5-7-2020

Propensity score matching and generalized boosted modeling in the context of model misspecification: A simulation study

Briana G. Craig
James Madison University

Propensity Score Matching and Generalized Boosted Modeling in the Context of Model Misspecification: A Simulation Study

Briana G. Craig

A Thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the Degree of

Master of Arts

Department of Graduate Psychology

May 2020

FACULTY COMMITTEE:

Committee Chair: Dr. S. Jeanne Horst

Committee Members/ Readers:

Dr. Dena Pastor

Dr. Christine DeMars

Acknowledgements

I would first like to thank the indispensable Dr. Jeanne Horst, without whom this thesis would not exist. Jeanne, you pushed me out of my comfort zone with this one, but by doing so you've truly helped me learn and grow. For the past two years, you have supported me academically, emotionally, professionally, and personally. You have been my favorite thing about graduate school, because you're an incredible advisor and advocate. Not only is it an honor to work with you, but it has been an absolute pleasure and I'm proud of all that we have accomplished.

I would also like to thank my stellar committee members, Dr. Dena Pastor and Dr. Christine DeMars. Dena, your thoughtful feedback truly pushed me professionally in the best way. You challenged me to think more deeply about the subject, to critically examine the coding, and to write my thesis with the upmost clarity. Christine, your presence on my committee was a true comfort. Through your feedback, your evident attention to detail and skill with statistics truly shined; I felt so lucky to have you in my corner. Thank you both for helping to build this thesis into something I could feel confident about.

Mom and Dad, thank you for you support, for your encouragement, and for your positivity. Without you both, I would not be the person I am now, and I certainly would not have made it this far. Mom, thank you for always answering the phone when I called, and for always making me laugh— I am lucky to have such an astounding emotional support. Dad, thank you for always telling me I could do anything, and be anything— I held those positive mantras close to my heart throughout this process. To my brother, Alex, thank you for making time to talk to me, even on the other side of the globe. Our

conversations were always a lovely reprieve from the stressors of graduate school and helped give me something to look forward to.

Thank you to my wonderful fiancée, Jonathan Kilgore. Jon, thank you for your unconditional love and support, for being there during the highs and lows of this experience, and importantly, thank you for making the coffee every morning! Your positivity and humor lifted my spirits, time and time again. Your presence was an endless comfort in my life, and I cannot express my gratitude for that enough.

Finally, acknowledgements to my cat, Navi, my most adorable supporter.

I thank you all dearly.

Table of Contents

List of Tables

List of Figures

Abstract

In the absence of random assignment, researchers must consider the impact of selection bias – pre-existing covariate differences between groups due to differences among those entering into treatment and those otherwise unable to participate. Propensity score matching (PSM) and generalized boosted modeling (GBM) are two quasi-experimental pre-processing methods that strive to reduce the impact of selection bias before analyzing a treatment effect. PSM and GBM both examine a treatment and comparison group and either match or weight members of those groups to create new, balanced groups. The new, balanced groups theoretically can then be used as a proxy for the balanced groups achieved via random assignment. However, in order to successfully employ GBM and PSM, researchers must properly specify the models used to reduce selection bias. Not only do researchers need to account for all covariates related to bias, but they also need to properly specify polynomial terms or interactions. This study investigated scenarios where either a quadratic term or an interaction term contributed to selection bias, and questioned: (1) how incorrectly specified PSM models, correctly specified PSM models, and GBM approaches compare in their ability to create balanced treatment and comparison groups; and (2) how much these methods reduce treatment effect estimation bias. Ultimately, this study found that PSM methods achieved adequate balance, even when misspecified to omit an interaction or quadradic term. In terms of reducing bias, the correctly specified PSM model performed the best, followed by the incorrectly specified PSM model and then the GBM model. All methods had a more accurate treatment effect estimate than the baseline model, which included no pre-processing for selection bias. Recommendations and implications are offered for researchers.

Chapter One

Introduction

The famous philosopher, David Hume, described causality with the statement: "We may define a cause to be an object precedent and contiguous to another, and where all the objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter" (Hume, 2003[1], p. 469). This definition suggests that causality requires two things: (1) that the "cause" precedes the "effect", and (2) without the "cause" there would be no "effect" (Wainer, 2015). The first requirement is relatively simple to establish, because with a proper research design, one can examine an effect prior to treatment, and then again after treatment. If the effect was not present until after treatment, then it can be said the treatment (the hopeful "cause") preceded the effect. If the effect existed before treatment and stayed the same after treatment, then the treatment cannot have caused the effect.

The second requirement of causality, that without the cause there would be no effect, is more difficult to establish. Who is to say what may have occurred had there been no "cause?" How can one observe two alternate realities in which something happens and simultaneously does not happen? Questions such as these introduce the concept of counterfactuals, and how they apply to cause and effect arguments.

The Counterfactual

"Counterfactuals are at the heart of any scientific inquiry" (Guo & Fraser, 2015, p. 23). Counterfactuals address what *could have happened* had some event not occurred (or occurred differently). Going back to Hume's definition, without the cause, would

[1] This quote came from a republishing of Hume's original 1740 book.

there be an effect? Rubin (1975) once concluded, “No causation without manipulation” (p. 234). In other words, the counterfactual is only relevant when the cause is manipulated (e.g., giving treatment or not giving treatment), rather than when the cause is something fixed, or otherwise unchangeable (e.g., someone’s race). Wainer (2015) further elaborates on this point in his book, *Truth or Truthiness*:

> Thus the statement “she is short because she is a woman” is causally meaningless, for to measure the effect of being a woman we would have to know how tall she would have been had she been a man. The heroic assumptions required for such a conclusion removes it from the realm of empirical discussion. (Wainer, 2015, p. 23-24)

Only variables that can be (ethically) manipulated by a person or researcher can be the “causes of interest.” In the context of the current study, the “cause of interest” will be an individual undergoing treatment, presuming that treatment participation may cause a certain effect. Using an applied example, if a “treatment” is taking a practice test in preparation for an examination, then the comparison would be not taking that practice test. If researchers expect that taking a practice test increases a score on a final, then the “cause of interest” would be taking the practice test, and the “effect” would be an increase of score on the final.

The counterfactual would then examine the effect with and without the “cause of interest.” In the applied example, the counterfactual investigates the final exam score (outcome) of a practice test taker had they never taken the practice test to begin with; or conversely, the final exam score of someone who did not take a practice test, had they

taken the practice test. This definition of counterfactuals is the reason counterfactuals often go by the alibi of "potential outcomes."

However, an individual cannot simultaneously go into a final having taken the practice test *and* having never taken the practice test. Likewise, individuals can never simultaneously be in both the treatment and comparison conditions. So, to observe counterfactuals would be to observe something impossible. Researchers can never directly estimate the size of an effect for an individual without the true counterfactual, so they must rely on research design and proper statistical analysis to approximate the counterfactual for a group instead (Wainer, 2015). Frequently, researchers use randomized experiments for their research design in order to estimate the counterfactual for a group.

Randomized Experiments

The randomized experiment is often the considered the "gold-standard" of research design. Although some researchers use the terms "randomized experiment" and "true experiment" interchangeably, a randomized experiment refers to a study in which contrasted treatments (e.g., treatment and control) are assigned to experimental units by chance (e.g., coin toss), while a true experiment is vaguely defined as any study that includes a manipulated independent variable and an observed dependent variable (Shadish, Cook, & Campbell, 2002). In short, randomized experiments require random assignment, while "true experiments" do not.

Random assignment refers to the process of assigning treatment group membership independently of baseline characteristics. When a researcher conducts a randomized experiment, random assignment ensures that each study participant has the

same probability of treatment group membership (and consequently the same probability of control group membership) as any other individual in the study. If assignment were decided by a coin flip, then everyone has a 50% chance of being assigned into the treatment group and a 50% chance of being assigned into the control group. Random assignment is not to be confused with random selection (or random sampling), which refers to the process of picking a sample from the broader human population. While random assignment strengthens the argument for causality, random selection strengthens the argument for generalizing the results to a larger population.

Randomized experiments are able to approximate the counterfactual at the group level because of random assignment. When treatment assignment is completely random, both the treated and untreated groups should have similar distributions of baseline covariates. Because the baseline covariates did not influence assignment and are similarly distributed between groups, the control group is theoretically similar to what the treatment group would have been without treatment. Thus, the control group is a proxy to the treatment group's counterfactual (Rosenbaum & Rubin, 1983; Wainer, 2015).

In applied research, sometimes barriers arise that prevent random assignment of participants (e.g., ethical standards, resource limitations, etc.). When it is unethical or otherwise infeasible to conduct a randomized experiment, researchers often turn to non-randomized, or observational data. However, without random assignment, researchers lose the plausible claim of group similarity in baseline covariates, and therefore lose a strong argument for approximating the counterfactual. In order to maintain scientific rigor in the observational setting, researchers must carefully consider whether they can account for the counterfactual using a quasi-experimental research design.

Quasi-Experimental Design

Although quasi-experiments do not employ randomization, they still have many basic similarities to randomized experiments; both methods test hypotheses and attempt to make causal claims, just in different ways. Quasi-experimental designs need to compensate for the selection bias introduced by the absence of random assignment.

Selection bias occurs when groups systematically differ in baseline characteristics due to the processes by which individuals become a member of those groups. Often, this is conceptualized as self-selection, where certain characteristics may increase the likelihood of an individual *choosing* to select into treatment (e.g., a highly motivated student may be more likely to complete an optional practice test). However, selection bias may also be the result of myriad factors. Financial selection may occur if participating in treatment requires a certain degree of disposable income. If someone cannot afford transportation, childcare, technology (e.g., computers, phone, internet), or treatment fees, then they cannot participate in a study, even if they desire to (e.g., a highly motivated student cannot afford the fee to take the optional practice test). Geographic selection may occur if a treatment exists only in specific geographic locations (e.g., a student's town does not have a testing center for taking the optional practice test). Selection biases such as these are a major threat to quasi-experimental methodology, because they damage the ability to make casual claims, and thus weaken internal validity (Austin, 2011; Austin et al., 2007; Guo & Fraser, 2015; Shadish et al., 2002).

According to Shadish et al. (2002), there are four types of validity: (1) internal validity, (2) external validity, (3) construct validity, and (4) statistical conclusion validity. Although all types of validity have implications for causal inferences, internal validity is

most directly related to causality. Some researchers interpret internal validity as the *sine qua non,* or the type of validity that is a necessary element to proper research. This is likely because causation is both at the heart of internal validity and at the heart of scientific inquiry. Internal validity examines further whether a causal relationship exists between the treatment and the outcome within the context of the study. Internal validity is often confounded by forces that could have occurred in the absence of treatment, which touches on the second piece of Hume's (2003) definition, that without the cause, the effect would not occur (Shadish et al., 2002).

Selection bias introduces an alternative explanation of an effect; did the treatment cause the effect, or did the a priori group differences cause the effect? In order to make causal claims in the face of selection bias, researchers must consider, then rule out all possible alternative explanations and confounds. Confounds refer to extraneous variables that covary with the outcome, or variable of interest (Shadish et al., 2002). To rule out alternative explanations and the effect of confounds, researchers and statisticians developed a series of quasi-experimental techniques.

Techniques to Reduce Selection Bias

When circumstances prohibit the use of random assignment, researchers have three options for controlling selection-related confounders: (1) use a research design that rules out alternate explanations for the cause and effect relationship (e.g., pre-tests or observations over time), (2) use statistical models to adjust treatment effects to account for sources of bias (e.g., ANCOVA), or (3) pre-process groups to balance them on specific covariates (e.g., stratification or matching) before analysis. Unfortunately, flaws exist for each of these techniques. Although incorporating strong elements into quasi-

experimental designs (such as a pre-test) can strengthen causal inferences, designing such a study requires a considerable amount of resources and advance planning. Additionally, an improved research design still does not always rule out alternative explanations. The second technique has theoretical and practical issues, as statistical models such as ANCOVA do not directly model bias, and statistical power decreases as each new covariate is incorporated. The third technique runs into problems if a researcher wants many levels of stratification or matching on many specific covariates (Bai & Clark, 2018; Shadish et al., 2002). However, one solution for the third technique, is to create a single value that summarizes a series of covariates, or in other words, to create a propensity score (Rosenbaum & Rubin, 1983).

Propensity scores denote an individual's probability of treatment, conditional on observed distributions of baseline covariates (Austin, 2009). Therefore, two individuals with the same "true" propensity score have similar distributions of covariates, regardless of treatment assignment (Rosenbaum & Rubin, 1983). Because propensity scores describe the distributions of multiple confounding covariates in a single composite value, several statistical methods use propensity scores to analyze quasi-experimental data with the aim of mimicking the rigor of a randomized experiment.

In randomized experiments, the propensity scores are fixed by the study design. For example, a researcher may determine that each individual has a 50% chance of being assigned treatment. If random assignment is done correctly, each individual has the same probability of treatment, and thus the same true propensity score (Rosenbaum & Rubin, 1983). In quasi-experimental designs, however, the propensity scores must be estimated, as they are not fixed by the study design, due to selection bias. Propensity scores can be

estimated with a variety of statistical techniques (Austin, 2009; Rosenbaum & Rubin, 1983). This paper will focus on the traditional logistic regression technique and the newer generalized boosted model (GBM) technique.

Logistic regression is frequently used to create propensity scores, as it predicts a binary outcome (e.g., treatment or comparison group) by modeling a series of researcher-chosen covariates. GBM is a machine learning-based technique that can model complex relationships to create propensity scores. GBM produces propensity scores by splitting (and classifying) data iteratively and "boosting" misclassifications in order to improve predictions. In GBM, the resulting propensity scores are the "average" of many propensity score models. Unlike logistic regression, GBM is entirely data-based (Bai & Clark, 2018; McCaffrey et al., 2013; Sinharay, 2016).

Propensity Score Matching. Propensity scores have many applications but are most commonly used to conduct propensity score matching (PSM; Austin, 2009). PSM is one technique that attempts to replicate the covariate balance achieved via random assignment. To do this, propensity scores are estimated for each individual. Afterwards, individuals in the treatment and comparison group are matched based on these scores. Ideally, once a new, matched sample is created, the treatment and comparison groups will have similar propensity score, and thus similar distributions of baseline covariates (Rosenbaum & Rubin, 1983). By employing PSM, a quasi-experimental design can mimic the group composition achieved through random assignment; therefore, the matched comparison group emulates the matched treatment group's counterfactual and vice versa.

Propensity Score Weighting. Alternatively, the technique of propensity score weighting also provides a way of preprocessing data to approximate the counterfactual. Once created, propensity scores may be used to assign weights to individuals in the comparison group – such that individuals in the comparison group who are more similar to the treatment group will receive a larger weights and count for more than their less similar peers who receive smaller weights. Ideally, the new weighted comparison group mimics the group composition of the treatment group; therefore, the weighted comparison group emulates the treatment group's counterfactual.

Purpose of the Study

The current study will be a simulation study that compares propensity score estimators and techniques in the context of model misspecification. One example of how these simulated data could be related to real world situations, would be in psychometric studies which examine the influence of a practice test on a student's exam score. In such an example, the practice tests operate as the treatment, while the score on the exam would operate as an outcome. As students may opt into taking practice tests, the treatment group may be qualitatively different than the comparison group on selection-related covariates. Previous literature considering SAT test preparation finds that already privileged students (i.e., students with unearned advantages based on group membership and parental economic status) are most likely and able to select into treatment (test preparation). Thus, a student's race, ethnicity, gender, family income, parental education, and geographic region all relate to both the levels of test preparation and the final SAT score (Alon, 2010; Buchmann, Cirndron, & Roscingno, 2010; Park, 2012; Park & Becks, 2015). The practice test example is provided in order to ground this simulation study in actual quasi-

experimental designs, and to provide real-world implications of how PSM decisions (e.g., the use of logistic regression or GBM to calculate propensity scores) may influence the inferences drawn under various covariate conditions.

Chapter Two

Review of the Literature

As Chapter 1 focused on the logic of causality, traditional issues with quasi-experimental methods, and potential solutions, this literature review will discuss propensity score techniques in greater depth. This chapter provides a comprehensive review of the decisions made when conducting propensity score matching (PSM), and introduces the fundamentals of generalized boosted modeling (GBM) as a propensity score weighting technique. Finally, the literature review will also briefly review past studies that examine the differences between PSM and GBM.

Although previous literature has compared PSM and GBM on their ability to create a balanced sample, no literature has examined how the two compare when higher-order relationships (i.e., interactions and powers) exist in the data. Thus, this paper seeks to unpack how both methods work, and how propensity score model specification may affect how the balance achieved by PSM compares to that achieved by GBM.

Propensity Score Techniques

Some researchers use regression-based, or covariate adjustment techniques (e.g., ANCOVA) to model and correct for a priori covariate imbalance between groups. However, McCaffrey et al. (2013) laid out five main advantages to using propensity score techniques instead: (1) by summarizing a group of covariates, propensity scores offer a succinct way for evaluating treatment effects; (2) propensity scores methods offer a formal model for causal inference, (3) bias from mispecifying the model for the mean can be avoided, as propensity score techniques do not require modeling the mean; (4) while parametric regression modeling may extrapolate whenever the treatment and comparison

groups differ, propensity score methods do not extrapolate; and (5) propensity score adjustments can be implemented without any use of the outcomes, only a priori covariates and treatment assignment, and this removes the potential for covariates to be chosen based on their impact on the estimated treatment effect.

A variety of propensity score techniques can be found in the literature. Propensity scores have been used for covariate adjustment, stratification, inverse probability of treatment weighting, and matching (Austin & Mamdani, 2006; Austin 2009). In the medical literature, propensity score matching (PSM) is used most frequently as a way to handle observational data (Austin, 2009).

