Spring 2017

Evaluating the performance of propensity score matching methods: A simulation study

Jessica N. Jacovidis
James Madison University

Evaluating the Performance of Propensity Score Matching Methods: A Simulation Study

Jessica N. Jacovidis

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2017

FACULTY COMMITTEE:

Committee Chair: Christine E. DeMars

Committee Members/ Readers:

Allison J. Ames

S. Jeanne Horst

Dena A. Pastor

Dedication

I dedicate this dissertation to my husband, Scott Jacovidis and my son, Aiden Jacovidis. I love you both so much; you are my world. You have been through this whole experience with me. I would not be where I am today without your love and support. Thank you for helping to make this possible. I look forward to the next chapter of our life together.

Acknowledgements

I would like to acknowledge several people who have made this dissertation possible. First, I would like to thank my advisor and dissertation chair, Dr. Christine DeMars. I am forever indebted to you for being my advisor and for all of the advice and guidance you have given me throughout the last three years. You are brilliant and I hope that one day I can be half the researcher that you are. Additionally, I am grateful that you gave me the freedom to be involved in a variety of projects as I was figuring out my own research interests. I also appreciate your willingness to work with me on this dissertation; especially since this is not your research area. Further, your patience has helped to keep me sane and allowed me to have some level of work-life balance. I could not have asked for a better advisor.

I would also like to thank the other members of my dissertation committee, Dr. Allison Ames, Dr. Jeanne Horst, and Dr. Dena Pastor. I appreciate the thoughtful and thorough feedback that you have given me throughout this process. I know this is a better dissertation because of your involvement. It has been a pleasure to work with you throughout my doctoral studies. I have learned so much and I cannot thank you enough.

Next, I would like to thank the Assessment and Measurement faculty that I have had the pleasure of working with throughout the last three years. I have grown so much during my time at JMU and each and every one of you have contributed to my personal and professional development. I would like to provide a special thanks to Dr. Sara Finney. I would not be at JMU if it was not for you. Further, I would like to thank the other students in CARS. Having others to share in this experience has made graduate

school more manageable. I have learned so much from our discussions both inside and outside of the classroom. I look forward to working with you in the future.

I would also like to thank my family and friends. First and foremost, I want to thank my mom. I cannot express how appreciative I am for all that you have done for me. None of this would have been possible without your love and support. I would also like to thank my grandmother, Judy Dotson. I wish she could be here to see how far I have made it. I would be remiss if I did not also thank Dr. Hall "Skip" Beck. I cannot begin to tell you how much of an impact you have had on my life. I am fortunate to call you my mentor, my colleague, and my friend.

Last, but certainly not least, I must thank my husband, Scott Jacovidis, and my son, Aiden Jacovidis. Scott, I cannot thank you enough for everything you have done for me. I am eternally grateful for your love, support, and encouragement throughout this process. I appreciate how understanding you have been when so much of my time was dedicated not only to this dissertation, but to my doctoral program. I cannot wait until we can have the life we want. We are almost there! Aiden, you are my little love. I know you do not fully understand what I am doing or why I work so much, but I hope someday you realize that I did this all for you. I want you to have a great life and I will do everything I can to make that possible.

Table of Contents

List of Tables

List of Figures

Abstract

In education, researchers and evaluators are interested in assessing the impact of programs or interventions. Unfortunately, most education programs do not lend themselves to random assignment; participants generally self-select into programs. Lack of random assignment limits the claims that researchers can make about the impact of the program because individuals who self-select into the program may be qualitatively different from individuals who do not self-select into the program. Propensity score matching allows researchers to mimic random assignment by creating a matched comparison group that is similar to the treatment group on researcher-identified variables.

There are a number of matching methods to choose from when employing propensity score matching. Matching methods vary in distance measures, matching algorithms, and rules for comparison group member selection that are used. Thus, the purpose of this study was to examine common matching techniques to determine how they differed in terms of the *quantity* and *quality* of matches and whether the results of subsequent group comparisons (e.g., significance test results, estimated effect sizes) varied across the different matching techniques. Differences across effect size, treatment group sample size, comparison-to-treatment ratio, and analysis technique were also examined.

To empirically investigate the performance of common matching methods under known and systematically manipulated conditions, data were simulated to reflect values found in higher education, using a recent study by Jacovidis and her colleagues (in press). The choice of matching method dictates both the quality and quantity of the matches obtained and the resulting outcome analyses (e.g., statistical significance tests and

estimated effect sizes). Although nearest neighbor matching with calipers produced better quality matches than the other matching methods, it also resulted in the loss of treatment group members. If treatment group members are excluded from the matched groups, representation of the treatment group could be compromised. If this happens, the researcher may want to select a matching method that does not result in a loss of treatment group members. It is up to the researcher to decide how to best balance the quality and quantity of matches, while recognizing that this decision can impact the accuracy of the outcome analyses.

CHAPTER 1

Introduction

Throughout the last decade, there has been increased use of propensity score matching in education research and evaluation (e.g., Branda & Xieb, 2010; Melguizo, Kienzl, & Alfonso, 2011; Schochet, D'Amico, Berk, Dolfin, & Wozny, 2012; Titus, 2007). This is largely because education researchers and evaluators are attempting to assess the impact of programs or interventions in situations where random assignment is not possible. Propensity score matching is one option for establishing an equivalent comparison group when random assignment is not feasible. Moreover, many federal and state agencies that fund education programs have increased their demand for rigorous research and evaluation designs, often including the explicit requirement of an equivalent comparison group (National Science Foundation, 2016; US Department of Education, 2015; US Department of Labor, 2014; What Works Clearinghouse, 2014).

Although there is extensive research related to propensity score matching, there is little guidance on some of the decision points in the propensity score matching process. Specifically, additional direction is needed on how to select a matching method, how that selection may impact the obtained matches, and ultimately, how that selection may impact the outcome analyses. The purpose of this study was to examine common matching techniques to determine how they differed in terms of the *quantity* and *quality* of matches and whether the results of subsequent group comparisons (e.g., significance test results, effect sizes) varied across the different matching techniques and conditions.

Background Context

Both K-12 and higher education personnel implement programs to improve instruction and pedagogy and promote student learning and development. Given that education agencies (e.g., federal, state, and local departments of education, foundations, and other funding agencies) invest substantial resources in these types of programs, it is important to evaluate whether the participants change in the expected ways. More importantly, evidence is needed to demonstrate that these changes are attributable to program participation, as opposed to maturation or other life experiences. To make these causal claims, research and evaluation on the programs must use rigorous methodologies that warrant such claims.

