower < beta[[i]]  & sub.upper > beta[[i]]  )

   cov_0<-rbind(naive_0,ols_0,psm_0,ipw_0,sub_0)
```