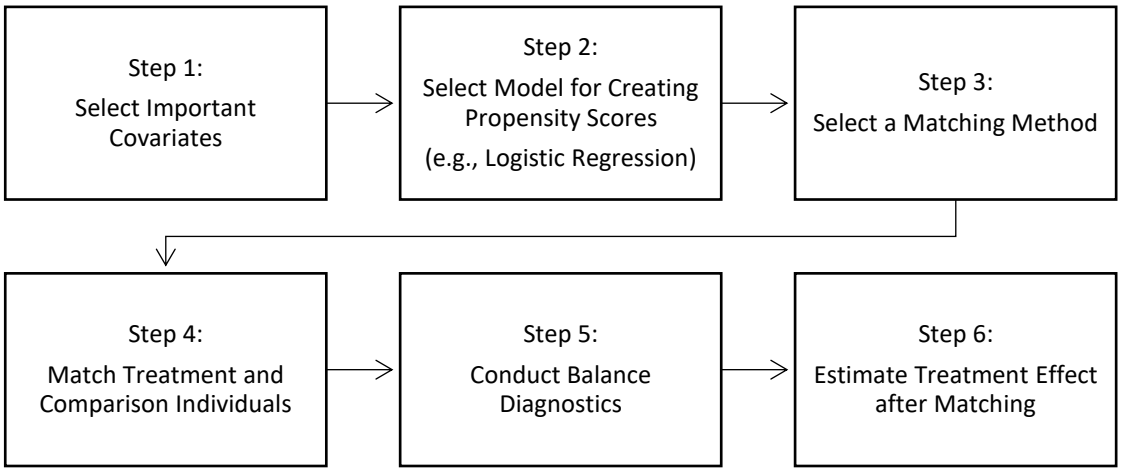
Traditional Propensity Score Matching

PSM is a technique that involves using propensity scores to match individuals in the treatment group with individuals in the comparison group. By doing so, researchers create a new, matched sample, which theoretically controls for the systematic bias of covariates related to self-selection (Rosenbaum & Rubin, 1983).

PSM involves several steps, which have been laid out by various authors (Benedetto et al., 2018; Caliendo & Kopeinig, 2008; Stuart, 2010). Harris and Horst (2016) endorsed a six-step model that summarizes the general process for PSM found in the literature (Figure 1). The first step is to examine the literature and run baseline analyses to select covariates that are important for creating a propensity score. The second step is to incorporate these covariates into a model, such as a logistic regression model (or a generalized boosted model). After you have created those propensity scores, the third step is to select a method for matching the treatment and comparison group individuals. Once that method is selected, then the fourth step is matching individuals to

create a new, matched sample which will be used for the rest of the process. In the fifth step, the researcher assesses balance in the matched sample, in order to ensure that the PSM process successfully reduced selection bias. Finally, the sixth step involves analyzing the new matched sample to estimate the treatment effect.

Figure 1
Six Steps for Propensity Score Matching



Note. Figure adapted from Harris & Horst (2016).

Step 1: Covariate selection. Steiner, Cook, Shadish, and Clark (2010) stated that, "…choice of covariates is more important than the choice of analytic method, assuming that the analysis is competent and sensitive to the assumptions required" (p. 264). Although it is possible for propensity score analyses to yield the same results as randomized experiments, propensity scores will only effectively reduce selection bias if propensity scores are adequately modeled (Bai & Clark, 2018). Ideally, covariates should be chosen based on theoretical foundation and statistical relationships with the outcome. To establish a theoretical basis, a thorough literature review should always be the first step in selecting covariates. This literature review allows researchers to familiarize

themselves with what variables have historically influenced selection and the treatment effect estimation (Bai & Clark, 2018).

After familiarizing oneself with the literature, a researcher should then consider how the covariates, treatment assignment, and outcome statistically relate to each other. Researchers should consider preliminary statistical assessment in determining appropriate covariates. Doing so not only allows researchers to examine which variables relate to the outcome and treatment group selection but can also hint to whether collinearity may be an issue in the chosen propensity score estimation model (Bai & Clark, 2018).

The accuracy of estimates depends on the assumption of strong ignorability. Strong ignorability relies on the idea that each person in a study has two potential outcomes [$Y = (Y_0, Y_1)$]: an outcome that would occur if given no treatment (Y_0), and an outcome that would occur if given treatment (Y_1). Strong ignorability is met when two things happen: (1) treatment assignment (Z) and the potential outcomes are conditionally independent given the observed covariates **X** [$Pr(Z|\mathbf{X}, Y) = Pr(Z|\mathbf{X})$], and (2) there is a nonzero probability of being in either condition [$0 < Pr(e(\mathbf{x}_i)) < 1$, for all $\mathbf{x}_i$, where $e(\mathbf{x}_i)$ represents the propensity scores], implying that each individual has some chance of either outcome (Rosenbaum & Rubin, 1983; Shadish, 2010).

In other words, the propensity score model should have no unmeasured confounders. In this context, confounders refer to covariates that may be influencing the independence of the outcome and treatment assignment. As the goal of propensity score techniques is often to isolate the influence of confounders, researchers, ideally, hope to include all possible confounders. When these confounders are not included in the model, the propensity score model has violated a key assumption of no unmeasured confounders.

Failure to include an important confounder, or other types of misspecification in a model (e.g., failure to include an interaction or polynomial) can lead to biased estimation of treatment effects (Austin, 2007; Drake, 1993).

The strong ignorability assumption is clear in theory, but in applied propensity score research it is difficult to determine whether the included covariates capture the selection bias, or even to what extent a bias actually exists. In most observational studies, strong ignorability is assumed rather than directly tested, because there are no tests that can determine whether the covariates allow condition selection to have the same independence as random assignment (Shadish, 2010; Steiner et al., 2010). Some researchers incorrectly believe that attaining good balance is indicative of meeting the strong ignorability assumption, but Shadish (2013) stated that, "balance may be necessary, but it is not sufficient for strong ignorability to be met" (p. 134).

Some researchers try "kitchen sink" methods of choosing covariates with the logic that if all variables are included into the model, then there should not be any unmeasured confounders. However, the inclusion of more covariates does not always lead to a reduction in selection bias (Brookhart et al., 2006; Steiner et al., 2010). For example, if a researcher uses a large number of covariates with certain PSM methods then the matched sample size may be dramatically reduced, as finding matches becomes more difficult (Austin, 2009). Other researchers build propensity score models using predictors of convenience, or covariates that are readily available (e.g., gender, marital status, age). This is considered bad practice, because propensity score based on predictors of convenience does not reduce bias well on average (Shadish, 2010).

Researchers examining an outcome should only include true confounders or potential true confounders in the propensity score model. True confounders are covariates that relate to the chosen outcome, as well as the selection bias. Therefore, the propensity score model should not include policy or temporal variables associated with selection but not the outcome (Austin, Grootendorst, & Anderson, 2007). Additionally, the propensity score model should not include any discriminatory covariates that were used as part of the criteria for entering treatment, as this would introduce propensity score with a zero value (Austin, 2011; Stuart, 2010). For example, if a treatment is only offered to female-identifying individuals, then gender should not be included in the model, because a male-identifying individual would have a zero probability of treatment [$Pr(e(\mathbf{x}_i) = 0]$.

Step 2: Propensity Score Estimation. Propensity scores have been estimated with a variety of techniques, including but not limited to, discriminant analysis, multiple regression, and logistic regression (Austin, 2009; Rosenbaum & Rubin, 1983; Stuart, 2010). Of these, researchers use logistic regression most frequently (Austin, 2009, 2011).

Logistic regression is a statistical technique used to predict a binary outcome (e.g., 0 or 1, treatment or no treatment) from a set of predictors that may be categorical or continuous. Due to the binary outcome, the errors will not be normally distributed, which fails an assumption of the commonly used general linear model, which is typically estimated using ordinary least squares (OLS) estimation. Instead, logistic regression must use a generalized linear model, which uses maximum likelihood estimation (MLE), rather than the OLS estimation method (Azen & Walker, 2011; Cohen, Cohen, West, & Aiken, 2003).

To elaborate further on the difference between the general and generalized linear model, the general linear model includes a model for the means (fixed effects) and a model for the variances (random effects). The model for the means is often what researchers are interested in when they are testing hypotheses, as it models the relationship between the predictors and the outcome. The model for the variance is often what researchers must make assumptions about and describes how the residuals are distributed across cases. In the general linear model, the assumption is that errors are normally distributed (so when errors are not normally distributed researchers must use the generalized linear model). The general linear model is sometimes written as:

$$Y_{outcome} = (\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K) + e \quad (1)$$

where $Y_{outcome}$ represents the value of the outcome, $(\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K)$ represents the model for the (conditional) mean of $Y_{outcome}$ and e represents the model for the conditional variance of $Y_{outcome}$. When researchers are *predicting Y*, the formula then becomes:

$$Y'_{outcome} = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K \quad (2)$$

where $Y'_{outcome}$ represents the predicted value of the outcome given the *k* predictors in the model, $(\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K)$ represents the model for the (conditional) mean, and the error can be found by subtracting $Y'_{outcome}$ from $Y_{outcome}$

As logistic regression assumes a Bernoulli distribution[2] for the errors, it uses the generalized linear model. The generalized linear model includes models for the mean and variance as well, but additionally includes a link function (that is not an identity link

[2] The Bernoulli distribution is a simple probability distribution for categorical data that can be used to determine the probability of success (e.g., treatment group assignment) for a single trial (e.g., one individual; Azen & Walker, 2011).

function). The link function transforms a non-normal (or in this context, binary) expected outcome into something that can be modeled as a linear function of the predictors (Azen & Walker, 2011; Cohen, Cohen, West, & Aiken, 2003). The generalized linear model is sometimes written as:

$$g(Y'_{Outcome}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K \quad (3)$$

where Y'_{outcome} represents the predicted value of the outcome given the *k* predictors in the model, $(\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K)$ represents the model for the (conditional) mean, and $g(.)$ represents the link function. When the link function is equal to 1, then the generalized linear model simplifies down to the general linear model.

In logistic regression, the link function is a logit link. The logit link is a logit transformation of the expected value of Y. This transforms a bounded, dichotomous expected value of an outcome (0,1) to an unbounded value that ranges from negative to positive infinity. This transformation allows the association between each predictor and the transformed expected value of the outcome and the predictors to be modeled using a linear model. The logit is the natural log of the odds of the event occurring, and is the default predicted score given by logistic regression, as it is the result of the logit link function. Therefore, the simple logistic regression equation becomes:

$$\text{Logit}\,(Y'_{Outcome}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K \quad (4)$$

Unfortunately, logits are difficult to interpret; therefore, researchers often transform the values to odds or probabilities for ease of interpretation. Odds can be derived by exponentiating the logit, or by dividing the probability of an event occurring by the probability of the event not occurring. Odds and odds ratios can be as low as 0 and increase to positive infinity. Probability can be derived from the odds by dividing the

odds by one plus the odds and is often the unit most familiar to the general public (Meyers, Gamst, & Guarino, 2005; Osborne, 2012).

To estimate a propensity score via logistic regression, a model should be built using covariates theorized to influence selection and the outcome as the predictors of the propensity score, and treatment group membership as the binary outcome (i.e., treatment or comparison). Researchers can choose to incorporate interactions and polynomials of the covariates into the logistic regression model, if such relationships are theorized to exist. The logistic regression model assigns a logit value to each individual, which can be transformed into a probability to operate as that individual's propensity score.

The focus of this section so far has been on logistic regression as the propensity score estimator, as logistic regression is most commonly used for PSM (Austin, 2009, 2011). Although this study examined traditional logistic regression estimation approaches, these methods were also be compared to the approach of using generalized boosted models for quasi-experimental analysis. As generalized boosted models are not typically used for PSM, they will not be discussed in this section about the six steps of PSM. Instead, generalized boosted models will be covered more extensively later in the literature review, along with propensity score weighting.

Step 3 and 4: Matching. After the propensity scores are estimated, individuals in the treatment group will be matched to individuals in the comparison group, to create a new, matched sample that hopefully resolves the threat of selection bias. This paper will focus on one-to-one matching methods, as they are more common in the literature than one-to-many matching. One-to-one matching involves matching one treatment group individual to one comparison group individual, while one-to-many matching involves

matching one treatment group individual to many comparison group individuals (i.e., 2 or 3). In sequential one-to-many matching (without replacement), everyone in the treatment group matches with someone in the comparison group (just as in one-to-one matching), then additional second, third, and higher-level matches are made from the remaining individuals in the comparison group (Parsons, 2004; Rassen et al., 2012). One-to-many matching is used less frequently than one-to-one matching, but there are several matching methods equipped to handle such a design (e.g., radius matching or nearest neighbor matching).

A variety of one-to-one matching methods are available, each with its own pros and cons. Researchers may decide on a matching technique depending on specifications and expectations for the study's matched sample size and quality of matches. Matched sample size is important when the treatment group is small prior to matching, as the matched sample will likely be small already and decrease further if treatment group individuals are lost.

Nearest Neighbor. Nearest neighbor (NN) matching relies on a greedy algorithm to match individuals in the treatment and comparison groups. The greedy algorithm starts with the first individual in the treatment group (typically sorted in descending order by propensity score) and matches them to the individual in the comparison group who has the propensity score closest in value; both of those individuals are then removed so that they will not be matched again in the following iterations. Afterwards, the algorithm continues down the list, matching each individual in the treatment group with the remaining unmatched comparison group individual with the closest propensity score.

NN is popular because it is easy to use, and it will match every individual in the treatment group[3]. By matching everyone in the treatment group, the matched sample size stays as large as possible. However, every individual is simply matched to the "best option" remaining in the larger group, regardless of how different the propensity score values may be. Because NN never re-evaluates those matches to determine if better ones could have been selected, this matching process is dependent on the order of the participants. For example, if two treatment group participants have a propensity score of .55, then the one listed earlier in the dataset may match a comparison group individual propensity score of .51, and the one listed later may be matched with the next closest individual, who has a propensity score of .13. Situations such as the above make nearest neighbor matching methods less appealing, as the risk of poor-quality matches may bias the treatment effect (Harris & Horst, 2016; Smith, 1997); although, typically, NN is still a decent option for PSM (Gu & Rosenbaum, 1993).

To fix potential quality issues with NN matching, some researchers incorporated changes into the NN approach. One such change is the introduction of replacement, where an individual from the comparison group could be matched multiple times, if they were closest to several treatment group participants. Although some researchers suggest that matching with replacement is better than matching without replacement (Bai, 2015), others do not recommend this approach as the data become dependent (Austin, 2009).

Nearest Neighbor with Caliper. Instead of using nearest neighbor or replacement techniques, researchers can use NN with a caliper adjustment. With a caliper adjustment, researchers can specify an "acceptable" distance within which matches can be made. This

[3] An exception would be in situations where the comparison group is smaller than the treatment group, but this is not recommended (see *Sample Size Ratio*, p. 41-42).

distance is often a value created by multiplying a fixed amount (e.g., .1 or .2) by the standard deviation of a logit from the propensity score model. Unlike NN without a caliper, this approach does not match everyone in the treatment group. If no individual in the comparison group has a propensity score within the caliper distance of a treatment group participant, then the treatment group participant will not be included in the resulting matched sample. Thus, it is important for researchers to carefully consider whether the higher quality matches of a smaller caliper are worth the loss in sample size (Jacovidis, Foelber, & Horst, 2017). It is also difficult to determine what size difference in propensity score should be considered tolerable in the first place (Caliendo & Kopeinig, 2005; Smith & Todd, 2005).

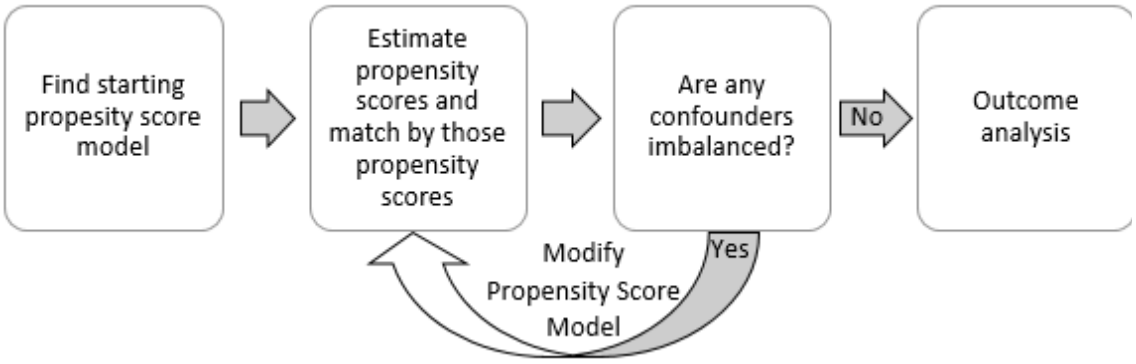
Optimal Matching. Optimal matching also offers an alternative for the potentially poor-quality matches made by NN without caliper adjustment. Optimal matching allows matches to be reconfigured to increase the global fit. In other words, after the initial matching process, pairs may be broken up and reassigned in order to minimize the overall distance between propensity scores among the matches (Rosenbaum, 1989; Stuart, 2010).

Genetic Matching. The above matching methods all rely on proper specification of the propensity score model. These methods have no definitive process for reconsidering the propensity score model, except for researchers to try a variety of models if the balance is not ideal. Because outcome data are not included in the propensity score model, creating multiple models is not often viewed as a sequential testing problem (Diamond & Sekhon, 2013).

Genetic matching eliminates the need to manually check the propensity scores by employing an iterative process which checks the model for misspecification (Figure 2).

An evolutionary search algorithm proposes iterative batches of weights. In each batch, many matched samples are produced, and then evaluated for loss (e.g., individual discrepancy measured by Kolmogorov-Smirnov tests). The model converges towards the weights which produced the smallest amount of loss, which is considered the "optimal solution." In short, genetic matching uses multiple iterations to find the best weights in a propensity score model to improve the balance between matched treatment and comparison groups (Diamond & Sekhon, 2013).

Figure 2
Iterative Estimation of a Propensity Score Model



Note. Adapted from Diamond and Sekhon's (2013) flowchart.