True experimental design (also called randomized controlled trials) is seen as the "gold standard" in research and evaluation methodologies. At the cornerstone of true experimental design is random assignment—participants are randomly placed into the treatment or comparison groups. Theoretically, random assignment ensures that any variation between the two groups prior to treatment is random. That is, treatment and comparison group members vary only randomly on background and experience variables, effectively controlling for the effect of these variables on the outcome.

Unfortunately, education programs often do not lend themselves to random assignment; generally, participants self-select (or are selected by administrators) into the program instead of being randomly assigned. As such, individuals who self-select into the program may be qualitatively different from individuals who do not participant in the program (Cook, 1999; Davies, Williams, & Yanchar, 2008). For example, suppose researchers have a teacher professional development program in which teachers volunteer

to participate. First, the researchers would be limited by the fact that only those teachers who are interested in the program will participate. It seems reasonable that teachers interested in participating in a professional development program would differ on key characteristics that may also relate to the outcome. For instance, these teachers may be more motivated to learn about novel instructional strategies and may be more willing to try new strategies in their classrooms, even before participating in the professional development program. If this is the case and the researchers compared participants and non-participants, they may erroneously conclude that the program *caused* teachers to use more diverse instructional strategies in their classrooms. Conversely, suppose that the professional development program was focused on teacher confidence and only early career teachers self-selected into the program. If the researchers compared these early career teachers to a group comprised of experienced teachers, then they may incorrectly conclude that the program was ineffective, as experienced teachers will likely have higher confidence levels than early career teachers. Ultimately, in both scenarios, the two groups were qualitatively different prior to the treatment. As such, researchers cannot parse out the effects of the program from preexisting differences—they are confounded.

Random assignment addresses this issue of confounding and allows researchers and evaluators to make causal claims regarding the effects of programs (Shadish, Cook, & Campbell, 2002). However, as already mentioned, true experimental design is often impractical in education contexts. As such, researchers and evaluators are forced to employ quasi-experimental designs or observational studies to assess the effectiveness of their programming. This limits the causal claims that researchers can make about the

impact of their programs and makes it difficult to differentiate the effect of the program from systematic preexisting differences (i.e., self-selection bias; Winship & Mare, 1992).

Overview of Propensity Score Matching

Propensity score matching provides one approach for creating comparable groups based on students' *propensity* for participation in the intervention (regardless of whether or not they actually participated). Propensity score matching creates a matched comparison group that is similar to the treatment group on a set of covariates (Austin, 2011b; Guo & Fraser, 2015; Luellen, Shadish, & Clark, 2005; Rosenbaum & Rubin, 1983; Stuart, 2010; Stuart & Rubin, 2008). Thus, propensity score matching allows researchers to mimic random assignment by balancing the distributions of the covariates across the treatment and matched comparison groups. Theoretically, balancing the groups on the propensity scores controls for the impact of the covariates on the outcome and allows for more meaningful group comparisons (e.g., more accurate estimates of the treatment effect) than if the propensity scores, and thus, covariates, were unbalanced.

Propensity score matching involves a series of steps: 1) select appropriate covariates, 2) compute a distance measure, 3) select a matching method (e.g., nearest neighbor, optimal matching), 4) create matched groups, 5) diagnose the quality of matches, and 6) examine group differences on the outcome (Harris & Horst, 2016). Although each step in propensity score matching requires careful consideration, this study focused on selecting matching methods (Step 3), how selection of the matching method influenced the quality of matches (Step 5) and treatment effects estimated in the outcome analyses (Step 6).

Matching Methods

Researchers have noted that additional research is needed to systematically examine what propensity score matching methods perform well under what data conditions (Austin, 2013; Bai, 2015). Matching methods employ different distance measures (i.e., propensity scores or Mahalanobis distances), matching algorithms (i.e., greedy or optimal), and rules for comparison group member selection. Thus, different matching methods could result in the selection of different comparison group members from the overall comparison pool. Selection of matching method will not only affect the quality of matches, but may also affect the results of any outcome analyses. As such, it is important to understand how the matching methods differ.

Distance Measure. There are various ways to compute the distance measure; however, two of the most common distance measures are propensity scores calculated via logistic regression (Guo & Fraser, 2015; Olmos & Govindasamy, 2015) and Mahalanobis distances (Cochran & Rubin, 1973; Guo & Fraser, 2015; Rubin, 1979). A key difference between propensity scores and Mahalanobis distances concerns the weighting of the covariates. Mahalanobis distances equally weight all covariates, taking into consideration variances and covariances of the covariates, whereas propensity scores weight covariates by how well they predict group membership.

Matching Algorithm. Propensity score matching typically employs one of two matching algorithms: greedy or optimal. The distinction between the greedy and optimal algorithms is whether matches are re-evaluated and modified throughout the matching process. The greedy algorithm proceeds sequentially, matching each treatment group member to the closest available comparison pool member based on the distance measure.

Matches are not modified at later stages in the matching process (Gu & Rosenbaum, 1993; Stuart, 2010; Stuart & Rubin, 2008). Conversely, the optimal algorithm re-evaluates the total distance between matched groups at each step and may alter earlier matching decisions, if the change will result in the smallest average absolute distance across all matched pairs (Gu & Rosenbaum, 1993; Guo & Fraser, 2015; Ho et al., 2007, 2011; Pan & Bai, 2015; Stuart, 2010).

Matching Methods. Variations of four matching methods were included in this study: random sampling, nearest neighbor matching, nearest neighbor matching with calipers, and optimal matching. In random sampling, as the name implies, a subset of the larger comparison group reservoir is randomly selected; this technique does not consider distance measures or covariates. Nearest neighbor uses a greedy algorithm to match each treatment group member to the closest available comparison pool member (Gu & Rosenbaum, 1993; Stuart, 2010; Stuart & Rubin, 2008). Nearest neighbor matching can be used with propensity scores or Mahalanobis distances and this study included both. When calipers are applied to nearest neighbor matching, treatment group members are only matched to comparison pool members if the propensity scores are within the researcher-specified caliper distance. Three calipers were applied in the current study: 0.1, 0.2, and 0.3 standard deviations of the logit of the propensity score. Optimal matching uses an optimal algorithm to match each treatment group member to the closest available comparison pool member, thus matches are re-evaluated and may be modified throughout the matching process (Gu & Rosenbaum, 1993; Guo & Fraser, 2015; Ho et al., 2007, 2011; Stuart, 2010). This study examined the performance of optimal matching

with one comparison to one treatment group member (optimal 1:1) and two comparison to one treatment group members (optimal 2:1).

Comparing Matching Methods. Studies comparing the matching methods described above are limited. Typically, performance of matching methods is determined by how well the matching method can balance the groups on the distance measure and the covariates or by how well the matching method reduces selection bias (Pan & Bai, 2015; Stuart & Rubin, 2008). Propensity scores result in better balanced groups than Mahalanobis distances when there are a large number of covariates (e.g., 20; Gu & Rosenbaum, 1993); however, the two distance measures result in comparable balance when there are a small number of covariates (e.g., 2 to 8; Gu & Rosenbaum, 1993; Zhao, 2004). Further, when treatment group members compete for comparison group members, the optimal algorithm outperforms the greedy algorithm (Gu & Rosenbaum, 1993); otherwise, the greedy and optimal matching approaches perform comparably in creating groups with balanced covariates (Austin, 2009b, 2013; Bai, 2013; Gu & Rosenbaum, 1993).

When calipers are applied to nearest neighbor matching, covariates and propensity scores are more balanced than when calipers are not applied (Austin, 2009b, 2013; Bai, 2015; Jacovidis et al., in press; Rosenbaum & Rubin, 1985). However, nearest neighbor matching with calipers also generally results in a loss of treatment group members (e.g., Austin, 2009b, 2013; Bai, 2015; Jacovidis et al, in press), as treatment group members who are not able to be matched are excluded from the matched data set. Moreover, as the caliper becomes more stringent, propensity score and covariate balance improves (Austin, 2009b, 2010a; Dehejia & Wahba, 2002; Jacovidis, in press), but loss of

treatment group members tends to be greater (e.g., Austin, 2009b, 2013; Dehejia & Wahba, 2002; Jacovidis et al., in press). This often results in tension between obtaining equivalent groups and maintaining representation of the original treatment group.

Few researchers have included the impact of matching methods on outcome analyses as part of the evaluation of matching method performance (Austin, 2013; Jacovidis et al., in press; Stone & Tang, 2013). As such, it is difficult to make general conclusions about matching method performance. Further, different decisions could be made about whether there was a statistically significant difference between groups, depending on which matching method was used (Jacovidis et al., in press). However, additional research is needed in this area and that is a primary goal of this study.

Limitations of Current Matching Method Research. Although comparison of matching methods have received some attention in the propensity score literature, there are a few limitations of note. First, studies comparing matching methods have not been systematic. Thus, it is difficult to determine what matching method should be used in what situation. Second, many studies are conducted using applied data. These studies compare matching methods on propensity score and covariate balance and bias reduction; however, it is difficult to include an examination of outcome analyses, as true group differences are typically unknown. Finally, in simulation studies, the simulated data are often unrealistic in that the simulated covariates are either all continuous or all binary. Additionally, the covariates are almost always simulated to be independent (e.g., Austin, 2011a, 2013). This may or may not make a difference in matching method performance; however, that is an empirical question that has not been investigated.

Other Study Conditions

This study also included an examination of the influence of treatment group sample size, comparison-to-treatment group ratio, and type of outcome analysis on matching method performance (e.g., covariate and distance measure balance and treatment effect estimates). Relevant information regarding these areas is described below.

Sample Size and Comparison-to-Treatment Ratio. The propensity score matching literature regarding sample size and comparison-to-treatment group ratio is unclear. There appears to be a complex interplay among total sample size, treatment group sample size, and comparison-to-treatment group ratio. This is also intertwined with how similar on the covariates the members of the comparison group reservoir are to the treatment group members (e.g., common support), as the similarity between groups heavily influences whether or not adequate matches can be found. Although these issues have been examined, the examination has not been systematic (e.g., Bai, 2015; Dehejia & Wahba, 2002; Rosenbaum & Rubin, 1983; Rubin, 1979). Thus, clear guidelines do not exist for researchers and evaluators as they conduct propensity score matching studies. Additional research is needed in this area. As such, the current study included an examination of the performance of the matching methods with different treatment group sample sizes and comparison-to-treatment group ratios.

Outcome Analyses. Ultimately, the goal of propensity score matching is to obtain comparable groups so that the researchers can examine group differences on the outcome of interest. There appears to be misalignment between the recommended approach to outcome analyses and the approach that researchers have taken in applied

practice. Propensity score researchers recommend that any covariates included in the matching model that remain unbalanced after matching should be included in the outcome analyses (Pan & Bai, 2015; Rosenbaum & Rubin, 1985). This technique has been shown to produce accurate estimates of treatment effects regardless of the choice of propensity score matching methods (Schafer & Kang, 2008; Shadish, Clark, & Steiner, 2008). However, researchers and evaluators using propensity score matching in applied settings often conduct group comparisons without including unbalanced covariates (e.g., Clark & Cundiff, 2011; Lu, Zanutto, Hornik, & Rosenbaum, 2001; Morgan, Frisco, Farkas & Hibel, 2010; Olitsky, 2013). Although the decision on whether to include unbalanced covariates will make little difference if the groups are balanced, it could influence the inferences made if the groups are still unbalanced on the covariates after matching.

The Current Study

Given that matching methods employ different distance measures, algorithms, and rules for selecting comparison group members, each technique could potentially lead to the selection of different comparison group members from the overall comparison pool to create the matched comparison group. Moreover, matched comparison group composition could vary considerably depending on the matching algorithm used. This will not only impact the quality of matches, but may also impact the results of any outcome analyses (e.g., Austin, 2013; Jacovidis et al., in press; Stone & Tang, 2913). As noted, additional research is needed on matching methods to provide guidance to practitioners on which matching methods perform the best under which conditions, and this study was meant to be one in a line of research on matching methods. The purpose

of this study was to examine and compare common matching techniques under systematically manipulated conditions, representative of program evaluation and effectiveness studies. Specifically, the current study addressed four research questions.

Research Question 1: How do the most common matching methods differ, in terms of quantity (i.e., number of matches) and quality (i.e., covariate balance) of matches? Each matching method selects comparison units from the comparison pool reservoir in a different manner. Some of the matching techniques result in the best match regardless of how close the match is (e.g., nearest neighbor, optimal matching), whereas other techniques require that the match fall within a specified distance from the treatment unit (e.g., caliper matching). If no matches can be found within the specified distance, then the treatment unit is dropped from further analyses. It is quite possible for different matching techniques to create comparison groups that are each composed of different individuals from the overall comparison pool. Further, it is possible that the matching technique that results in the best covariate balance also results in the loss of treatment units (e.g., Austin, 2009b, 2013; Bai, 2015; Jacovidis et al, in press). This research question explored these issues.

Research Question 2: Once matched comparison groups are formed, how do the results of group comparisons (e.g., significance tests) on the outcome compare across the different matching techniques? Given that the matched comparison groups could be composed of different individuals from the overall comparison pool, it stands to reason that the results of any outcome analyses may differ depending on the selected matching algorithm (Austin, 2013; Jacovidis et al., in press; Stone & Tang, 2013). This question addressed issues of Type I error and power for various effect sizes.