Other Forms of Matching – An Aside. There are other quasi-experimental matching methods that do not directly incorporate propensity scores but are often used in conjunction with PSM techniques. These matching methods include approaches such as exact matching and matching on Mahalanobis distance. Although neither of these techniques will be used in the current study, they are worth mentioning for the sake of comprehensiveness.

Exact Matching. Exact matching involves matching individuals who have the same value on specific covariates, rather than matching them on a propensity score. Exact

matching is used to match important covariates, often categorical variables. Choosing continuous variables, or categorical variable with many levels may result in a greater loss of individuals from the matched sample, because it will be more difficult to find an exact match when the covariate has more variety between individuals. For example, exact matching on whether an individual passed or failed a test would result in a greater sample size than exact matching on the score each individual received on that test. Additionally, it is more difficult to exact match as the number of covariates increase, because two individuals will only match if they have the same values for every covariate chosen for exact matching. For example, when exact matching on gender and education, a female with a Ph.D. could only be paired with another female with a Ph.D. If matching on gender, education, and state, then a female with a Ph.D. from Nebraska, could only be paired with a female with a Ph.D. from Nebraska, which reduces the pool of potential matches from the sample.

Mahalanobis Distance Matching. Sometimes referred to a Mahalanobis metric matching, Mahalanobis distance matching (MDM) was a predecessor to PSM (Guo & Fraser, 2015; Rubin, 1979). MDM is distance-based, rather than model-based (no logistic regression). Treatment and comparison group individuals were matched based on the Mahalanobis distance $d(i, j)$ calculated with the following formula:

$$d(i,j) = (\mathbf{u} - \mathbf{v})T\,\mathbf{C}^{-1}(\mathbf{u} - \mathbf{v}) \quad (5)$$

where **u** and **v** correspond to the vector of matching variables for treatment group participant *i* and comparison group participant *j*, respectively, and **C** corresponds to the sample covariance matrix of the matching variables from the full comparison group (although some researchers define **C** differently). When many covariates are included,

Mahalanobis distances between observations tend to be larger, and it is increasingly difficult to find matches (Guo & Fraser, 2015).

Once the Mahalanobis distance has been calculated, then MDM can be achieved through the greedy matching on the Mahalanobis distance values (NN). Rosenbaum and Rubin (1985) found that MDM reduced standardized differences for individual coordinates of x better than PSM methods but did not reduce standardized differences along the propensity score as well. As a result, Rosenbaum and Rubin recommended a hybrid approach which used MDM with calipers defined by propensity scores.

A Brief Comparison of Propensity Score Matching Methods. Researchers most commonly use NN and NN with caliper adjustment for PSM (Austin, 2009; Harris & Horst, 2016; Stuart, 2010); however, between the two, NN with caliper produces higher quality matches (reduces selection bias more) than NN without caliper (Bai, 2011). When there is a large comparison group to treatment group ratio (i.e., many more individuals in the comparison group), then optimal matching performs similarly to NN in terms of balance achieved. However, when there is a smaller ratio of comparison group to treatment group individuals, optimal matching methods will perform better (Austin, 2011; Gu & Rosenbaum, 1993). In simulation studies genetic matching was also more effective than NN matching at reducing selection bias (Diamond & Sekhon, 2013). Given that the purpose of matching is to create balanced groups, researchers must evaluate whether the groups are truly balanced to determine whether the proper matching methods were used.

Step 5: Assessing Balance Diagnostics. Ultimately, propensity scores are balancing scores, so logically, the quality of propensity score estimation is directly

connected to the quality of balance that is achieved after matching. To suggest that a matching method has achieved balance, is suggesting the distributions of baseline covariates are similar between the matched treatment and comparison groups (Ho et al., 2007). Austin (2009) examined the efficacy of several numeric and visual methods for assessing balance diagnostics when propensity score matching.

Significance Testing. Some researchers have argued for using significance tests (e.g., t-tests) to determine whether the covariates have similar distributions, and thus, balance (Pan & Bai, 2015). However, this is not a theoretically sound approach to balance diagnostics for two reasons. First, the matched sample will be a reduced version of the unmatched sample, which decreases statistical power and consequently, the ability to detect imbalance. Therefore, any perceived improvement in balance from the unmatched to the matched sample, may actually be an artifact of reducing sample size and power. Second, inferential statistics are intended to be used when a researcher desires to make inferences about a larger population. Balance, however, is a property of a particular sample rather than a larger population. Because inferential statistics are intended for inferences about populations, not samples, then they should not be used for determining properties of samples (Austin, 2009; Imai, King, & Stuart, 2008).

Comparing Means. One method of numerically diagnosing balance is to compare the standardized difference in propensity score (and individual covariates) between groups. Also known as standardized bias, the standardized difference for continuous variables can be found with the following formula, based on Cohen's *d*:

$$d = \frac{(\bar{X}_{Treatment} - \bar{X}_{Comparison})}{\sqrt{\frac{s^2{}_{Treatment} + s^2{}_{Comparison}}{2}}} \quad (6)$$

where $\bar{X}$ denotes the sample mean of the covariate of interest in the treatment and comparison group, and s^2 denotes the sample variance of the covariate in the treatment and comparison group (Austin, 2009).

Less commonly used, a similar formula finds the standardized difference for dichotomous variables:

$$d = \frac{(\hat{p}_{Treatment} - \hat{p}_{Comparison})}{\sqrt{\frac{\hat{p}_{Treatment}(1 - \hat{p}_{Treatment)} + \hat{p}_{Comparison}(1 - \hat{p}_{Comparison)}}{2}}} \quad (7)$$

where $\hat{p}$ denotes the mean (i.e., proportion) of the variable in the treatment and comparison group (Austin, 2009).

Unlike statistical tests, the standardized difference is not influenced by sample size. Austin (2009) suggested, "In observational studies, as in randomized experiments, balance is a large-sample property; moderate imbalance can be expected in small samples, even if the propensity score is correctly specified." Currently, there is no consensus on what value constitutes balance or imbalance. Normand et al. (2001) suggested that a difference of .1 denoted meaningful imbalance, and this criterion has been resounded in other literature (Austin, 2009, 2011). What Works Clearinghouse proposed more stringent guidelines for achieving baseline equivalence; standardized differences should be less than a quarter of the standard deviation when the analysis includes acceptable statistical adjustment[4] [<.25(*sd*)], or below one twentieth of the

[4] What Works Clearinghouse considers a variety of statistical adjustments to be acceptable, depending on the relationship between the outcome and the covariate in question. One example of this could be including the imbalanced covariate into an ANCOVA model.

standard deviation when the analysis does not include statistical adjustment [<.05(*sd*); What Works Clearinghouse™ Standards Handbook][5].

***Percent Bias Reduction*.** Another helpful indicator of balance is to examine the percent reduction in bias from the unmatched sample to the matched sample. This value should be calculated for the propensity scores, as well as for each of the covariates used in the propensity score modeling and matching process. The percent bias reduction (PBR) can be calculated with the following, equivalent formulas:

$$\text{PBR} = \frac{\left(\bar{X}_{Treatment} - \bar{X}_{Comparison}\right)_{before} - \left(\bar{X}_{Treatment} - \bar{X}_{Comparison}\right)_{after}}{\left(\bar{X}_{Treatment} - \bar{X}_{Comparison}\right)_{before}} \times 100\% \quad (8)$$

$$PBR = \frac{|B| - |B_m|}{|B|} \times 100\% \quad (9)$$

where B is the mean difference before matching, and B_m is the mean difference after matching (Pan & Bai, 2015).

Using this formula, a positive percent value indicates that the PSM process reduced bias, and therefore improved balance. A negative value indicates that the PSM process increased or overcorrected for bias, and therefore, balance was made worse. Although there are no established cutoffs for PBR, some recommendations suggest a value of 80% indicates sufficient reduction in bias (Cochran & Rubin, 1973; Pan & Bai, 2015). However, the PBR is greatly dependent on the baseline (unmatched) sample's balance, such that covariates with only mild balance problems before matching will likely not have a large PBR or may overcorrect; however, a small balance improvement may still be important if the covariate greatly influences the outcome or treatment selection.

[5] What Works Clearinghouse uses a measure of standardized difference based on Hedge's G, rather than Cohen's d.

Variance Ratios. Another numeric method of diagnosing balance is through variance ratios. The variance ratio is calculated with the following formula (Stuart & Rubin, 2008):

$$\text{Variance Ratio} = \frac{s^2_{Matched\ Treatment\ Group}}{s^2_{Matched\ Comparison\ Group}} \quad (10)$$

where s^2 denotes the variance of the propensity score (or the individual covariates) for the matched treatment or comparison group, respectively. Rubin (2001) recommended that the ratios of the variance of the propensity scores be close to one, with a deviation of .5 being too extreme. It is recommended that a comparison of means for the propensity scores and covariates is used in tandem with the variance ratio (Harris & Horst, 2006; Ho et al., 2007).

Five-Number Summary. The last numerical method of assessing balance is the examination of the five-number summary, which was suggested by Hoaglin et al. (1983) as an adequate summary of distribution. In the context of PSM, the five-number summary includes the minimum, 1st quartile, median, 3rd quartile, and maximum of each continuous covariate for both the treatment and comparison groups.

Five-number summaries are not commonly used (or reported) in PSM studies, likely because interpretation is difficult. There is no statistical way of determining what amount of variation is reasonable, and what amount suggests a misspecification of the propensity score model. This technique gives researchers a rough, quantitative look at distribution and skew, but may only be useful for assessing issues with balance if the propensity model is grossly misspecified (Austin, 2009).

Visual Analysis. Multiple graphical methods of assessing balance exists, including, but not limited to, side by side boxplots, quantile-quantile (Q-Q) plots, jitter

plots, side by side histograms, density plots, and cumulative distribution functions (Austin, 2009; Ho et al., 2007; Stuart, 2010). These visual methods of assessing balance can be used to compare propensity scores between groups, as well as the balance among individual covariates. Like the five-number summary, graphical comparisons are interpretationally limited, because it involves simply "eyeballing" a graphical summary for any disparity between the treatment and comparison groups. Therefore, it is difficult to determine what amount of deviance is expected from a correctly specified model, and what amount of deviance indicated misspecification. Austin (2009) recommended that visual analyses should be used in addition to numeric methods, as a stronger argument for balance may be made with a combination of numeric and visual diagnostic tools.

Step 6: Treatment Effects. PSM is an approach intended for hypotheses regarding the Average Treatment effect on the Treated (ATT). If the goal is to estimate the treatment effect on the overall population, rather than just treated individuals, then researchers should consider the Average Treatment Effect (ATE; Ho et al., 2007; Rosenbaum & Rubin, 1983). If the researcher's hypothesis involves the ATE, then the data is best handled with propensity score methods other than matching, such as inverse propensity score weighting or stratification (Benedetto, Head, Angelini, & Blackstone, 2018). Despite the distinction between ATT and ATE, knowing one provides a good estimator of the other, and if the causal effects are constant, then the two are identical (Ho et al., 2007).

Regarding the use of inferential statistical methods to determine the treatment effect, there is some debate about whether matched groups should be treated as dependent or independent. Some researchers regard the matched groups to be dependent, as they

believe the matching process ensures similar propensity score values, thus, theoretically, the matched groups come from the same multivariate distribution (Austin, 2011). If the matched samples are considered dependent, continuous variables could be analyzed with paired *t* tests, while binary variables warrant the use of McNemar's test, or certain logistic regression techniques (Benedetto et al., 2018).

Other researchers consider the matched treatment and comparison group to be independent, as the matching process is conducted separately from the outcome, so the outcomes of matched individuals should not be correlated (Schafer & Kang, 2008). This study will borrow Stuart's (2010) justification for independence suggesting that an analysis does not need to account the matching process for two reasons: (1) the conditioning on the covariates used is sufficient, and (2) PSM does not guarantee that the individual pairs are well matched on all covariates, but rather the groups of individuals have similar distributions. Therefore, all of the individuals in the matched sample may be pooled together and incorporated into a regression analysis. After the regression analysis is conducted, the weighted averages of the regression coefficients are used to calculate the ATT.

Once the treatment effect has been estimated, the researcher has finished all six steps of PSM, and may continue on to analyze and discuss the implications of the study. Concluding the description of the PSM steps, the following sections will discuss additional considerations in the PSM process before branching out to cover generalized boosted modeling.

Additional Considerations in PSM

Each of the six steps of PSM discussed above introduces a series of decisions that need to be made at each step (e.g., "How many covariates should I include?" or "Which matching method should be used?"). However, there are also decisions that need to be made before the PSM process, which may impact the decisions made during the six steps. These decisions involve the collection of participants for the study, and consider aspects such as comparison group selection, sample size, and common support.

Comparison Group Selection. Although much attention is often given to the treatment group, the comparison group is equally important for successful PSM. Bias tends to be lower in studies that carefully select a comparison group to be maximally similar to the treatment group on certain characteristics (e.g., both groups are from the same location). Suppose a researcher's treatment group was comprised of individuals in a specific major at a certain university; the best comparison group would be formed from other individuals in that same major at that same university, rather than students from another major or university (Cook, 2008; Shadish, 2013, Shadish & Cook, 2009).

In simulations, there is usually no concern over whether the comparison group is fitting, because both the comparison and treatment groups are created based on theory. However, in applied studies this can become a larger concern, especially when the sampling process is not explicitly discussed – how does one know if proper consideration was given to the initial design of the comparison group? The process of comparison group selection is especially concerning in archival studies, which may pool together individuals who differ in important ways (Shadish, 2013).

Sample Size. PSM can sometimes inflate bias in the effect estimate, rather than reduce it. One method of minimizing this threat is to increase sample size. For this reason, PSM is considered to be a large sample method, but exactly how large has been a source of discussion in the literature. Simulations found that in total samples of $n = 200$, the analysis increased bias about 15-17% of the time. In samples of $n = 500$, this percentage dropped to around 1-3%, and at $n = 1000$, the percentage dropped further to less than 1%. Around $n = 1500$, the chance of increasing bias is completely negligible (Luellen, 2007; Shadish, 2013). This echoed the work of Feng et al. (2011), who simulated samples of $n = 100, 300, 1000, 3000$, and 10000, and recommended moderate to large sample sizes. Additionally, McCandless et al. (2012) simulated samples of $n = 100, 250, 500$, and 1000, and found poor performance when sample size was below 250.

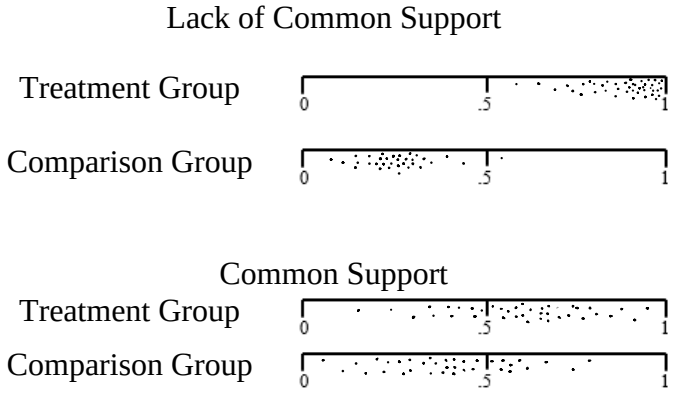
Sample Size Ratio. One variable that moderates the effect of sample size on bias, is the ratio of comparison to treatment group individuals – some researchers even suggest that ratio is more important than sample size (Bai, 2015; Rubin, 1979). It is often recommended to have a much larger comparison group than treatment, so that each treatment group individual has more potential matches to "choose from." The benefits of a higher ratio are most evident when comparing a 1:1 ratio to a 2:1 (comparison group *n*: treatment group *n*). Higher ratios (e.g., 3:1, 9:1) further reduce bias, but by a negligible amount considering the increase in cost that accompanies larger ratios (Rubin, 1979).

Common Support. Common support refers to the extent by which the propensity scores for the treatment and comparison group "overlap" in distribution. When there is more overlap, or high common support, better quality matches can be made. When there is less common support, there may be problematic differences in the distributions of

propensity scores. A lack of common support may result in fewer matched pairs (if using a caliper), which inadvertently leads to a loss of information, particularly with individuals who may be qualitatively different. When estimating treatment effects, a lack of common support damages the ability to make unbiased and representative ATE and ATT estimates (Caliendo & Kopeinig, 2005; Stuart, 2010).

The most straight-forward way of examining common support is through visual analysis. Researchers can create a jitter plot of propensity scores comparing the treatment and comparison groups and look where the propensity scores cluster and overlap (Figure 3).

Figure 3
Jitter Plot Comparison for Common Support



Note. An example of two jitter plots representing the propensity score distributions for the treatment and comparison groups in a scenario that lacks common support (top), and a scenario that has common support (bottom).

Generalized Boosted Models

Generalized boosted modeling (GBM) was developed in the late 1990s and has recently gained popularity with the growing interest in machine learning. GBM is a supervised learning technique, which refers to a type of machine learning where a researcher supplies input (**X**) and an output variable (Y), and an algorithm is employed to

map the two [Y = f(**X**)]. This is different from unsupervised machine learning, when only input data is supplied (Brownlee, 2016; Lison, 2015).

In short, generalized boosted modeling is a decision tree-based boosting technique that provides probabilities of group membership that can be applied to estimate propensity scores (Westriech, Lessler, & Funk, 2010). Researchers have used the probabilities generated by GBMs to create propensity scores, effectively offering an alternative to the logistic regression approach (McCaffrey, Ridgeway, & Morral, 2004). Although GBMs can be used in situations with multiple treatments (McCaffrey et al., 2013), this paper will continue to focus on treatment assignment as a binary outcome (i.e., treatment group and comparison group).

How GBMs Work. To understand GBM, a handful of data mining techniques need to be described first for context, as the GBM method builds on Classification and Regression Tree (CART) models, random forests, and boosting.