Research Question 3: How well do the matching methods recover the true treatment effect (e.g., difference between the group means)? Of particular interest was the bias and root mean squared error of effect size estimates.

Research Question 4: What conditions are optimal in obtaining accurate estimates of parameters? Specifically, what combinations of true difference between the means, matching method, comparison-to-treatment ratio, sample size, and outcome analysis affect the treatment effect estimates?

CHAPTER 2

Literature Review

Often, education researchers and evaluators want to assess the impact of programs or interventions in situations where random assignment is not possible. When participants self-select into programs, variables related to participation in the intervention, particularly those related to the outcome of interest, can influence the inferences drawn from the findings. If researchers conclude that the intervention is impactful (attributing group differences to the intervention when the intervention was not impactful), they could be making flawed conclusions. Propensity score matching provides one approach for dealing with this situation by creating comparable groups based on an individual's *propensity* for participation in the intervention, regardless of actual participation. Although there is extensive research on propensity score matching, there is little guidance on how to select a matching method, how that selection may impact the obtained matches, and ultimately, how that selection may impact the outcome analyses. The purpose of the current study was to examine and compare the performance of common matching techniques under manipulated conditions and make recommendations regarding the use of those matching methods.

Overview of Propensity Score Matching

Propensity score matching is a technique that allows researchers to create a matched comparison group that is similar to the treatment group on a set of researcher-identified characteristics, called covariates (Austin, 2011b; Guo & Fraser, 2015; Luellen et al., 2005; Rosenbaum & Rubin, 1983; Stuart, 2010; Stuart & Rubin, 2008). Propensity score matching reduces selection bias by controlling for covariates related to self-

selection into the treatment group, the outcome of interest, or both (Austin, 2007a; Austin, Grootendorst, & Anderson, 2007; Caliendo & Kopeinig, 2008; Guo & Fraser, 2015; Stuart & Rubin, 2008). Propensity score matching allows researchers to mimic random assignment by balancing the distributions of the covariates across the treatment and matched comparison groups. That is, propensity scores are calculated from a set of covariates, and then matches are created based on those scores, effectively controlling for groups differences on the covariates.

Propensity scores are defined as the probability of treatment group membership, conditional upon a set of observed covariates (Joffe & Rosenbaum, 1999; Rosenbaum & Rubin, 1983). The formal definition of a propensity score is shown in Equation 1,

$$p(\boldsymbol{X_i}) = \Pr(T_i = 1|\boldsymbol{X_i}) \quad (1)$$

where Pr represents the probability of treatment group membership, T_i represents binary group membership (0 for comparison, 1 for treatment) for person i and $\boldsymbol{X_i}$ represents the vector of covariates for person i. Theoretically, balancing on the propensity scores controls for the impact of the covariates on the outcome. This allows for more meaningful group comparisons (e.g., more accurate estimates of the treatment effect) than if the propensity scores were unbalanced (Rosenbaum & Rubin, 1983).

Logic of Propensity Score Matching. In education research and evaluation, researchers are interested in estimating the effects of their programs. Thus, they want to make causal statements, attributing group differences on some outcome of interest between treatment and comparison groups to their program. As noted in Rubin's Causal Model, there are two possible outcomes for each individual (Rubin, 1974): each individual could serve as a participant or a comparison group member. Ultimately,

researchers want to know the outcome score for each individual under both group assignments. By comparing the two potential outcomes, researchers can obtain an estimate of the causal effect (Rubin, 1974). However, for any one individual, researchers can only observe one of the outcomes (Rubin, 1974, 1978). This is a fundamental problem in causal modeling: it is impossible to observe the outcome of interest for the same individual in both the treatment and comparison group simultaneously (Rosenbaum & Rubin, 1983; Rubin, 1974).

To make a causal linkage, researchers need to obtain some estimate of how the individuals would have performed on the outcome had they not received the treatment, known as the counterfactual (Rosenbaum & Rubin, 1983; Rubin, 1974). Without a viable estimate of the counterfactual, researchers cannot rule out that the observed differences would have happened regardless of the program. Thus, researchers attempt to obtain a credible estimate of the counterfactual in order to estimate the causal effect (Caliendo & Kopeinig, 2008; Holland, 1986; Pattanayak, 2015; Rubin, 1974).

When random assignment is used, the comparison group serves as a proxy for the counterfactual because the two groups vary only randomly on observed and unobserved covariates. Thus, the causal effect can be obtained by directly comparing the outcomes of the treatment and comparison groups (Rosenbaum & Rubin, 1983; Rubin, 1974). However, when random assignment is not possible, direct comparisons could be misleading, as treatment and comparison group members may differ systematically (Rosenbaum & Rubin, 1983).

Propensity score matching is one option for creating a counterfactual group when random assignment cannot be employed (Rosenbaum & Rubin, 1983). As noted

previously, propensity score matching allows researchers to create a matched comparison group that is similar to the treatment group on a set of researcher-identified characteristics, thus mimicking random assignment by balancing the distributions of the covariates across the treatment and matched comparison groups. That is, propensity score matching results in an estimate of the counterfactual by allowing researchers to create a matched comparison group that is similar to the treatment group on a set of covariates. If propensity score matching assumptions are met, the matching results in a comparison group that differs from the treatment group solely on assignment to the program. Thus, balancing on the covariates allows for a direct comparison between the participant and matched comparison groups that is more meaningful than if the covariates were unbalanced (Rosenbaum & Rubin, 1983).

There are two indices that are frequently used to estimate the average treatment effects in propensity score matching: average treatment effect (ATE) and average treatment effect on the treated (ATT; Caliendo & Kopeinig, 2008). The ATE is the average treatment effect estimated for a given population. That is, ATE is used to make inferences about the potential impact of the program for the whole population, if the whole population received the treatment (Austin, 2011b; Caliendo & Kopeinig, 2008; Ho et al., 2007). The formula for the ATE is presented as Equation 2 (Ho et al., 2007, p. 204),

$$ATE = \frac{1}{n}\sum_{i=1}^{n} E[Y_i(1) - Y_i(0)|X_i] \quad (2)$$

where Y_i (1) is the expected value of the outcome for person *i* when treated, Y_i (0) is the expected value of the outcome for person *i* when untreated, X_i is the vector of covariates for person *i*. In propensity score analyses, ATE cannot be computed directly. Instead,

estimation of the ATE requires an extrapolation of treatment across levels of the covariates and to the whole sample.

The ATT is an estimate of the average treatment effect for the population represented by the group who actually participated in the program. That is, ATT only involves making inferences about the individuals who participated or would be interested or eligible in participating in the program (Austin, 2011b; Caliendo & Kopeinig, 2008; Ho et al., 2007). The formula for the ATT is presented as Equation 3 (Ho et al., 2007, p. 204),

$$ATT = \frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} T_i E[Y_i(1) - Y_i(0) | X_i] \quad (3)$$

where T_i is treatment group membership, Y_i (1) is the expected value of the outcome for person *i* when treated, Y_i (0) is the expected value of the outcome for person *i* when untreated, X_i is the vector of covariates for person *i*. This is the approach that is typically taken when propensity score matching is used to create matched groups, and is the approach used in the current study. The ATT is straightforward to estimate: the researcher compares the treatment and matched comparison group by conducting the appropriate inferential tests dictated by the research question of interest (Caliendo & Kopeinig, 2008; Gu & Rosenbaum, 1993; Ho et al., 2007; Stuart, 2010; Stuart & Rubin, 2008).

Researchers and evaluators should decide whether the ATE or the ATT is of interest in their particular research. For example, in the teacher professional development program discussed earlier, researchers may want to estimate the potential impact of that program on all teachers in a given school or district. In this situation, the researcher would be interested in estimating the ATE. However, researchers may only want to

estimate the impact of the program on the teachers like those who participated; researchers may not be interested in the effects of the program on those who were not eligible to participate or who chose not to participate. In this situation, the researcher would be interested in estimating the ATT. Generally, in education research and evaluation studies and in studies that use propensity score matching to create matched groups, researchers are typically interested in the ATT. Moreover, given that *matching* is the focus of the current study, ATT was estimated as an index of the average treatment effects.

Assumptions. Although propensity score matching can be a useful technique, it relies on strong assumptions. One assumption of propensity score matching is conditional independence. When conditional independence is assumed, treatment group assignment is "strongly ignorable" (Burgette, McCaffrey, & Griffin, 2015; Rosenbaum & Rubin, 1983, p. 43). That is, after controlling for covariates, assignment to the treatment group is essentially random and each individual has the same probability of treatment, as in random assignment (Rosenbaum & Rubin, 1983). However, this also requires that all relevant covariates related to participation are included in the matching model (e.g., there are no unmeasured covariates). It is this assumption that allows the matched comparison group to be used as the counterfactual for the treatment group (Rosenbaum & Rubin, 1983). However, this assumption is often unrealistic given that researchers can never be certain that all key covariates have been included in the matching model.

Another related assumption of propensity score matching is common support. Common support addresses the extent to which the participant and comparison groups are similar on their distributions of propensity scores (Caliendo & Kopeinig, 2008; Stuart,

2010). Common support is required to find adequate matches and is a necessary, but not sufficient, condition for local independence. A lack of common support may result in too few matched pairs (Caliendo & Kopeinig, 2008; Stuart, 2010). However, when there is adequate overlap in the distribution of propensity scores, most matching techniques will produce similar results (Bai, 2015). Researchers can examine the densities of the propensity scores to determine whether there is sufficient common support to produce adequate matches. Although there is no standard for common support, Rubin (2001) suggested that there should be less than a 0.5 standard deviation unit difference between the groups on their average propensity scores. If there are large differences between the minimum and maximum propensity scores between groups, some researchers (e.g., Caliendo & Kopeinig, 2008; Guo & Fraser, 2015) suggest removing cases from the comparison group that lie outside of the region of support for the treatment distribution.

As with any study designed to make causal inferences, propensity score matching studies must also meet the stable unit treatment value assumption (Rosenbaum & Rubin, 1983; Rubin, 1986). This assumption essentially deals with contamination of the comparison group. Ultimately, if participants share information about the program with comparison group members, this could lead to an effect on the outcome of interest. For example, in the teacher professional development program scenario, if participating teachers shared instructional strategies with teachers in the comparison group, then this may impact the instructional strategies that comparison teachers use in their classroom, thus the outcome of interest has been contaminated by the sharing of information.

Advantages and Disadvantages. Propensity score matching has a few notable advantages and disadvantages. One advantage of propensity score matching is that it

uses a linear combination of covariates to form a composite that can be used to balance the treatment and comparison groups. As such, researchers can match on a large number of covariates without the decrement in treatment group sample size that would occur if the researcher matched treatment and comparison group members only when individuals had identical values on the covariates (e.g., exact matching). Another advantage of propensity score matching is that it allows researchers to obtain a credible estimate of the counterfactual, when random assignment is not possible. Propensity score matching results in a more precise estimate of the treatment effect than comparing groups with unbalanced covariates. However, this is only true if propensity score matching assumptions are met (Rosenbaum & Rubin, 1983).

One disadvantage is that propensity score matching relies on fairly stringent assumptions. Researchers can never be certain that all key covariates have been included in the matching model. Another related disadvantage of propensity score matching is that it only accounts for observed covariates (Austin, 2011b). Variables that influence self-selection into treatment or the outcome that have not been measured cannot be accounted for in the matching procedure. Thus, any hidden bias due to the unmeasured variables may remain after matching. Further, it is important to note that propensity score matching does not establish causation (Austin, 2011b). Another disadvantage of propensity score matching is that it requires large samples with substantial overlap between the treatment and comparison groups on the covariates (Bai, 2015; Rubin, 1979).

Propensity Score Matching Steps

Numerous researchers have outlined the steps involved in propensity score matching (e.g., Caliendo & Kopeinig, 2008; Guo & Fraser, 2015; Harris & Horst, 2016;

Pan & Bai, 2015; Stuart & Rubin, 2008). The general steps are shown in Figure 1. Specifically, when conducting a propensity score matching study, researchers must consider the following: 1) selecting appropriate covariates, 2) computing the distance measure, 3) selecting a matching method (e.g., nearest neighbor, optimal matching), 4) creating matched groups, 5) diagnosing the quality of matches, and 6) examining group differences on the outcome (Harris & Horst, 2016). Although each step requires careful consideration, this study focused on selecting matching methods (Step 3), diagnosing matches (Step 5), and examining group differences on the outcome (Step 6). For a guide to the decisions at each step of the propensity score matching process, see Harris and Horst (2016).