CART Models. Classification and Regression Trees both can use the same inputs and work in the same way, but they differ in the outcome they produce. Classification trees produce categorical outcome estimates, and regression trees produces continuous outcome estimates. CARTs take a dataset and use a series of binary splits to create subsets of the data. The goal of splitting is to get similar values of the outcome within subsets and values as different as possible between subsets.

The first split of a CART model is based on a chosen value of a single input variable. If the input variable is categorical, then subsets will be split as belonging to a category or not belonging to a category (e.g., if examining education as an input, the data may be split by having a high school diploma or not having a diploma). If the input

variable is continuous, the binary splits can occur between any pair of consecutively ordered observed values (e.g., if examining age as an input, the data may be split by persons younger than 18, and persons 18 and older). Out of all possible splits, the algorithm selects the split that is most discriminatory. For regression trees, the most discriminatory value is one that minimizes prediction error, or the discrepancy between the predicted outcome value and actual outcome. For classification trees, the most discriminatory values consider misclassification error, deviance, and the total variance across the classes (measured by Gini index). The tree continues to split the data until the researcher-set "allowable" number of splits has been reached (McCaffrey, 2004). Each split down the tree can use the same input variables (with a different split value) or separate input variables – whatever produces the best split. A predicted value is decided upon by following a pathway (i.e., a series of splits) for an individual based on their covariates, until the end of the tree (i.e., final node/subgroup) is reached. If predicting a categorical outcome (using a classification tree), then the predicted outcome would be whatever outcome was shared by the majority of the individuals in that subgroup/node. If predicting a continuous outcome (using a regression tree), then the predicted outcome would be an average of all individuals in that subgroup/node.

An educational application for regression trees could be electronic essay scoring, while classification trees may be used to examine drop-out status given a set of covariates. Unlike traditional prediction methods (e.g., multiple linear regression or logistic regression), CARTS require no distributional assumption, allowing them to explain more complex interactions among predictors. The flaw of CART modeling is that they can be biased in unbalanced datasets, prone to overfit, and small changes in the data

can lead to very different splits. Ensemble methods, such as random forests and boosting, can be employed to alleviate some of these concerns (Sinharay, 2016).

Random Forests. To further approach how GBMs work, the application of random forests to CART models warrants discussion. In this application, random forests are essentially CART models with bootstrapping. In supervised learning methods, prediction models are constructed from a sample called a training set. To create random forests, a certain number (*B*) of bootstrap samples are drawn from the training set with replacement, so that each bootstrap sample has the same sample size as the training set. Then, a tree is constructed from each bootstrap sample, resulting in *B* trees. Each tree uses a random subset of the available *p* predictors ($\sqrt{p}$ for classification and *p*/3 for regression; Hastie et al., 2009) so that the trees are different from each other (a process known as decorrelating). Afterwards, a predicted value of the response for an observation is decided upon by "combining" the predictions from the *B* trees. For regression trees, the predicted values from the *B* trees are averaged, and for classification trees, the "majority vote" from the *B* trees is used (Sinharay, 2016).

Boosting and GBM. Similar to random forests, boosting also combines predictions from *B* trees. However, boosting accomplishes this in a different way. Instead of using bootstrapping to construct many trees and then combining them, boosting creates several trees sequentially, such that information from the previous tree is used to modify the next tree (Sinharay, 2016). There no longer needs to be bootstrapping or a random subset of predictors, because each tree "learns" from the mistakes (misclassifications) of the trees before it instead.

When GBM is used for propensity score methods, the important baseline covariates are used as the input variables, and treatment group membership is used as the categorical outcome. So, GBM starts with a weak model that guesses whether an individual is in the "treatment" or "comparison" group with an error rate only marginally better than chance. Individuals who have been misclassified (e.g., a treatment group individual who has split into the comparison group category) are "boosted," or given a larger weight in the next iteration. The larger weight increases the chance that the next tree will correctly classify that individual (Sinharay, 2016). This process continues for thousands of iterations until a "stopping rule" has been met. In the context of PSM, GBM's iterative process stops when covariates are balanced. The optimal iteration of GBM (most balance in covariates) is achieved when either the absolute standardized bias is minimized or the Kolmogorov-Smirnov statistic is maximized (McCaffrey et al., 2013).

The optimal iteration chosen by the stopping rule is the one that produces the propensity scores for each individual. These propensity scores are then used to weight the observations when estimating the treatment effect. The propensity score weights adjust the groups so that the treatment and comparison groups have similar distributions of covariates. Therefore, individuals in the comparison group who are more similar to individuals in the treatment group may be given a larger weight, so their covariate distribution counts as "more." Individuals in the comparison group who are less similar to individuals in the treatment group may be given a smaller weight, so their covariate distribution counts as "less." The propensity score weights can be used to produce a weighted ATT estimate, or a weighted ATE estimate (McCaffrey et al., 2004).

Pros and Cons to GBM. GBM is a technique that can effectively model complex relationships, due to its non-reliance on a distributional assumption (ability to model non-normal data). Therefore, trees can handle non-linear relationships, large numbers of covariates, interactions, variable transformation (e.g., log(x) or x^2), and a variety of variable types (e.g., continuous, nominal, ordinal; McCaffrey et al., 2004). Because GBM is a nonparametric model, the chance of model misspecification errors is reduced and, therefore, the treatment effects are less likely to be biased (Drake, 1993; McCaffrey, 2004).

However, when the sample size is small, and the number of covariates is large, then the algorithm may not be able to reach an optimal iteration or find balance. Another downside to the GBM approach is that it is purely data-driven by nature, and like many machine learning techniques, GBM can be criticized for modeling relationships with a numerical, rather than theoretical basis (Burgette et al., 2015).

Propensity Score Weighting

GBM operates best in tandem with propensity score weighting techniques, rather than matching. Propensity score weighting is a technique where observations are multiplied by a derivative of the propensity score in order to achieve balance between groups. The theory behind propensity score weighting is that a sample is weighted such that a new, synthetic sample is created where the distribution of baseline covariates is independent of treatment (i.e., approximate the counterfactual better). Commonly, this is done via inverse probability of treatment weighting (IPTW), where an individual's weight is determined by the inverse probability of receiving the treatment (Austin, 2011; Clark, 2015).

The steps for conducting propensity score weighting are similar to those of PSM (Olmos & Govindasamy, 2015):

1. Examine outcomes[6] and balance before weighting
2. Select method of propensity score estimation
3. Weight estimation using propensity scores
4. Conduct balance diagnostics
5. Outcomes analysis

The similarities to PSM exist in the importance of covariate selection, how balance is assessed, and the importance of picking an adequate propensity score estimator. As with PSM, the propensity scores for weighting can be estimated in a variety of ways, including logistic regression and GBM. However, this study will pair propensity score weighting with GBM as the propensity score estimation method, while logistic regression will be the estimation method for the PSM technique. This plays to the strengths of both estimation methods, as logistic regression is better for matching, and GBM is better for weighting (Bai & Clark, 2018; Stone & Tang, 2013).

One benefit to using propensity score weighting over matching, is that you do not need as large a sample to effectively use it. Unlike matching methods, which tend to lose comparison group or treatment group individuals, weighting methods allow an entire sample to be factored into the final analysis to some degree (Olmos & Govindasamy, 2015).

[6]Although some researchers advise against examining the outcome before employing a propensity score technique, as to avoid researcher bias.

Logistic Regression and GBM

Previous literature comparing logistic regression and GBM's ability to estimate propensity scores found that the best estimation method depended on the propensity score method (e.g., matching, stratifying, weighting). Both logistic regression and GBM typically work well with most datasets, but logistic regression tends to perform better when matching or stratifying, and GBM tends to perform better when weighting (Bai & Clark, 2018; Stone & Tang, 2013).

Several authors have suggested that when GBM uses a "stopping rule" based on minimizing the difference between the weighted distributions of the covariates in the two groups (i.e., treatment and comparison), then GBM estimates propensity score weights that yield better balance scores and smaller mean square error than other propensity estimation methods (Harder, Stuart, & Anthony, 2010; McCaffrey et al., 2004; McCaffrey et al., 2013).

Additionally, logistic regression models can be problematic estimators of propensity scores when the model is misspecified (Lee et al., 2009; McCaffrey et al. 2013), whereas GBM can compensate for misspecification as long as the right covariates have been included. In simulation studies, when logistic regression models are misspecified to omit non-linear and non-additive data, boosted models have been shown in simulations to have substantially better bias reduction (Lee et al., 2009). This is especially important, as logistic regression models are organized by a human researcher, who may not think to include higher order relationships and accidentally incorrectly specify a model. GBM is data-driven, so higher order relationships are more likely to be

factored into the propensity score model, but at the cost of capitalizing on sample-dependent error and losing generalizability.

The Current Study

The current study seeks to elaborate on previous literature comparing logistic regression and GBM as propensity score estimators. As the literature has suggested that logistic regression is more appropriate in the context of PSM, and that GBM is more appropriate for propensity score weighting (Pan & Bai, 2015), each estimation method was paired with the technique it is best suited for. This study examined the differences in balance, and estimated treatment effect between PSM paired with logistic regression, and propensity score weighting paired with GBM.

Research Questions. Specifically, this study is investigating two research questions. In scenarios where either a quadratic term or an interaction term contributes to selection bias:

1. How do incorrectly specified PSM models, correctly specified PSM models, and GBM approaches compare in their ability to achieve covariate balance between the treatment and comparison groups?
2. How much do the above methods reduce treatment effect estimation bias, compared to a baseline model with no matching or weighting?

Chapter Three

Method

The present study compares logistic regression as a propensity score estimator for propensity score matching (PSM) with the newer technique of generalized boosted modeling (GBM) as a propensity score estimator in the context of weighting. Given that logistic regression propensity score models are researcher-set, and therefore prone to misspecification related to missing quadratic relationships and interaction terms, how do logistic regression models compare to GBM in the presence of misspecification? These techniques were evaluated and compared on the quality of matches produced (balance) and the accuracy of estimated treatment effects. This study is an elaboration on a simulation study performed by Austin (2009) who estimated balance and bias differences after matching on correctly and incorrectly specified logistic regression-based propensity score models.

Conditions

To answer the research question, I manipulated two main factors: (Factor 1) the "true" propensity model, and (Factor 2) the propensity score technique or lack thereof (Table 1). Factor 1 contained two levels, which are hereafter referred to as scenarios. In Scenario A, a quadratic relationship exists between one of the covariates and the true propensity score. In Scenario B, an interaction exists between the two covariates with respect to their relationship with the true propensity score.

Factor 2 was therefore comprised of four levels: (1) correctly specified logistic regression as the model for PSM; (2) incorrectly specified logistic regression as the model for PSM, which did not include a polynomial nor interaction; (3) GBM with

weighting; and (4) a baseline model which involved no manipulation of the samples. Therefore, the combination of Factor 1 and Factor 2 results in a total of 8 fully crossed conditions.

Table 1
The 2x4 Design of the Current Study

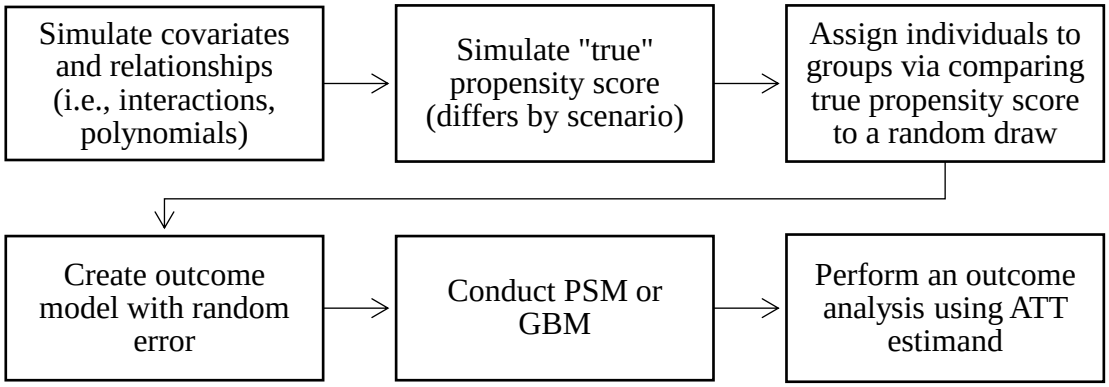
Factor 1: "True" Propensity Score Model	**Factor 2:** Model			
	Correctly Specified PSM	Incorrectly Specified PSM	GBM	Baseline
Scenario A: $Probit(Y_{Group}) = b_0 + b_1X_1 + b_2X_2 + b_3X_2^2$ (Quadratic Relationship)	Condition 1	Condition 2	Condition 3	Condition 4
Scenario B: $Probit(Y_{Group}) = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$ (Interaction)	Condition 5	Condition 6	Condition 7	Condition 8

Note. Scenario A and B represent the structure of the "true" logistic regression models which predict treatment group membership. The variables X_1 and X_2 will be simulated with the cumulative normal distribution. The correct or incorrect specification of PSM refer to the specification of the logistic regression model that will produce propensity scores for the PSM group. The incorrectly specified logistic regression model in Factor 2 will be one that does not include polynomial or interaction terms, $Logit(Y_{Group}) = b_0 + b_1X_1 + b_2X_2$.

Simulation of Data

The current study used RStudio version 1.1.463 (RStudio Team, 2016) to create and analyze the simulated data. Figure 4 outlines the process for simulating data in this study.

Figure 4
Six-Step Process for Simulating Data.



I conducted Monte Carlo simulations that included 1000 replications with 1000 simulees per replication to examine situations where the true propensity score models included (A) a quadratic term or (B) an interaction term. For both scenarios, X_1 and X_2 were obtained from bivariate normal distributions with means of 0 and standard deviations of 1. The correlation between X_1 and X_2 as specified to be 0.3, a correlation intended to emulate relationships often found among real world variables in the educational psychology setting. According to Osborne (2003), the mean effect sizes (d = .68, SD = .37) reported in the educational psychology literature are equivalent to an r = .32. If one considers effect sizes one standard deviation above and below .68, then the range of equivalent rs would be from .16 to .46. I chose an r = .3 to be within that range and similar to what is average in the literature.

In addition to X_1 and X_2, Scenario A included a third variable defined by squaring X_2, or X_2^2. Scenario B included X_1 and X_2, as well as their product, X_1X_2. Therefore, within a single replication, Scenario A and B each included 1000 simulees, and scores from the same simulees were tested across all three conditions of Factor 2 (i.e., correctly

specified PSM, incorrectly specified PSM, GBM). The current study included 1000 replications of this process.

The relationship between the covariates and latent propensity was fixed across the models; X_1 and X_2 had a relationship of r = .2 with the latent propensity and the third variable (X_2^2 in Scenario A, X_1X_2 in Scenario B) had a relationship of r = .5 with the latent propensity, such that the quadratic and interaction terms were more strongly related to treatment assignment compared to the initial two variables. Below I describe separately for Scenarios A and B how I simulated the latent propensity values to align with the aforementioned specifications and to yield the desired proportions of simulees in the treatment and control groups.

Scenario A. Treatment status was generated by first creating "true propensity scores" via a three-step process[7]. First, because I set the relationships among the covariates, as well as the relationship between the covariates and the latent propensity, I was able to produce the probit regression coefficients through matrix algebra[8]. Second, multiplying the data matrix by the vector of probit regression coefficients produced a

[7] An important distinction must be made between $Y_{outcome}$, Y'group, Y_{group}, and the "true propensity scores." $Y_{outcome}$ refers to the overall outcome or the dependent variable that may have been influenced by selection bias. In order to reduce the influence of selection bias, I conducted PSM or GBM to predict group membership based on baseline characteristics. The group membership predicted by either PSM or GBM is denoted, Y'_{group} while the actual, simulated group membership is denoted Y_{group}. The "true propensity score" refers to a simulee's probability of treatment group membership, regardless of whether they were assigned or predicted to be in that group. This "true propensity score" is equivalent to probit(Y_{group}) converted into a probability metric. Both PSM and GBM then produce an "estimated propensity score," which is equivalent to logit(Y'_{group}) converted to a probability metric.

[8] The logic here follows the equation, $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}'_{group}$, where **B** represents the weights of the coefficients, **X**'**X** represents the covariate correlation matrix and **X**'**Y**'$_{group}$ represents the correlations between the covariates and latent probability of treatment group membership, which was hard coded to be .2, .2, and .5 for X_1, X_2, and the third variable (X_2^2 or X_1X_2) respectively. **X**'**X** was procured in a preliminary step, where I obtained values of X_1 and X_2 for 1,000,000 simulees from a bivariate normal distribution with means of 0 and standard deviations of 1 for each variable. X_1 and X_2 were correlated, r = 0.3. Values of X_2^2 in Scenario A and X_1X_2 in Scenario B were then calculated and the correlations among all predictors estimated. The values of all three predictors were then standardized before calculating the beta coefficients.

predicted Y_{group} for each simulee[9]. Third, each simulee was assigned a value from the cumulative probability density function value (on a 0 to 1 scale), which indicated the proportion of scores in the normal curve that fell at or below the predicted value (Y_{group}) for that simulee. This value from the cumulative density function represented their true latent propensity, which was then labeled their "true propensity score." This process outputs propensity scores theoretically similar and empirically, nearly identical to creating true propensity scores via a correctly specified logistic regression model that predicted propensity scores from X_1, X_2, and the quadratic term, X_2^2.

After creating true propensity scores, I assigned a random draw to each simulee (between 0 and 1), such that if the true propensity score was greater than the random draw, then the simulee was assigned to the treatment group (group = 1). If the propensity score was less than or equal to the random draw, then the simulee would be assigned to the comparison group (group = 0). This is the same as pulling a random number for group assignment from a Bernoulli distribution with its defining parameter (p_i) equal to the true propensity score for each simulee. When assigning group membership, the treatment:comparison group ratio was fixed to be approximately 200:800 or 1:4. This was done by rescaling the latent propensity distribution prior to random draw and group assignment[10]. The final propensity for treatment correlated with the true propensity scores, r = 0.998 for both scenarios. This correlation is not a perfect one, as the true

[9]Because the true propensity score is a continuous variable, the relationship between the data matrix and the predicted Y_{Group} is linear, such that $\boldsymbol{Y'}_{Group} = \boldsymbol{XB}$ where **X** is the data matrix and **B** is the vector of probit regression coefficients.