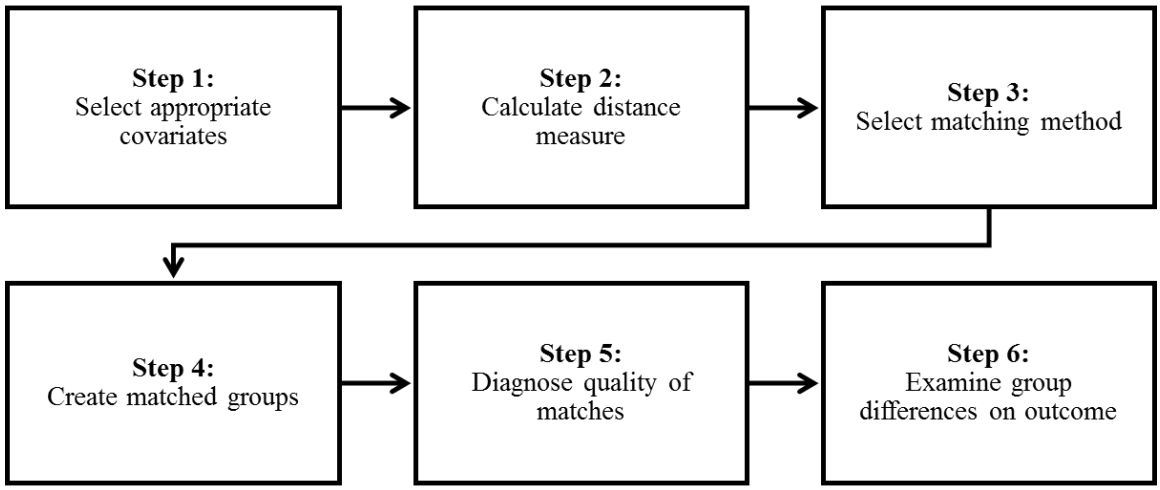

Figure 1. Steps in propensity score matching.

Step 1: Select appropriate covariates. The first step in propensity score matching is selecting appropriate covariates. As noted above, covariates should be related to selection into the treatment group, the outcome of interest, or both (Austin, 2007a; Austin et al., 2007; Caliendo & Kopeinig, 2008; Guo & Fraser, 2015; Stuart & Rubin, 2008). Unlike random assignment, propensity score matching does not balance

on unmeasured covariates (i.e., variables not used in the matching model; Austin, 2011b). If all relevant covariates were included in the matching model, then researchers could completely control for self-selection bias (Steyer, Gabler, von Davier, & Nachtigall, 2000), and propensity score matching would produce a more precise estimate of the treatment effect than would be obtained without matching (Rosenbaum & Rubin, 1983; Rosenbaum & Rubin, 1984). However, if researchers omit important covariates, the treatment and matched comparison group may still be qualitatively different on the unmeasured covariates (Austin, 2011; Steiner et al., 2010; Steiner, Cook, & Shadish, 2011). Thus, excluding key covariates can lead to biased estimates of the treatment effect if those covariates are related to self-selection into the treatment group or the outcome of interest (Austin et al., 2007; Brookhart et al., 2006; Steiner et al., 2010). Although the selection of covariates was not of interest in the current study, selection of appropriate covariates is vital to the meaningfulness of the obtained matches (Caliendo & Kopeinig, 2008; Steiner et al., 2010), and subsequently the inferences about the outcome measures made from the matched groups.

Selection of covariates has received considerable attention in literature and a number of researchers have made recommendations on when to include or exclude certain covariates (Brookhart et al., 2006; Caliendo & Kopeinig, 2008; Rubin, 2001; Steiner et al., 2010; Steiner et al., 2011). Researchers have suggested that covariates that influence the self-selection process should be included in the matching model. For example, in a study comparing experimental and quasi-experimental estimates of treatment effects for mathematics and vocabulary training programs, researchers examined common categories of covariates: demographic variables, proxy-pretest

variables, prior academic achievement, topic preference, and psychological predisposition (Steiner et al., 2010). Researchers found that self-selection into the mathematics training program could be fully explained by topic preference; however, self-selection into the vocabulary training program was more complex, requiring both topic preference and proxy pretest covariates to fully explain self-selection. Further, although the researchers acknowledged that demographic and prior achievement variables were frequently included as covariates in in education, these covariates had little impact on removing self-selection bias (Steiner et al., 2010). Although the importance of these particular covariates may not generalize to other programs, this study illustrates the importance of understanding which covariates are related to the self-selection process.

Other researchers have recommended that covariates that are unrelated to self-selection but are related to the outcome of interest should always be included in the matching model; however, covariates that are related to self-selection, but unrelated to the outcome of interest can bias estimates of the treatment effect. Additionally, the inclusion of variables that are strongly related to self-selection, but only weakly related to the outcome can bias estimates of the treatment effect when total sample size is small (Brookhart et al., 2006). Thus, it is important that researchers understand the literature and program theory to select theoretically sound covariates (Brookhart et al., 2006; Rubin, 2001; Steiner, Cook, Shadish, & Clark, 2010).

Researchers generally recommended that propensity score matching models include a large set of covariates (Austin et al., 2007; Brookhart et al., 2006; Stuart, 2010; Stuart & Rubin, 2008). However, use of a large set of covariates should be balanced with

the fact that covariates need to be observable and measureable (i.e., covariates cannot be included if they have not been measured). This often results in education researchers including covariates that are readily available or easy to measure such as demographic variables (e.g., gender, ethnicity), experience variables (e.g., age, grade, number of years of teaching experience), standardized test scores (e.g., SAT, ACT, GRE, state standardized test scores), dispositional measures (e.g., motivation), and personality traits (e.g., conscientiousness, openness to experience). Given that this is a typical approach to selecting covariates and the goal of this study was to provide recommendations for propensity score matching methods under typical circumstances, the data for this study were modeled using this approach.

Step 2: Compute distance measure. Once covariates have been selected, they can be used to compute the distance measure used for creating matched groups. There are various ways to compute the distance measure, such as logistic regression (Guo & Fraser, 2015; Olmos & Govindasamy, 2015), Mahalanobis distances (Guo & Fraser, 2015; Zhao, 2004), discriminant analysis (Pan & Bai, 2015), boosted regression (Burgette et al., 2015; McCaffrey et al., 2013), Bayesian regression (Stone & Tang, 2013), and classification and regression trees (Lee, Lessler, & Stuart, 2010). The most frequently used method for creating propensity scores, and one of the methods that was used in the current study, is logistic regression (Austin, 2011b; Stuart, 2010). Thus, the propensity score is the probability of participating in a program, given a set of covariates (Luellen et al., 2005). As shown in Equation 4, to compute propensity scores via logistic regression, the researcher simply includes the covariate scores as predictors of treatment group membership.

$$\widehat{p_i} = \frac{1}{1+e^{-(\beta_0+\beta_1(X_{1i})+\cdots+\beta_k(X_{ki}))}} \quad (4)$$

In Equation 4, $\hat{p}_i$ is the predicted probability of being in the treatment group and $\beta_0 + \beta_1(X_{1i}) + \cdots + \beta_k(X_{ki})$ represent the unique contribution of each of the *k* covariates to treatment group membership (Cohen, Cohen, West, & Aiken, 2003, p. 486).

A propensity for treatment (e.g., propensity score) is estimated for each treatment and comparison group member. Individuals with the same propensity score are considered to have the same propensity for participating in the program, regardless of whether or not they actually participated. Moreover, members from different groups with the same propensity score have identical distributions on the set of covariates (Austin, 2011b; Caliendo & Kopeinig, 2008; Ho et al., 2007; Stuart, 2010). Thus, researchers can compare outcomes between individuals who did participate in the program with individuals who did not participate but who have the same propensity for treatment, conditional upon the covariates included in the model.

Another distance measure frequently used to create matched groups, and the second distance measure included in the current study, is Mahalanobis distances (Cochran & Rubin, 1973; Guo & Fraser, 2015; Rubin, 1979). Mahalanobis distance matching is not a propensity score technique. Instead, Mahalanobis distances are used to match treatment and comparison group members. Mahalanobis distance matching was developed prior to propensity score matching (Cochran & Rubin, 1973; Guo & Fraser, 2015). The formula for calculating Mahalanobis distances is presented in Equation 5 (Guo & Fraser, 2015, p. 146),

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

where the distance $MD(i, j)$ is the Mahalanobis distance between treatment group member i and comparison group member j, $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors of covariates for treatment group member i and comparison group member j, respectively, and $\boldsymbol{C}$ is the sample covariance matrix from the full comparison group reservoir.

A primary difference between propensity scores and Mahalanobis distances pertains to the weighting of the covariates. Mahalanobis distance matching equally balances all covariates, also taking into consideration variances and covariances of the covariates, regardless of their relationship with group membership. That is, all covariates are equally important in the calculation of the distance measure and contribute equally to matching (Rosenbaum & Rubin, 1983; Rubin, 1979; Stuart, 2010). Conversely, in propensity score matching, the covariates are weighted by how well they predict treatment group membership. Thus, covariates that have a stronger relationship with treatment group membership are weighted more heavily than covariates that have a weaker relationship with treatment group membership (Gu & Rosenbaum, 1993). That is, covariates are *not* equally important.

Researchers do not agree on whether propensity scores or Mahalanobis distances should be used for matching. Results from simulation studies have suggested that if there are a large number of covariates (e.g., 20), then propensity score matching results in better balanced matches than Mahalanobis distance matching. When there are few covariates (e.g., 2 to 8), the two distance measures result in comparable balance (Gu & Rosenbaum, 1993; Zhao, 2004). Intuitively, this makes sense. Propensity score matching weights covariates, giving greater importance to the ones that can better differentiate between groups (Rosenbaum & Rubin, 1983). Mahalanobis distance

matching balances all covariates equally, which becomes more difficult as the number of covariates increases (Gu & Rosenbaum, 1993). However, some researchers (e.g., King & Nielsen, 2016; Zhao, 2004) still advocate for Mahalanobis distance matching over propensity score matching. Thus, the current study included both propensity score and Mahalanobis distance matching techniques.

Step 3: Select Matching Method. The next step is to select the matching method that will be used to create the matched groups. There are general considerations for researchers irrespective of which matching method they choose. These considerations include one-to-one versus one-to-many matching, matching with or without replacement, sample size, and comparison-to-treatment group ratio. Additionally, there are a variety of matching methods, each employing different algorithms and matching rules. This section provides a discussion of the general considerations, followed by a description of each matching method employed in the current study.

When selecting a matching method, researchers should consider the number of comparison group members that will be matched to each treatment group member (i.e., one-to-one matching or one-to-many matching). Generally, each treatment group member is matched to one comparison group member (one-to-one or pair-matching; Austin, 2013). However, treatment group members can be matched to multiple comparison group members (one-to-many matching; Austin, 2010b). Many of the matching methods apply one-to-one matching by default; however, one-to-many matching can easily be specified in current software packages (e.g., the R package 'MatchIt'; Ho, Imai, King, & Stuart, 2011).

Another consideration is whether to match *with* or *without* replacement. When matching *without* replacement, each comparison pool group member can only be matched to one treatment group member. When matching *with* replacement, there is the potential for the same comparison pool group member to be matched to multiple treatment group members. Findings from simulation studies have suggested that if the treatment group size is less than half of the comparison group reservoir, treatment group members rarely compete for the same comparison group member (Carpenter, 1977).

Some researchers have suggested that matching *with* replacement can result in better quality matches than matching *without* replacement (Caliendo & Kopeinig, 2008; Stuart, 2010). However, there is mixed evidence on whether matching *with* replacement is more effective at reducing bias than matching *without* replacement (Austin, 2013; Bai, 2015; Dehejia & Wahba, 2002). Further, matching *with* replacement may cause a violation of the assumption of independence of observations (i.e., each match is unrelated to the other matches), as some comparison group members could be included more than once (Austin, 2007a, 2009b; Bai, 2015; Caliendo & Kopeinig, 2008; Stuart, 2010). Matching *with* replacement is rarely used in practice (Austin, 2009b; Caliendo & Kopeinig, 2008). Therefore, the current study employed matching *without* replacement for all methods.

Researchers should also consider sample size and comparison-to-treatment group ratio. It is important to note upfront, the literature regarding sample size and comparison-to-treatment group ratio is unclear. Much of the confusion centers around whether researchers are focused on the total sample size (treatment and comparison group, collectively), the treatment group sample size, or the comparison group sample size

compared to the treatment group sample size (comparison-to-treatment ratio). This is further complicated by the different definitions of "small" that researchers use when discussing sample size, without articulating whether they mean total or treatment group sample size. Moreover, it is difficult to disentangle the effect of sample size and comparison-to-treatment group ratio because common support heavily influences whether or not adequate matches can be found.

Propensity score matching was developed to be a large sample size technique; however, it has been applied in small sample size situations (e.g., small-scale program evaluations; Stone & Tang 2013). There are mixed perspectives on whether propensity score matching should be used with small sample sizes and what constitutes a small sample size (e.g. Bai, 2015; Dehejia & Wahba, 2002; Rubin, 1979, 1997; Stone & Tang, 2013; Zhao, 2004). For example, findings from one study suggested that propensity score matching did not perform well when total sample size was "small" (defined by the researcher as $n = 500$) and the comparison-to-treatment group ratio was 5:1 for all sample sizes. However, the correlations between the covariates and group membership was low (Zhao, 2004). Thus, it likely that propensity score matching failed because the researcher violated the assumption of common support, not because the sample size was "small."