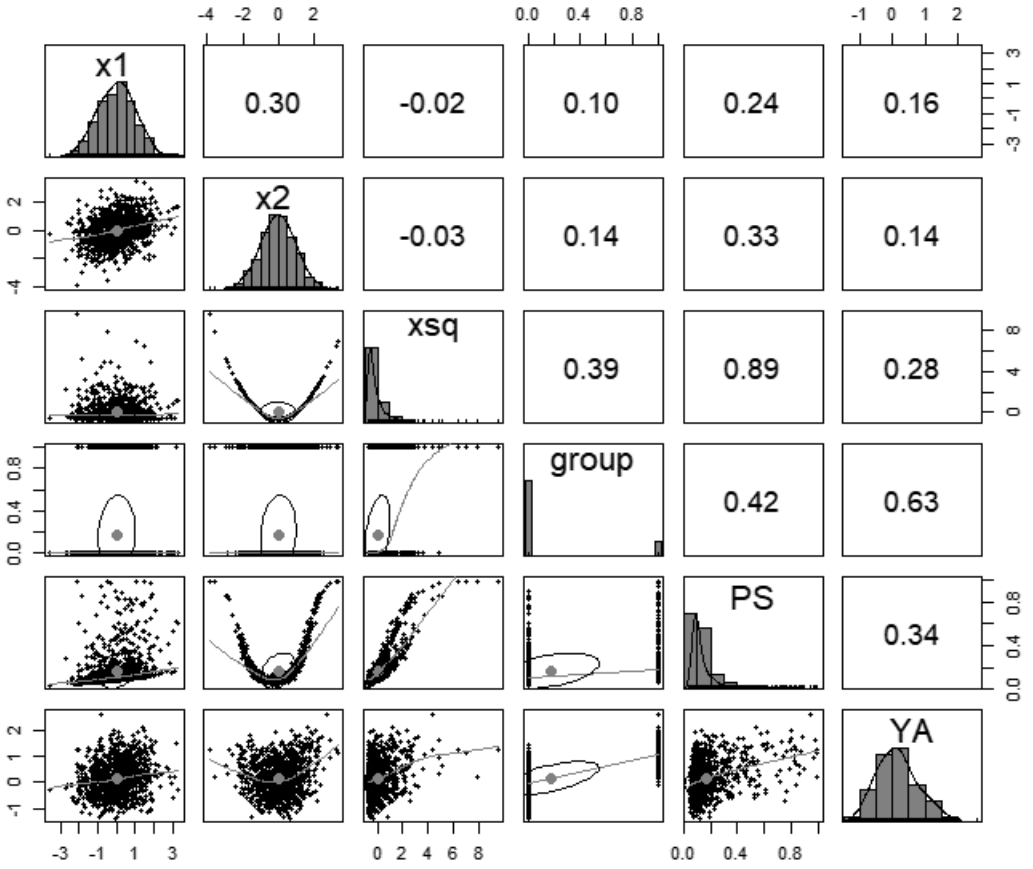
[10] The latent propensity distribution was linearly rescaled by subtracting the constant value of the intercept of the probit model. This intercept was calculated by taking the z-score of the standard normal distribution corresponding to .80 and dividing it by $\sqrt{(1 - R^2)}$ with Scenario A and B each having their own R^2. This R^2 was calculated by B'RB/(B'RB+1))

propensity model was simulated with a probit model, and final propensity for treatment was created with a logit model.

The continuous outcome was generated using the linear regression model, $Y_{Outcome} \sim 1\ (Y_{group}) + .05X_1 + .05X_2 + .05X_2^2 +$ v, where v represents random error in the model. The values of v were simulated randomly to follow a normal distribution with a mean of 0 and a standard deviation of 0.50. Figure 5 displays information from the validation sample (a single simulee sample of 1000), including the distributions of X_1, X_2, X_2^2, group membership, the true propensity score (PS), the outcome, and the correlations among variables. Figure 6 displays the relationship between X_2, X_2^2 and the propensity score in the logit metric.

Figure 5

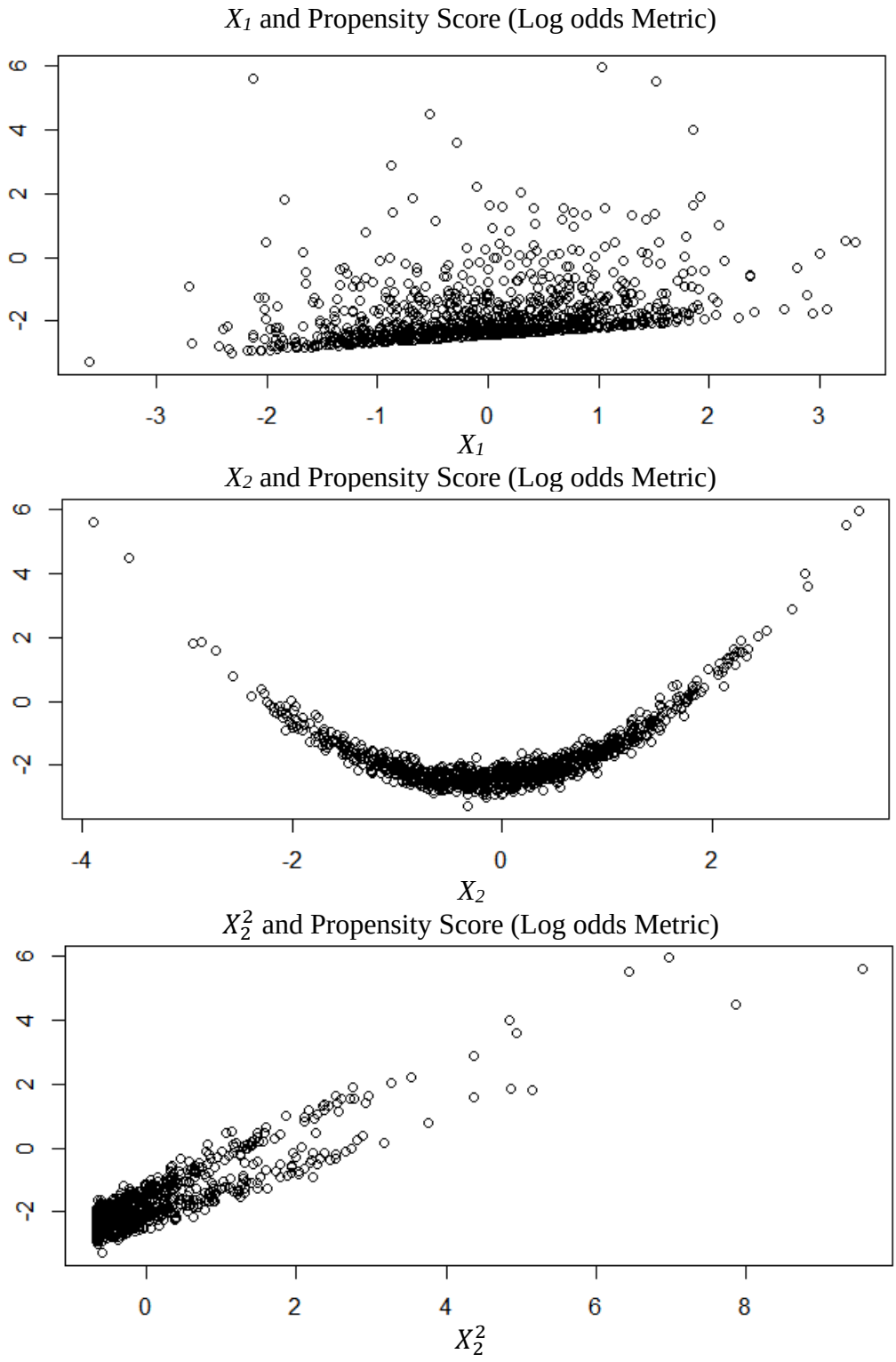
Scenario A's Correlation Matrices, Histograms, and Scatterplots



Note. Scenario A's Correlation matrices, histograms, and scatterplots of the simulated covariates, X_1, X_2, and polynomial term. In this figure, X_1 and X_2 are normally distributed, xsq represents X_2^2, group represents treatment group assignment, PS represents the "true" propensity scores (in probability metric), and YA represents the simulated outcome for Scenario A.

Figure 6

Relationships Between Propensity Scores and Covariates in Scenario A

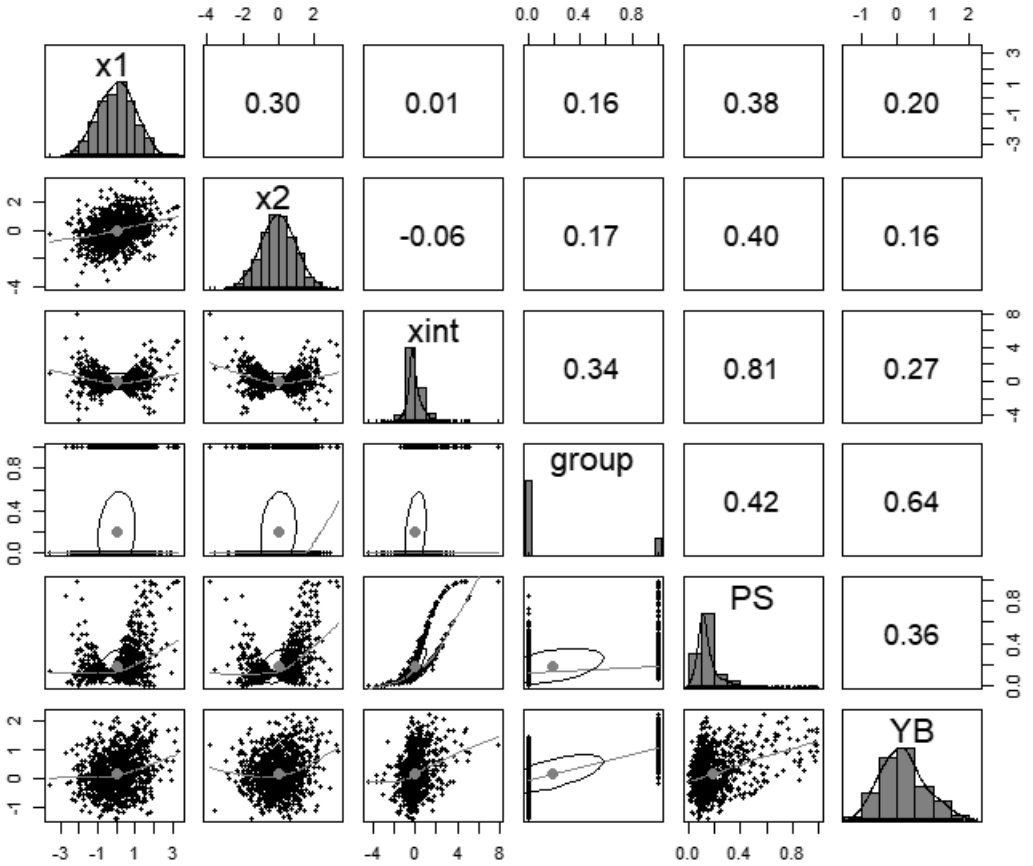


Scenario B. Treatment status was generated by first creating "true propensity scores," which were produced in the same manner described above, except that it included the interaction term, X_1X_2, rather than a quadratic term. Then, a random draw was assigned to each simulee, such that if the true propensity score was greater than the random draw, then the simulee was assigned to the treatment group (group = 1). Otherwise, a simulee would be assigned to the comparison group (group = 0). When assigning group membership, the treatment:comparison group ratio was fixed to be approximately 200:800 or 1:4. This was done by rescaling the latent propensity distribution, similar to Scenario A. The final propensity for treatment also correlated with the true propensity scores, $r = 0.998$.

The continuous outcome was generated using the linear regression model, $Y_{Outcome} \sim 1\ (Y_{group}) + .05X_1 + .05X_2 + .05\ X_1X_2 + v$, where v represents random error in the model. The values of v were simulated randomly to follow a normal distribution with a mean of 0 and a standard deviation of 0.50. Figure 7 displays the distributions of X_1, X_2, X_1X_2, group membership, the propensity score in probability metric (PS), and outcome from the validation sample, as well as the correlations among each variable. Figure 8 displays the relationship between X_2, X_1X_2 and the propensity score in the log odds metric.

Figure 7

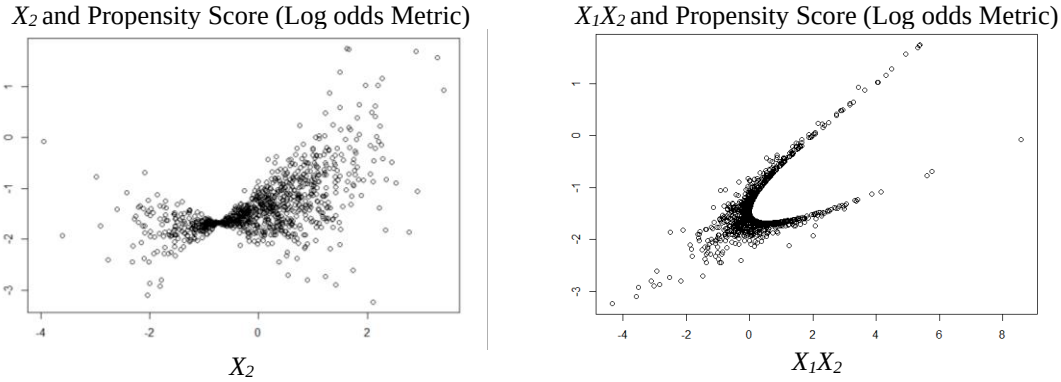
Scenario B's Correlation Matrices, Histograms, and Scatterplots



Note. Simulation B's Correlation matrices, histograms, and scatterplots of the simulated covariates, X_1, X_2, X_1X_2, and the interaction and polynomial terms. In this figure, X_1 and X_2 are normally distributed, xint represents the product (interaction) between X_1 and X_2 , group represents treatment group assignment, PS represents the "true" propensity scores (in probability metric), and YB represents the simulated outcome for Scenario B.

Figure 8

Relationships Between Propensity Scores and Covariates in Scenario B



Validation Data Sets

A validation data set was produced from both of the simulated scenarios. These datasets were used for visual balance diagnostics and to ensure the data were simulated correctly. To investigate whether the data were simulated correctly, I examined the number of simulees assigned to treatment and comparison groups, descriptive statistics for the relevant covariates, and the relationships among the variables (Table 2; Figures 5,7).

Table 2
Descriptive Statistics by Scenario and Group

	n	*M*	*SD*	Min	Max
Scenario A					
Group: 0					
X_1	830	-0.05	1.00	-3.61	3.32
X_2	830	-0.06	0.86	-2.87	2.52
X_2^2	830	-0.17	0.68	-0.67	4.87
PS	830	0.14	0.11	0.04	0.90
YA	830	-0.02	0.51	-1.37	1.42
Group: 1					
X1	170	0.23	0.95	-2.13	3.24
X2	170	0.31	1.48	-3.90	3.39
X_2^2	170	0.85	1.67	-0.67	9.54
PS	170	0.32	0.22	0.07	~1.00
YA	170	1.05	0.47	-0.09	2.62
Scenario B					
Group: 0					
X_1	813	-0.08	0.95	-3.61	2.93
X_2	813	-0.08	0.92	-3.56	2.91
X_1X_2	813	0.16	0.81	-4.40	3.65
PS	813	0.15	0.11	~0.00	0.85
YB	813	-0.02	0.50	-1.37	1.45
Group: 1					
X_1	187	0.33	1.13	-2.70	3.32
X_2	187	0.35	1.24	-3.90	3.39
X_1X_2	187	0.71	1.37	-1.44	7.86
PS	187	0.33	0.27	0.06	0.99
YB	187	1.04	0.49	-0.17	2.23

*Note. "*Group: 0" indicates the simulated comparison group, and "Group:1" indicates the simulated treatment group. PS indicates the "true" propensity score on the probability metric. YA and YB indicate the outcome variable in Scenario A, and Scenario B respectively. Notably, the maximum value of PS in Scenario A, group 1, appears to violate the assumption that propensity scores should not be equal to 1 or 0. The value of the maximum propensity score is less than one when it is not rounded to two decimal places (0.99739). Similarly, the minimum PS value in Scenario B, group 0, is larger than 0 when not rounded to 2 decimal points (0.00389). However, these values are still worth discussing in regard to violating the assumptions.

Propensity Score Matching

For the PSM conditions, propensity scores were estimated via the MatchIt package in R (Ho et al., 2011). The MatchIt package allowed for the use of logistic regression as an estimation method, and nearest neighbor matching using a 0.2 caliper width as a matching method. I chose NN matching with a caliper adjustment as it

generally produces more balanced matches than nearest neighbor matching without a caliper and performs on par with optimal matching when there is a large ratio of comparison group to treatment group individuals (Austin, 2011; Bai, 2011; Gu & Rosenbaum, 1993). Additionally, the use of NN with caliper matching reflects the methods used by Austin (2009).

Generalized Boosted Modeling

I used the Twang package in R (Ridgeway et al., 2015) to conduct generalized boosted modeling. For GBM, X_1 and X_2 were the only two variables included in the model to predict group membership, Y_{Group}. The polynomial and interaction terms were not included, as GBM should incorporate interactions and polynomials into the model if they are relevant to the prediction (McCaffrey et al., 2013).

Following the practices of Ridgeway et al. (2015), within each replication of the GBM analysis, I chose to produce 5000 trees, with an interaction depth of 2, a shrinkage value of .01, and an ATT estimand. Additionally, I determined the optimal iteration by minimizing the average standardized absolute mean difference (effect size), a method recommended by McCaffrey et al. (2004) and supported by the Twang package (Ridgeway et al., 2015). The Twang package (Ridgeway et al., 2015) was then used to pull out the weights produced by GBM, and those weights were incorporated into an outcome model using the survey package (Lumley, 2004, 2019).

Evaluating the Research Questions

After simulating data for the two scenarios, relevant information was saved, assessed, and then collapsed across replications. This information allowed me to assess balance and estimate treatment effect across the treatment conditions.

Group Balance. To assess balance, I employed both visual and numerical diagnostic techniques, as recommended in the literature (Austin, 2009). Although obtaining and assessing visual diagnostics for all 2000 replications (1000 for each scenario) is impractical, the validation data were used to produce jitter, density, and Q-Q plots (via the ggplot2 and MatchIt R package; Ho et al., 2011; Wickham, 2016). The majority of the balance diagnostics therefore relied heavily on numerical interpretations. Numerically, the variance ratios, standardized differences, and PBRs were examined and compared for X_1, X_2, X_2^2 (Scenario A) and X_1X_2 (Scenario B) when applicable. The jitter plots and standardized difference plots produced in the validation sample through PSM are included below (Figures 9-16). In the jitter plots (Figures 9, 11, 13, and 15), there is an appropriate amount of common support between the matched treatment and comparison (labeled control) groups; however, due to the caliper matching method, a handful of treatment units were left unmatched in each PSM condition. In each of the standardized differences plots (Figures 10, 12, 14, and 16) the propensity score balance (PS), X_1, X_2, showed unbalance before matching (value above 0.2) and balance after matching (value below .2). The same could be shown in conditions that examined X_2^2 (Figure 10) and X_1X_2 (Figure 14). It is worth noting that when the models in the validation sample included the third variable (X_2^2 or X_1X_2), the PS and third variable had a greater degree of unbalance before matching, that was corrected after matching.

Figure 9

Jitter Plot from Matching on a Correctly Specified Propensity Score Model in Scenario A

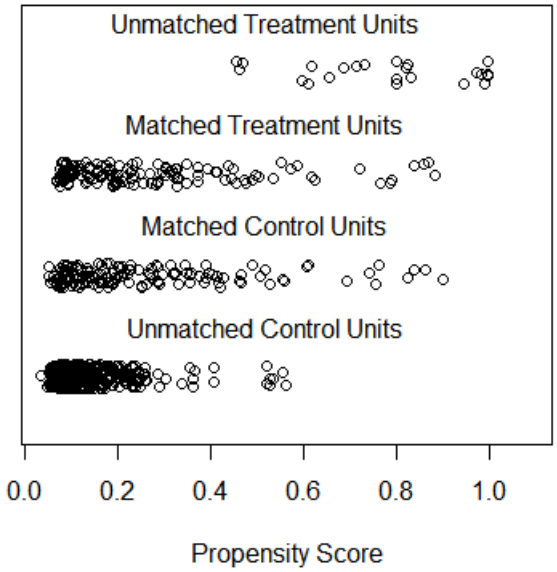


Figure 10

Standardized Differences after Matching on a Correctly Specified Propensity Score Model in Scenario A

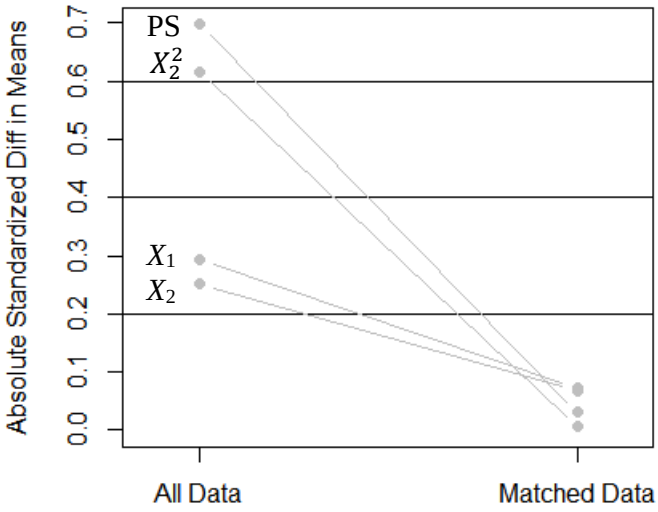


Figure 11

Jitter Plot from Matching on an Incorrectly Specified Propensity Score Model in Scenario A

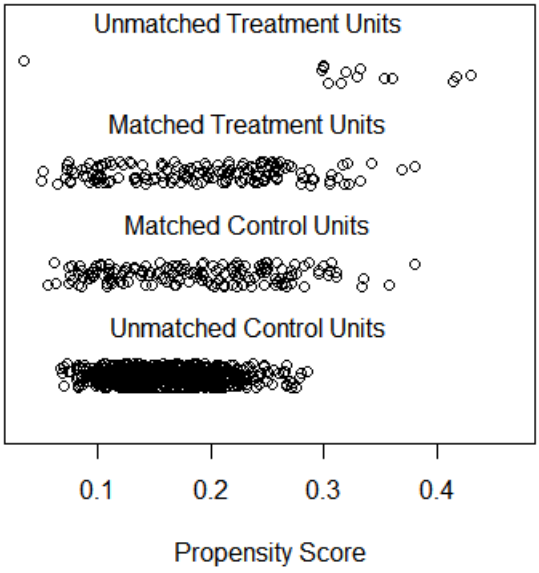


Figure 12

Standardized Differences after Matching on an Incorrectly Specified Propensity Score Model in Scenario A

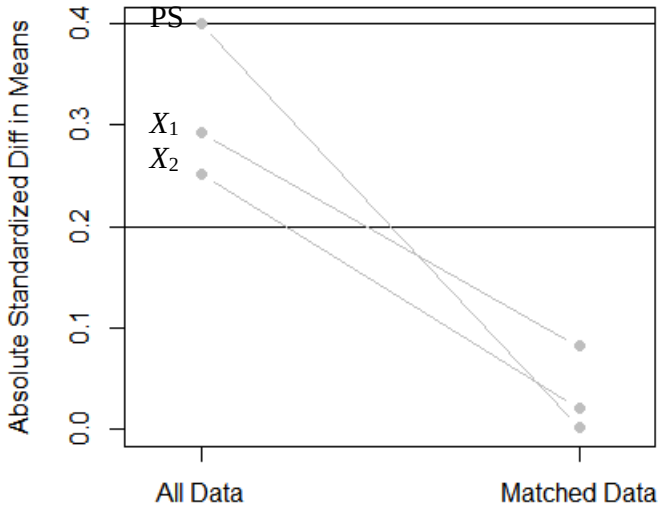


Figure 13

Jitter Plot from Matching on a Correctly Specified Propensity Score Model in Scenario B

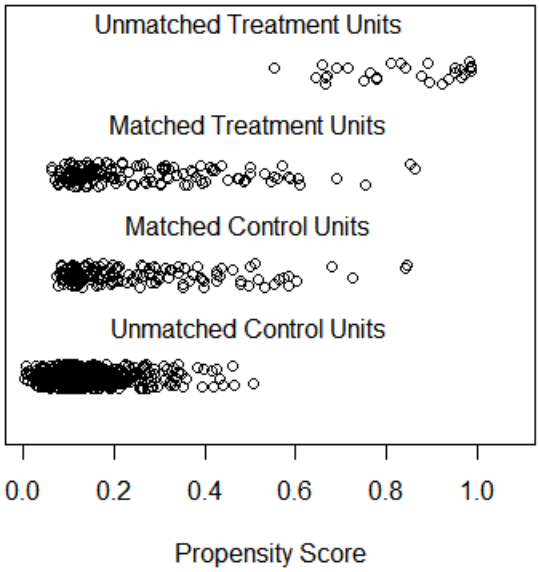


Figure 14

Standardized Differences after Matching on a Correctly Specified Propensity Score Model in Scenario B

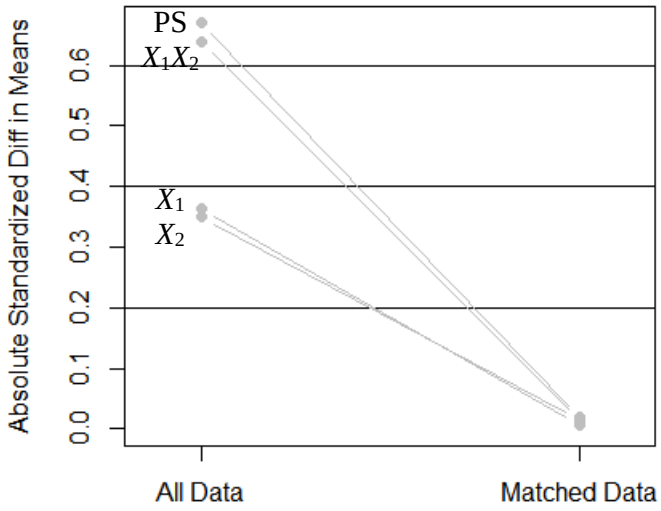


Figure 15

Jitter Plot from Matching on an Incorrectly Specified Propensity Score Model in Scenario B

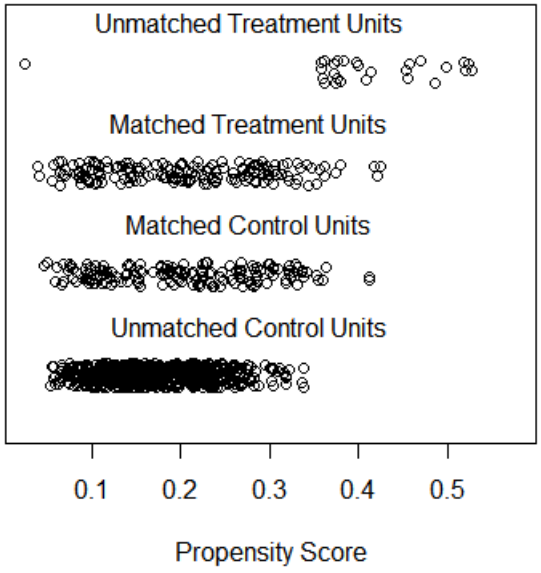
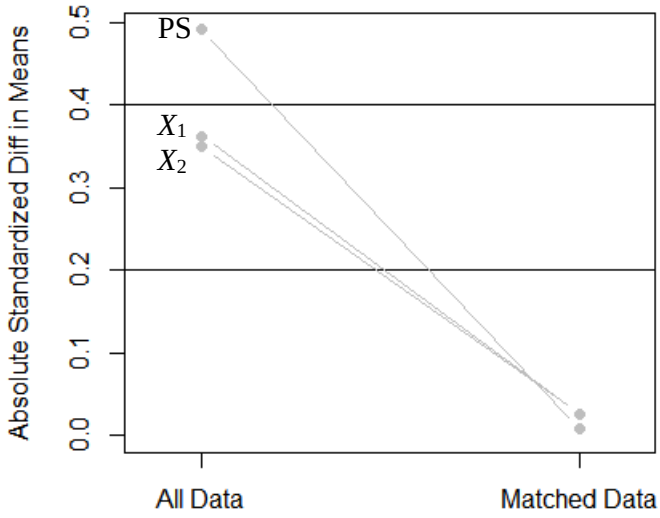


Figure 16

Standardized Differences after Matching on an Incorrectly Specified Propensity Score Model in Scenario B

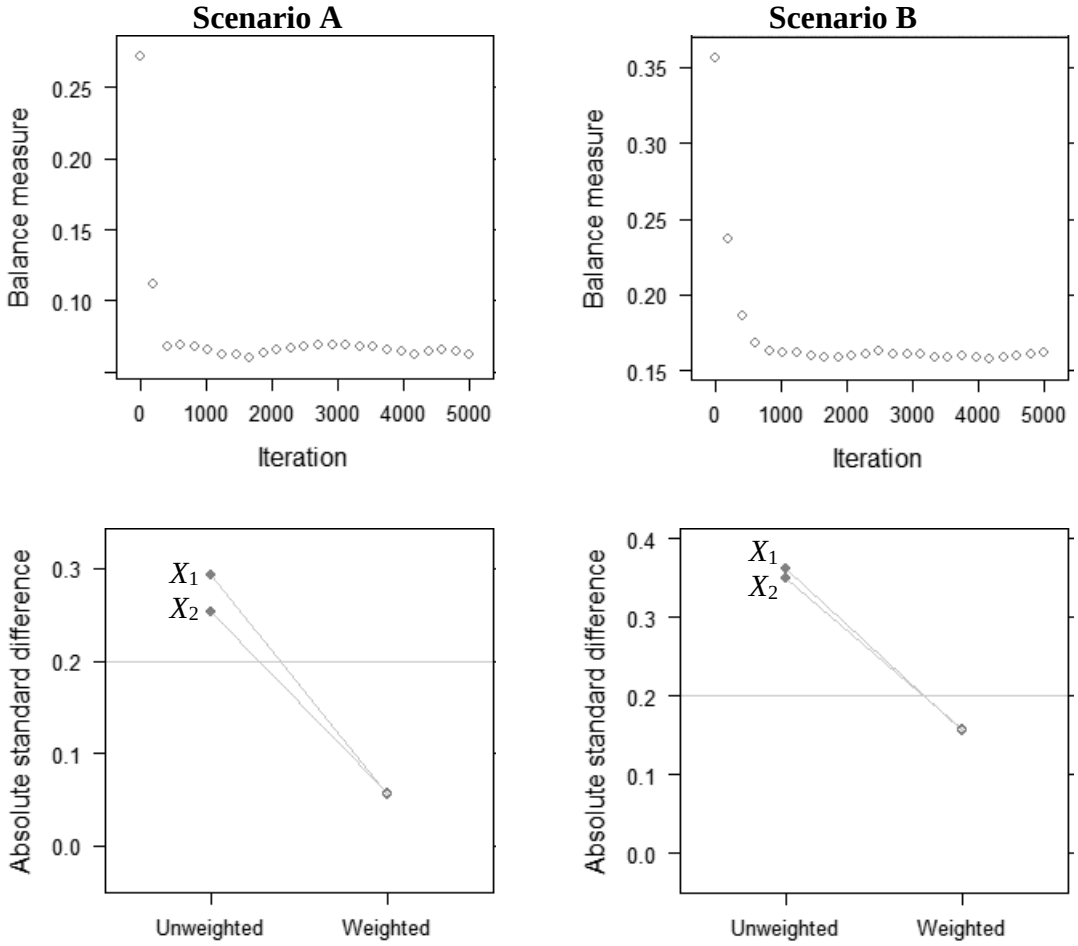


For GBM, the balance by iteration and standardized differences plots form the validation sample are all located in Figure 17. Only X_1 and X_2 were inputted into the GBM model in both scenarios. This is because, due to the nature of GBM, the Twang package specifies that, "there is no need to specify interaction terms in the formula"

(Ridgeway et al., 2017, p. 3) and McCaffrey et al. (2013) notes GBM's ability "to capture complex and nonlinear relationships between treatment assignment and the pretreatment covariates without over-fitting the data"(p. 3). Therefore, only the standardized differences for those two variables were included in the plot. Both figures showed improvement in balance from unbalanced (above .2) to balanced (below .2), but Scenario A reached an optimal value in fewer iterations (1772) than Scenario B (3962). This difference is notable in each of the balance by iteration plots, as an observable "dip" in the dots appears at each plots' optimal iteration. Table 3 further explores the balance in the validation samples, by displaying the percent balance reduction (PBR) for each variable included in each condition. When examining PBR, values closer to 100 indicate a greater reduction in unbalance (data are more balanced after matching or weighting) and values closer to -100 indicate an increase in unbalance (data are less balanced after matching or weighting).

Figure 17

Balance by Iteration for GBM Effect Size Stopping Rule & Standardized Differences



Note. The top two figures represent the balance measure by iteration of GBM. In the validation sample, the optimal iterations for Scenario A and B were the 1772 and 3962 iterations respectively. Absolute standardized differences from those iterations are displayed before and after weighting below.

Table 3
Percent Bias Reduction by Condition in Validation Sample

	PBR (%)
Scenario A – Correct PSM Model	
X_1	74.69
X_2	72.87
X_2^2	99.05
PS	95.64
Scenario A – Incorrect PSM Model	
X_1	71.93
X_2	91.77
PS	99.25
Scenario A – GBM	
X_1	81.00
X_2	77.27
Scenario B – Correct PSM Model	
X_1	98.05
X_2	95.15
X_1X_2	97.64
PS	97.00
Scenario B – Incorrect PSM Model	
X_1	92.64
X_2	92.45
PS	98.34
Scenario B – GBM	
X_1	56.69
X_2	48.16

Note. Percent balance reduction (PBR) is on a scale of -100 to 100, where negative values indicate that a worse balance was achieved (i.e., overcorrecting) after matching or weighting, and positive values indicate that a better balance was achieved after matching or weighting. Some researchers recommend an 80% criteria as sufficient reduction in bias (Cochran & Rubin, 1973; Pan & Bai, 2015).

Treatment Effect Estimation. Treatment effects were estimated for each replication of the final matched (or weighted) groups. I considered the mean difference in outcome between the treatment and comparison groups (i.e., coefficient for the grouping variable), to evaluate whether the simulated treatment effect was removed. Any difference between the average group coefficient in the outcomes model and the simulated group difference, 1, was considered to be residual bias. I compared this bias in the mean difference between the predicted and true outcome and examined the cell means

for each of the 8 conditions by conducting a 2x4 within-subjects ANOVA, considering effect size over statistical significance, due to the large sample size.

Chapter Four

Results

Sample Size

Before Matching. Although all simulated samples had a sample size of 1000, the average baseline treatment sample sizes were smaller than the goal size. The goal treatment:comparison group ratio was about 1:4 so the average treatment group sample should have had around 200 people. Instead, the treatment group sample size averaged between 161.78 and 191.23, or around a 1:4.6 ratio (Table 4). This discrepancy is permissible, as the change in ratio benefits the bias reduction, but only by a negligible amount (Rubin, 1979).

After Matching. Because the PSM models used NN matching with a .2 caliper, all PSM models tended to lose treatment group simulees who could not be matched. Particularly, the matched samples appeared to lose more treatment group simulees in conditions where the propensity score model was correctly specified than in conditions where the model was incorrectly specified (Table 5). This sample loss is explained when looking at the validation jitter plots (Figures 9, 11, 13, & 15); the correctly specified PSM model better explains group differences, so groups are further apart than the incorrectly specified PSM model. Thus, more simulees would reasonably have propensity scores that were greater than the .2 caliper apart from each other. Because GBM uses weighting rather than matching with a caliper, no treatment group members were dropped.

Table 4
Treatment Sample Sizes

	M	*SD*	Min	Max
Scenario A				
Baseline $n_{Treatment}$	189.24	11.03	158	223
Correctly Specified PSM Matched $n_{Treatment}$	161.78	11.06	131	195
Incorrectly Specified PSM Matched $n_{Treatment}$	176.06	12.54	136	217
Scenario B				
Baseline $n_{Treatment}$	191.23	11.23	153	226
Correctly Specified PSM Matched $n_{Treatment}$	165.57	11.23	128	202
Incorrectly Specified PSM Matched $n_{Treatment}$	172.96	11.54	135	209

Note. Baseline $n_{Treatment}$ refers to the number of simulees in the treatment condition before matching. Due to the nature of weighting, the GBM conditions would have kept everyone in the treatment group, so those conditions would have equivalent sample sizes as the baseline conditions.

Table 5
Treatment Sample Loss After Matching

	Mean Loss	% Loss	*SD*
Scenario A – Correct PSM Model	27.46	14.51	6.37
Scenario A – Incorrect PSM Model	13.19	6.97	6.75
Scenario B – Correct PSM Model	25.66	13.42	6.41
Scenario B – Incorrect PSM Model	18.27	9.55	5.84

Note. Mean loss was calculated by subtracting the matched treatment sample size from the respective baseline/unmatched treatment sample size for that scenario. Percent loss was calculated by dividing the mean loss by the respective scenario's baseline treatment group size (189.24 for Scenario A, 191.23 for Scenario B; Table 4) and multiplying that number by 100.

Examining Balance Between Models

To numerically examine the balance across the various conditions, I considered PBR and standardized mean differences. Additionally, I considered the variance ratio of the propensity scores for the PSM conditions in order to evaluate the width of the distribution of propensity scores and whether it was similar across the treatment and comparison groups.

Percent Bias Reduction. Although the mean PBR for all conditions appeared to be similar, the PBRs for propensity scores tended to have the lowest standard error (Table 6). Notably, all models except the Scenario B GBM model overcorrected some of the covariates on at least one occasion (denoted by negative minimum PBR values, which indicate worse group balance after matching/weighting). However, the Scenario B GBM model also did not have ideal PBR values (i.e., PBR values greater than 80%; Cochran & Rubin, 1973; Pan & Bai, 2015).

Standardized Mean Difference. Although the Twang and MatchIt R packages both calculate standardized mean differences, they use a different standardizer in their formulas. I chose to default to the formula used by the MatchIt package, which used the comparison group standard deviation as a standardizer, rather than the treatment group standard deviation. Thus, the chosen formula for standardized mean difference was:

$$SMD = \frac{\mu_{X|Treatment} - \mu_{X|comparison}}{s_{X|comparison}} \quad (11)$$

This value can be interpreted such that values close to zero indicate better balance among the covariates and propensity score than values further away from zero. This study's mean standardized differences after matching/weighting remained close to zero across all conditions (Table 7). Although discrepancies between calculations of SMDs and their benchmarks exist, I chose to still compare these calculations of SMDs to the .1 benchmark endorsed by Austin (2009, 2011) for the Cohen's *d* method of calculating SMD. Using this benchmark, there was adequate balance across conditions when evaluated using SMD, regardless of model or correct specification.

Variance Ratios. To evaluate the width of the propensity score distributions, the mean variance ratios of the propensity score should be close to 1, with a standard

deviation that is less than .50 (Rubin, 2001). With this criterion in mind, all PSM conditions except the Scenario A incorrectly specified condition appeared to have similar variability on the propensity score (Table 8). Not only did the Scenario A incorrectly specified condition average a variance ratio nearly double the recommended value of one, but the standard deviation was greater than the recommended .50.

Table 6
Percent Bias Reduction by Condition

	M	*SE*	Min	Max
Scenario A – Correct PSM Model				
X_1	82.73	15.04	-33.37	99.95
X_2	84.92	13.56	-2.12	99.95
X_2^2	95.13	3.22	82.10	100.00
PS	96.14	0.89	92.93	99.16
Scenario A – Incorrect PSM Model				
X_1	87.74	11.78	-15.20	100.00
X_2	89.11	10.46	13.13	99.97
PS	95.73	1.72	88.79	99.85
Scenario A – GBM				
X_1	84.23	11.73	1.09	100.00
X_2	75.96	19.11	-13.64	100.00
Scenario B – Correct PSM Model				
X_1	85.47	15.04	-180.47	100.00
X_2	87.31	10.42	34.28	100.00
X_1X_2	95.58	3.14	82.09	100.00
PS	96.22	0.94	93.33	98.82
Scenario B – Incorrect PSM Model				
X_1	89.17	13.60	-210.93	100.00
X_2	91.63	8.99	-10.75	100.00
PS	95.77	1.63	89.51	99.96
Scenario B – GBM				
X_1	78.00	13.37	22.17	100.00
X_2	79.10	12.28	25.66	100.00

Note. The above values are in percent metric. The upper bound of PBRs are constrained to a maximum value of 100.

Table 7
Standardized Mean Differences by Condition Before and After Matching/Weighting

	M_{Before}	M_{After}	SE_{After}	Min_{After}	Max_{After}
Scenario A – Correct PSM Model					
X_1	21.56	0.01	0.07	-0.25	0.21
X_2	20.10	0.01	0.05	-0.14	0.14
X_2^2	58.41	0.03	0.02	-0.01	0.01
PS	-	0.03	0.01	0.01	0.05
Scenario A – Incorrect PSM Model					
X_1	21.56	-0.02	0.05	-0.17	0.14
X_2	20.10	0.02	0.03	-0.07	0.13
PS	-	0.02	0.01	0.00	0.04
Scenario A – GBM					
X_1	21.56	0.05	0.04	-0.04	0.26
X_2	20.10	0.07	0.05	-0.12	0.25
Scenario B – Correct PSM Model					
X_1	22.10	0.02	0.05	-0.09	0.09
X_2	24.41	0.02	0.05	-0.17	0.18
X_1X_2	54.02	0.02	0.03	-0.06	0.11
PS	-	0.03	0.01	0.01	0.05
Scenario B – Incorrect PSM Model					
X_1	22.10	0.01	0.04	-0.10	0.15
X_2	24.41	0.01	0.04	-0.11	0.16
PS	-	0.02	0.01	-0.01	0.05
Scenario B – GBM					
X_1	22.10	0.07	0.05	-0.02	0.25
X_2	24.41	0.08	0.05	-0.03	0.25

Note. Standardized mean differences of 0 indicate more similarity between the matched (or weighted) comparison group and treatment group. Positive values indicate that the treatment group mean was higher than the comparison group mean, while negative values indicate the comparison group mean was higher than the treatment group mean. Propensity scores before matching/weighting were not saved out.

Table 8
Variance Ratios for Propensity Scores

	M	*SD*	Min	Max
Scenario A – Correct PSM Model	1.05	0.02	0.99	1.12
Scenario A – Incorrect PSM Model	2.16	0.72	1.06	6.13
Scenario B – Correct PSM Model	1.05	0.02	1.00	1.11
Scenario B – Incorrect PSM Model	1.30	0.21	0.87	3.87

Note. The above variance ratios refer to the variance of the treatment group's propensity scores, over the variance of the comparison group's propensity scores (after matching). To evaluate the width of the propensity score distributions, the mean variance ratios of the propensity score should be close to 1, with a standard deviation that is less than .50 (Rubin, 2001).

Treatment Effect Estimation

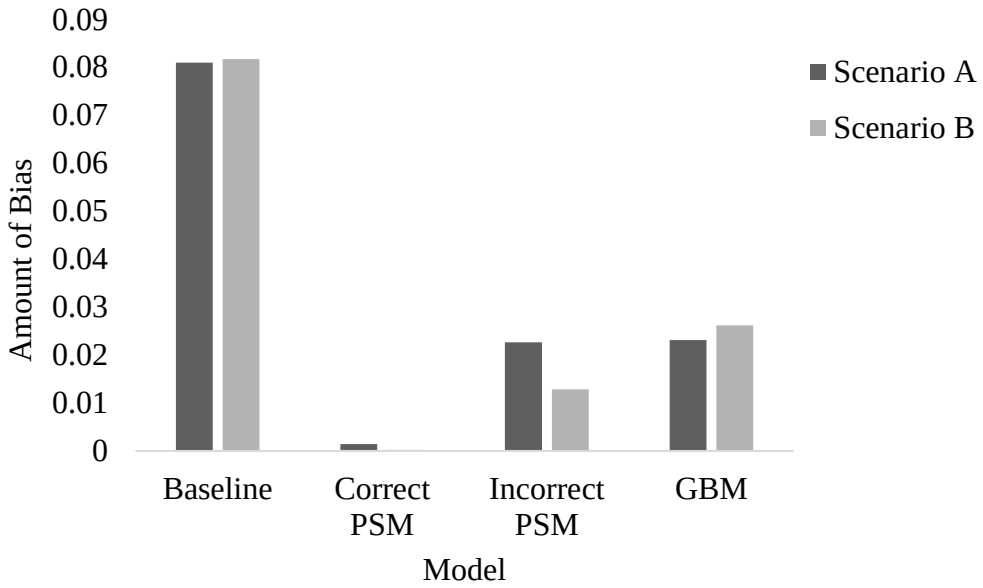
As there appeared to be adequate balance achieved by each of the propensity score methods, I proceeded on to the treatment effect estimation. To estimate how much each model reduced selection bias in the treatment estimate, I examined the regression model, which used either the matched or weighted sample to predict the outcome variable, $Y_{outcome}$, from treatment group membership. Particularly, I examined the coefficient that accompanied the group variable in the outcome regression model, to see whether the coefficient would be equal to one, the simulated group difference on the outcome variable. Table 9 provides descriptive statistics for the group coefficients across conditions. Amount of bias in the model could then be considered as the group coefficient minus one and is illustrated by condition in Figure 18.

Table 9
Group Coefficients by Model

	M	*SE*
Scenario A		
Baseline Model	1.081	0.041
Correct PSM Model	1.001	0.053
Incorrect PSM Model	1.023	0.054
GBM	1.023	0.049
Scenario B		
Baseline Model	1.082	0.041
Correct PSM Model	1.000	0.053
Incorrect PSM Model	1.013	0.053
GBM	1.026	0.047

Note. The means represent the mean regression coefficient for group when predicting the outcome from group membership across the 1000 simulations. The baseline model predicted the outcome from group membership before any matching or weighting was conducted, $Y'_{outcome} = b_0 + b_1 x_{group}$. The true group difference was simulated to be 1.

Figure 18
Average Amount of Bias by Condition



Note. Bias is a function of the respective models' group coefficient (after matching or weighting; Table 9) subtracting the true group difference, one.

To determine whether the difference in bias across conditions was statistically and practically significant, I conducted a 2x4 within-subjects ANOVA on the group coefficients, with Factor 1 consisting of the two scenarios (i.e., A and B) and Factor 2

consisting of the four approaches/models (i.e., correctly specified PSM, incorrectly specified PSM, GBM, and a baseline model with no alterations to the sample). A within-subject ANOVA suited this study better than between-subjects ANOVA because conditions within each replication were simulated in a way that made them dependent. Both scenarios were created from the same initial baseline covariates (i.e., X_1 and X_2) for each replication. I used those baseline covariates to create a third variable (X_2^2 or X_1X_2) and outcome (Y) for each scenario. Thus, all models/conditions used the same X_1 and X_2, and all models/conditions within a scenario were influenced by the same third variable (X_2^2 or X_1X_2) and outcome (Y).

Before running the ANOVA, I checked Mauchly's Test to evaluate the sphericity assumption for factorial within-subjects ANOVAs. Mauchly's test was significant for both Factor 2 [$W = .870$, $X^2(5) = 138.68$, $p < .001$] and the interaction between the factors [$W = .523$, $X^2(5) = 647.44$, $p < .001$]. The Greenhouse-Geisser Epsilon for both Factor 2 and the interaction between factors were $\varepsilon = .93$, and $\varepsilon = .76$, respectively. Due to the large sample size of this study, the Greenhouse-Geisser correction was chosen over the Huynh-Feldt due to its conservative nature.

Using the Greenhouse-Geisser Correction, the omnibus test for the Factor 1 main effect (Scenario) was not significant, but the test for the Factor 2 main effect and the interaction between factors was significant (Table 10). Based on the value of partial η^2 for the interaction, however, the interaction effect is likely not a practically meaningful contributor to bias due to its small effect size ($\eta^2 > .01$; Cohen, 1988), despite being statistically significant. Therefore, both the main effect of Factor 2 and the interaction

will be separately considered in the results, rather than picking one over the other. Both will be discussed further in the discussion section.

Table 10
Omnibus Tests of Within-Subjects Effects

	df	*F*	*p*	η^2	1 - β
Factor 1: Scenario	1, 999	2.403	.121	.002	.341
Factor 2: Model	2.785, 2782.042	2911.695	<.001	.745	~1.000
Factor 1 x Factor 2	2.291, 2289.182	25.111	<.001	.025	~1.000

Note. Used a Greenhouse-Geisser Correction.

Exploring the Interaction. To explore the interaction, I examined the simple effects, as recommended by Maxwell and Delaney (2004). To do this, I conducted a one-way within-subjects ANOVA for each scenario separately, and found significant differences in average bias across the conditions within Scenario A [$F(3.50, 2545.23) = 1486.12$, $p < .001$, $\eta^2 = .60$], and Scenario B [$F(3.87, 2645.42) = 1887.89$, $p < .001$, $\eta^2 = .89$] using the Greenhouse-Geisser correction ($\varepsilon_a = .85$, $\varepsilon_b = .88$, respectively). The Bonferroni adjusted pairwise comparisons conditional on Scenario are located in Table 11. Of interest, all comparisons were statistically significant except the comparison between the incorrectly specified PSM and the GBM models in Scenario A. In both scenarios, the baseline model consistently included more bias than all other models, and the correctly specified model consistently included less bias than all other models. In Scenario A, there was no significant difference in bias between GBM and the incorrectly specified model ($p = \sim1$). However, in Scenario B, there was a significant difference in bias, such that the incorrectly specified PSM model had less bias than the GBM model ($p < .001$).

Table 11
Pairwise Comparisons within Each Scenario

Model Comparison (I X J)	$M_{Diff.}$ (I – J)	*SE*	*p*	95% CI LB	95% CI UB
Scenario A - Quadratic					
Baseline X Correct	.080	.001	<.001	.076	.083
Baseline X Incorrect	.058	.001	<.001	.055	.061
Baseline X GBM	.058	.001	<.001	.056	.060
Correct X Incorrect	-.021	.002	<.001	-.025	-.017
Correct X GBM	-.022	.001	<.001	-.025	-.018
Incorrect X GBM	.000	.001	~1.00	-.004	.003
Scenario B - Interaction					
Baseline X Correct	.082	.001	<.001	.078	.085
Baseline X Incorrect	.069	.001	<.001	.066	.072
Baseline X GBM	.056	.001	<.001	.053	.058
Correct X Incorrect	-.013	.001	<.001	-.016	-.009
Correct X GBM	-.026	.001	<.001	-.029	-.023
Incorrect X GBM	-.013	.001	<.001	-.017	-.010

Note. Pairwise comparisons used a Bonferroni correction. LB and UB represent the lower bound and upper bound, respectively.

Main Effect of Model. If focusing on the main effect of model rather than the interaction, then each of the models are significantly different from the other (Table 12), such that the correctly specified PSM model had significantly lower bias than the rest of the models, and the baseline model had significantly higher bias than the rest of the models. The bias for the incorrectly specified PSM model averaged lower than the GBM model.

Table 12
Factor 2 Main Effect Model Pairwise Comparison Collapsed Across Factor 1

Model Comparison (I x J)	M_{Diff} (I – J)	*SE*	*p*	95% CI LB	95% CI UB
Baseline X Correct	.081	.001	<.001	.078	.083
Baseline X Incorrect	.064	.001	<.001	.061	.066
Baseline X GBM	.057	.001	<.001	.055	.059
Correct X Incorrect	-.017	.001	<.001	-.020	-.014
Correct X GBM	-.024	.001	<.001	-.026	-.021
Incorrect X GBM	-.007	.001	<.001	-.009	-.004

Note. Model comparisons used a Bonferroni correction. LB and UB represent the lower bound and upper bound, respectively.

Chapter Five

Discussion

The goal of this study was to compare PSM and GBM in their ability to create balanced groups and reduce treatment effect bias. This study considered the impact of selection bias that includes interactions or quadratic terms, as well as the impact of a common human error – omitting interactions and quadratic terms in a PSM model. To compare PSM and GBM-based methods, a simulation study was done, so that the objective "truth" could be compared to the outcomes of the models. In this section, I will briefly discuss findings from the balance metrics used, then discuss the reduction of bias found across conditions, the limitations of the study, and the recommendations for researchers moving forward.

Balance Diagnostics

As the propensity score is a balancing score (Austin, 2009), the rationale for employing PSM and GBM-based weighting is to create balanced treatment and comparison groups. Therefore, an important first step in comparing methods is to evaluate the balance achieved after matching or weighting. Per Austin's (2009) recommendations, multiple methods were used to numerically assess balance, including PBR, SMD, and variance ratios.

The average PBRs for each condition were above 75% for each covariate – indicating a decent improvement in balance on the whole. However, every condition except the GBM for Scenario B overcorrected the balance at some point – denoted by the presence of a negative value in the minimum column of Table 6. This suggests that in some replications, the group difference between the treatment and comparison group

were overcorrected in the opposite direction (e.g., if treatment group mean was greater than the comparison group mean before matching, then treatment group mean may be lower than the comparison group mean after matching). This is an important aspect to consider, as overcorrections such as this can bias the treatment effect more, rather than less. This is not to say that the Scenario B GBM model was without its flaws, as the PBRs tended to average below the recommended value of 80% for both X_1 and X_2. Because the average PBR for X_2 in the Scenario A GBM model was also below 80%, there may be a relationship between variable(s) involved in the creation of the third variable (i.e., Scenario A's X_2^2, and Scenario B's X_1X_2) and low PBR values in GBM.

The SMD examined whether the distributional centers (i.e., means) of the propensity scores and covariates were aligned in the treatment and comparison groups after adjustment (i.e., matching or weighting). The standardized difference in means were all between .01 and .08, when a value of 0 suggests no difference between means (i.e., balance in the distributions). Although all models exhibited good balance, the GBM models had the highest SMDs, as all the PSM models had lower SMDs that ranged between .01 and .03.

After considering whether the means were aligned, I evaluated the width of the propensity score distributions by considering the variance ratios. On Table 8, it is evident that the incorrect PSM model for Scenario A deviates from the other PSM models. Additionally, both incorrectly specified PSM models have a larger range of variance ratios than their correctly specified counterparts, and thus, a larger standard error.

Considering the numerical balance metrics above and the visual balance metrics displayed in Chapter 3 (Figures 9-17), it appears as though PSM and GBM both resulted

in improved balance over baseline scenarios. Although it cannot be said that one model would consistently achieve more balance in other situations, in this simulation the correct PSM models had more stability in the variance ratios than the incorrect PSM models (evidenced by Table 8). Additionally, GBM consistently had lower PBR averages and higher SMD after weighting (Table 6) than the PSM counterparts. Therefore, based upon the limited conditions of the current study, the correctly specified PSM model achieved the best balance – but it is worth noting that a correctly specified model was still prone to occasional overcorrections.

Treatment Effect Estimation

After adequate balance has been confirmed for propensity score methods, then one can evaluate the treatment effect estimate. I used an ANOVA to examine how the propensity score methods and the baseline models compared in the average difference found between the treatment and comparison group in the outcomes model. Although one could either favor the main effects or interaction interpretation of the ANOVA, the correctly specified PSM model reduced the most bias, but all models reduced a significant amount of bias from the baseline model.

ANOVA Interpretation. When examining results from a simulated study, it is important to consider the impact of sample size on frequentist tests of statistical significance. This study used 1000 replications of 1000 subjects, so the results may be prone towards Type 1 error (finding significance when it does not exist). Because of this, I used conservative adjustments (e.g., Greenhouse-Geisser and Bonferroni adjustments). Additionally, I evaluated effect sizes to differentiate between statistical and practical significance. I used partial eta-squared (η^2) for my effect size, considering Cohen's

(1988) benchmarks for effect size, where .01 indicates a small effect, .06 indicates a medium effect, and .14 indicates a large effect.

The interaction between the Scenario and the Model was statistically significant but had a small effect size ($\eta^2 = .025$). The main effect for model was also statistically significant, but with a much larger effect size than the interaction ($\eta^2 = .745$). This evokes the question of what interpretation of the within-subjects ANOVA is most relevant and meaningful. On one hand, it seems misguided to ignore a significant interaction. On the other hand, perhaps the main effect interpretation is more meaningful and practical for real-world applications, as the effect size is very large, and the interpretation is more intuitive. I favor of the main effect interpretation, but I will interpret both below to be thorough.

Interaction Interpretation. An examination of the interaction via the simple effects (Table 11) suggests that a correctly specified PSM model reduces bias the most and produces a treatment effect estimate that is closest to one (i.e., the population treatment effect). Additionally, a baseline model with no matching or weighting consistently has the most bias, and a treatment effect estimate furthest from one. The source of the significant interaction appears to be the comparison of the incorrectly specified PSM model and the GBM model across the scenarios. In the presence of a quadratic relationship (Scenario A), both the incorrectly specified and the GBM models reduced the same amount of bias. In the presence of an interaction (Scenario B), the incorrectly specified PSM model reduced bias more than the GBM model. Of interest, the incorrectly specified model in Scenario A also had the most extreme variance ratio, indicating an extreme difference in the distribution of the propensity score. The

implications of this in the interaction cannot be fully explored, because there were no calculations of variance ratio for the GBM conditions. However, as mentioned previously, GBM methods can be compared using other methods of balance diagnostics, such as PBRs. This comparison revealed a trend of unbalance among predictors involved in the creation of the third variable, such that X_2 was slightly unbalanced in Scenario A, and both X_1 and X_2 were slightly unbalanced in Scenario B (denoted by average PBR values below 80%; Cochran & Rubin, 1973; Pan & Bai, 2015).

Model Main Effect. Although the interaction had a small effect, the interpretation of the main effect is more practical, as there is an incredibly large effect size – so the statistical significance cannot be entirely attributed to the large sample. Ultimately, the interpretation of the main effect is similar to the interaction interpretation in that the correctly specified PSM model is the best at reducing bias, but all models perform better than baseline. While the interaction differentiates the utility of incorrectly specified PSM models in the scenario with a quadratic term, an examination of the main effect suggests that incorrectly specified PSM models perform better than GBM across scenarios (but only slightly, $M_{Diff} = -.007$).

Limitations

Limitations in design and execution point to opportunities for future research to elaborate on the methods in this study. The design of the study was limited in that only one matching method represented PSM, nearest neighbor with a caliper of .20. Because of this decision, the results of the PSM models cannot be extrapolated to other matching methods (e.g., nearest neighbor without caliper, optimal, genetic). Additionally, I cannot wholly separate the results of the PSM models from the influence of sample size loss due

to the stringent requirements imposed by the caliper (i.e., matched pairs had to be within .2 SD of each other). On average, the treatment group lost between 13.19 and 27.46 simulees, and given that the treatment group sample size was often less than 200, a considerable portion of that treatment group was lost. Although dropping some treatment group members assisted in creating balanced groups, losing group members risks changing the composition of the treatment group to something no longer reflective of the intended population. Thus, significant loss in treatment group members may bias treatment effect estimates and decrease power for detecting that treatment effect (Stone & Tang, 2013).

The treatment sample size is also a limitation of the study. Although I simulated the data with the intention of a 200-800 split, the treatment group sample size averaged below 200 (Table 4). This may be attributable to the linear rescaling of the latent propensity distribution. Additionally, it is worth noting that the latent propensity scores were created with a probit model, rather than a logit model. Thus, the distribution of the simulated latent propensity scores could not perfectly be estimated by the logit models used for the research question.

Another limitation of the simulation may be the amount of bias simulated. While the treatment effect is comparable to previous studies (Austin, 2009), it is unknown whether this amount of bias adequately reflects the amount of selection bias present in applied samples, or in what circumstances this amount of bias is concerning or not. Additionally, this study simplifies selection bias as a result of two covariates, when in reality selection bias can be complex and multidimensional.

Recommendations

Future Studies on Quasi-Experimental Techniques. Researchers hoping to elaborate on this topic should consider simulation studies that involve more covariates, and experiment with the magnitude of the covariates' relationships to each other, the true propensity score, and the outcome variable. Additionally, as mentioned in the limitations, I only used one matching method, rather than comparing the different methods that existed. Future research should consider adding additional matching methods, such as the well-performing optimal matching, or the commonly used NN without a caliper. Additionally, researchers could further explore the effects of different sized calipers, to better examine the tradeoff made between holding a strict caliper and maintaining the treatment group size. It would also be informative to see if there was additional bias in the treatment effect if not held to a strict caliper. Perhaps with a different caliper, or no caliper, there would have been a more definitive difference found between the performance of the incorrectly specified PSM model and the GBM model in how much the treatment effect bias was ultimately reduced.

Additionally, I used GBM without specifying an interaction or quadratic term, to test the claims that such relationships would still be included by the nature of the GBM processes (McCaffrey et al., 2013, Ridgeway et al., 2017). However, perhaps GBM may not fully capture such relationships unless they are more explicitly specified into the model. It could be interesting to compare how GBM would have performed when the quadradic or interaction term were explicitly specified, compared to the implicit specification from including X_1 and X_2.

Future Studies Using Quasi-Experimental Techniques. Before conducting quasi-experimental studies in which selection bias is bound to be present, researchers should first carefully examine the literature for what covariates may be related to selection bias. By doing so, researchers can arrange to measure all covariates theoretically related to selection bias. This way, researchers can assure a correctly specified model, which is an assumption that underlies any statistical method. One aspect worth noting is that the incorrectly specified PSM model achieved adequate balance, despite the obvious model misspecification. This supports the claim mentioned earlier in the literature review that, "balance may be necessary, but it is not sufficient for strong ignorability to be met" (Shadish, 2013; p. 134). Therefore, I would further caution applied researchers that balance diagnostics should not be used as an indicator for correct model specification – as it only provides information on what the researcher has chosen to include.

Once a researcher has collected data, then they can then examine the data closely for interactions and exponentiation before making decisions about a model. It is worth noting that checking the collected data for interactions and exponentiation cannot make up for model misspecification caused by a researcher never having measured an important covariate.

Once a researcher believes they have discerned important covariates, interactions, and exponentiation they can chose whether to use GBM or PSM. In this study, the correctly specified PSM model with NN matching using a .2 caliper produced the best reduction in bias; therefore, this approach is recommended for situations with interactions and exponentiation. However, GBM and the incorrectly specified PSM model still

produced a meaningful reduction in bias (although it is worth noting that these models included all covariates contributing to selection bias and no spurious or otherwise misleading covariates).

Applied researchers should consider running multiple analyses and reporting and comparing each in the context of the study. By doing so, the applications of this study can be better examined, and the usefulness of each technique may be evaluated in real-world contexts that have more nuance in selection bias and its effects. This suggestion is echoed by several other researchers in the literature, such as Austin (2011), who recommends an iterative approach to model building to achieve better balance in a sample.

Conclusion

By comparing statistical approaches for approximating the counterfactual such as PSM and GBM, these results should help inform researchers about best practices when making causal claims in the absence of random assignment. This study found that a correctly specified PSM model reduced selection bias better than an incorrectly specified PSM model or GBM – both in scenarios with quadradic terms and interactions. Therefore, with careful research and consideration of covariate relationships, a correctly specified PSM model provides the closest approximation to the treatment group's counterfactual. Although in applied research, it is immensely difficult to perfectly specify a model for selection bias, the performance of GBM and the incorrectly specified PSM model provide encouragement that even an omission of a higher-order term can still lead to bias reduction in the estimation of the outcome. However, nothing performs as well as a correctly specified model.

References

Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *The Stata Journal*, *4*(3), 290-311.

Alon, S. (2010). Racial differences in test preparation strategies: A commentary on shadow education, American style: Test preparation, the SAT and college enrollment. *Social forces*, *89*(2), 463-474.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083-3107.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399-424.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine, 26*, 734-753.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, *25*(12), 2084-2106.

Author (2017). What Works Clearinghouse™ Standards Handbook (Version 4). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Azen, R., & Walker, C. M. (2011). *Categorical data analysis for the behavioral and social sciences*. Routledge.

Bai, H. (2011). A comparison of propensity score matching methods for reducing selection bias. *International Journal of Research & Method in Education*, *34*(1), 81-107.

Bai, H. (2015). Methodological considerations in implementing propensity score matching. In W. Pan & H. Bai (Eds.), *Propensity Score Analysis: Fundamentals and Developments* (pp. 74-88). New York, NY: Guilford Publications, Inc.

Bai, H., & Clark, M. H. (2018). *Propensity Score Methods and Applications* (178). Sage Publications.

Benedetto, U., Head, S. J., Angelini, G. D., & Blackstone, E. H. (2018). Statistical primer: propensity score matching and its alternatives. European Journal of Cardio-Thoracic Surgery, 53(6), 1112-1117.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149-1156.

Brownlee, J. (2016, September 22). Supervised and Unsupervised Machine Learning Algorithms. Retrieved from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow education, American style: Test preparation, the SAT and college enrollment. *Social forces*, *89*(2), 435-461.

Burgette, L. F., McCaffrey, D. F., & Griffin, B. A. (2015). Propensity score estimation with boosted regression. *Propensity Score Analysis: Fundamentals, Developments and Extensions. New York: Guilford Publications, Inc*.

Caliendo, M., & Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31-72.

Clark, M. H. (2015). Propensity score adjustment methods. *Propensity Score Analysis: Fundamentals and Developments, New York and London: The Guilford Press*, 115-140.

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

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

Cohen, J. Cohen, P., West, S. & Aiken, L. (2003). Alternative regression models: Logistic, Poisson regression, and the generalized linear model. In *Applied Multiple Regression /Correlation Analysis for the Behavioral Sciences* (3rd ed) Hillsdale: Erlbaum.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, *27*(4), 724-750.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics, 95*(3), 932-945.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49,* 1231–1236.

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y., & Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine, 31*(7), 681-697.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405-420.

Guo, S., & Fraser, M. W. (2015). *Propensity Score Analysis*. Sage.

Harder V.S., Stuart E.A., Anthony J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal association in psychological research. *Psychological Methods 15*(3), 234-249.

Harris, H.D. & Horst, S.J. (2016). A brief guide to decisions at each step of the propensity score matching process. *Practical Assessment in Research and Evaluation, 21,* 1-10.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer. New York, NY.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199-236.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2015). MatchIt: Nonparametric preprocessing for parametric causal inference. Package "MatchIt" R documentation. https://cran.rproject.org/web/packages/MatchIt/MatchIt.pdf

Hoaglin, D.C., Mosteller, F., Tukey J.W. (1983). Understanding robust and exploratory data analysis.

Holzman, M. A., & Horst, S. J. (2019, April). *Treatment-comparison group ratio and accuracy of treatment effect estimates after propensity score matching.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, CA.

Hume, D. (2003). *A treatise of human nature*. Courier Corporation.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. The Review of Economics and Statistics, 86, 4–29

Jacovidis, J. N., Foelber, K. J., & Horst, S. J. (2017). The effect of propensity score matching method on the quantity and quality of matches. *The Journal of Experimental Education*, *85*(4), 535-558.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435-454.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337-346.

Lison, P. (2015). An introduction to machine learning. *Language Technology Group (LTG)*, *1*, *35*.

Luellen, J. (2007). A comparison of propensity score estimation and adjustment methods on simulated data (Unpublished doctoral dissertation). The University of Memphis, Memphis, TN.

Lumley, T. (2019). "survey: analysis of complex survey samples." R package version 3.35-1.

Lumley, T. (2004). "Analysis of Complex Survey Samples." *Journal of Statistical Software*, **9**(1), 1-19. R package version 2.2.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison approach* (2nd ed.). New York, NY: Psychology Press.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine, 32,* 3388-3414.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological methods, 9(4), 403.

McCandless, L. C., Richardson, S., & Best, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association*, *107*(497), 40-51.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2005). Logistic regression. *In Applied Multivariate Research*. Sage. Thousand Oaks, CA.

Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary

angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54*(4), 387-398.

Osborne, J. W. (2002). Effect sizes and the disattenuation of correlation and regression coefficients: lessons from educational psychology. *Practical Assessment, Research, and Evaluation, 8*(1), 11.

Olmos, A., & Govindasamy, P. (2015). A practical guide for using propensity score weighting in R. *Practical Assessment, Research, and Evaluation, 20*(1), 13.

Osborne, J. (2012). Logits and tigers and bears, oh my! A brief look at the simple math of logistic regression and how it can improve dissemination of results. *Practical Assessment, Research, and Evaluation, 17*, 1-10.

Pan, W., & Bai, H. (Eds.). (2015). *Propensity score analysis: Fundamentals and developments*. Guilford Publications.

Park, J. J. (2012). It takes a village (or an Ethnic economy) the varying roles of socioeconomic status, religion, and social capital in SAT preparation for Chinese and Korean American students. *American Educational Research Journal, 49*(4), 624-650.

Park, J. J., & Becks, A. H. (2015). Who benefits from SAT prep?: An examination of high school context and race/ethnicity. *The Review of Higher Education, 39*(1), 1-23.

Parsons, L. S. (2004, May). Performing a 1: N case-control match on propensity score. In *Proceedings of the 29th Annual SAS users group international conference* (pp. 165-29). SAS Users Group International, Cary, NC.

Rassen, J. A., Shelat, A. A., Myers, J., Glynn, R. J., Rothman, K. J., & Schneeweiss, S. (2012). One-to-many propensity score matching in cohort studies. *Pharmacoepidemiology and Drug Safety*, *21*, 69-80.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2017). Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package. *Santa Monica, CA: RAND Corporation*.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*(408), 1024-1032.

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In Social Statistics Section, *Proceedings of the American Statistical Association*, 233-239.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318-328.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169–188.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. Psychological Methods, 13, 279–313

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. New York, NY: Houghton Mifflin Company.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, *60*, 607-629.

Shadish, W. R. (2013). Propensity score analysis: Promise, reality and irrational exuberance. *Journal of Experimental Criminology*, *9*(2), 129-144.

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice*, *35*(3), 38-54.

Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, *27*, 325-353

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

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*, 250-267.

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. Practical Assessment, *Research & Evaluation*, *18*(13), 1-12.

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

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inferences. In J.W. Osborne (Ed.), *Best Practices in Quantitative Methods*, pp. 155-176. Los Angeles, CA: SAGE Publications.

Wainer, H. (2015). *Truth or truthiness: Distinguishing fact from fiction by learning to think like a data scientist.* Cambridge University Press.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology, 63*(8), 826.

Wickham, H. (2016). Ggplot2: Elegant graphics for data analysis. *Springer-Verlag. New York.*