There is contradictory evidence about whether propensity score matching performs well with small *treatment* group sample sizes. For example, results from one study indicated that when treatment group sample size was "small" (defined by the researcher as $n = 30$ or 60), propensity score matching did not perform well (Stone & Tang, 2013). The comparison group reservoir for this study contained more than 300 group members, a minimum comparison-to-treatment group ratio of 5:1 (Stone & Tang,

2013). However, results from another study showed that when the treatment group was a little larger (n = 100), but the comparison-to-treatment group ratio was smaller (2:1), propensity score matching still performed well with some matching methods, namely caliper matching (Bai, 2015).

Researchers have suggested that the size of the comparison group reservoir is more influential in matching than the total sample size (Bai, 2015; Dehejia & Wahba, 2002; Rosenbaum & Rubin, 1983; Rubin, 1979). Researchers have examined comparison-to-treatment group ratios from 2:1 to 6:1 and 9:1 and have consistently shown that as comparison-to-treatment group ratio increases, the quality of matches obtained with propensity score matching improves (Bai, 2015; Dehejia & Wahba, 2002; Rosenbaum & Rubin, 1983; Rubin, 1979). This makes sense intuitively—the larger the comparison pool, the more likely an adequate match can be found for the treatment group members, assuming adequate common support. However, the improvements in percent bias reduction from 2:1 to 9:1 were modest (Rubin, 1979).

The interplay among total sample size, treatment group sample size, and comparison-to-treatment group ratio is complex. Although researchers have examined these issues, the examination has not been systematic. As such, there are no clear guidelines for researchers and evaluators as they conduct propensity score matching studies. Additional research is needed to determine the appropriateness of propensity score matching with small *total* and *treatment* group sample sizes, as well as the *minimum* acceptable comparison-to-treatment group ratio. Thus, the current study examined the performance of the matching methods with different treatment group sample sizes and comparison-to-treatment group ratios.

Finally, researchers should consider which matching method to employ. The current study explored some of the most common matching techniques used in applied educational research: random sampling, nearest neighbor (with propensity scores and Mahalanobis distances), nearest neighbor with a caliper, and optimal matching (Austin, 2011a; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008). Although random sampling and nearest neighbor matching with Mahalanobis distances are not *propensity* score techniques, they are included because they are commonly used in practice and allow for the comparison of propensity score techniques to non-propensity score matching techniques.

It is worth noting that in recent years, there has been increased interest in other matching methods such as genetic matching (Diamond & Sekhon, 2013), stratification or subclassification (Rosenbaum & Rubin, 1984), and full matching (Gu & Rosenbaum, 1993; Hansen, 2004). These matching methods involve a different philosophical approach than nearest neighbor and optimal matching. These methods are used to estimate ATE, not ATT. As such, these techniques do not result in a matched comparison group; instead, comparison group members are weighted (e.g., all comparison group members are retained, but weighted according to the estimate obtain through the matching method). As such, discussion of these techniques is beyond the scope of the current study. A brief description of the matching methods included in this study is included below.

Random Sample. Random sampling involves randomly selecting a sample from the larger comparison group pool. Random sampling is not a matching technique, per se. Further, this technique does not involve propensity scores or covariates. Random

sampling is primarily used to create groups of comparable *size*, rather than to obtain comparable covariate distributions. For example, if a researcher has a treatment group that consists of 50 students and comparison pool greater than 50, equally-sized groups could be created by taking a random sample of 50 students from the comparison pool.

Given that covariates are not considered with random sampling, if the treatment group and comparison pool differ on the covariates, then the treatment group and randomly-sampled comparison group will also differ on the covariates (Rosenbaum & Rubin, 1985). That is, the randomly sampled comparison group will typically resemble the full comparison pool on the covariates. Creating groups of equal size can help researchers meet certain assumptions of outcome analyses; however, this technique does not control for selection bias. If there is uncontrolled selection-bias, the treatment effect estimates will also be biased. Although random sampling is not recommended, this technique was included as another point of comparison because of its prevalent use in applied practice.

Nearest Neighbor. One approach to creating a matched comparison group is to use the nearest neighbor matching method. Nearest neighbor uses a greedy algorithm to sequentially match each treatment group member to the closest available comparison pool member (Gu & Rosenbaum, 1993; Stuart, 2010; Stuart & Rubin, 2008). Nearest neighbor matching can be used with propensity scores or Mahalanobis distances, and this study examined the performance of the nearest neighbor matching method with both propensity score and Mahalanobis distance measures. The formula for nearest neighbor matching is provided in Equation 6 (Pan & Bai, 2015, p. 7),

$$d(i,j) = min_j\{\left|e(X_i) - e\left(X_j\right)\right|\} \quad (6)$$

where $d(i, j)$ is the distance (e.g. propensity score or Mahalanobis distance, depending on the selected distance measure) between treatment group member i and comparison group member j and $min_j\{\left|e(X_i) - e\left(X_j\right)\right|\}$ results in the selection of the comparison group member with the minimum absolute difference in the distance measure for treatment group member i.

Conceptually, nearest neighbor matching involves similar steps regardless of which distance measure is used. For propensity score matching, the nearest neighbor matching method starts with a treatment group member and selects the comparison pool member with the closest absolute difference between propensity scores (as defined in Equation 6). Once the match is created, the algorithm proceeds sequentially in the same manner until all treatment group members have been matched. For Mahalanobis distance, the matching algorithm starts with a treatment group member and calculates the Mahalanobis distance between that treatment group member and every comparison pool member on the vector of covariates. The comparison pool member with the minimum distance is chosen as the match for the treatment group member (i.e., Equation 6), and both are removed from the matching pool. This is repeated until all treatment group members have been matched (Guo & Fraser, 2015). It is important to note that even though Mahalanobis distances are computed for each pair of treatment and comparison group members, current software packages do not save or report this information.

The starting point for the matching algorithm varies depending on the software package used. The default in the MatchIt package matches treatment group members with the largest distance measures first and those with the smallest distance measures matched last (Ho et al., 2011). Research has shown minimal differences in the

performance of the greedy algorithm for different orderings of treatment group members for propensity score matching (e.g., high to low vs. low to high propensity scores; Austin, 2013). Regardless of distance measure, the nearest neighbor matching algorithm does not re-evaluate matches once a match has been made (i.e., a subsequent member could have a closer match, but the nearest neighbor algorithm will not adjust the matches; Stuart, 2010).

Nearest neighbor matching has been described as the most "straightforward" (Caliendo & Kopeinig, 2008, p. 41; Schuler, 2015) and understandable matching technique. Unsurprisingly, nearest neighbor matching is one of the more commonly used matching algorithms (Austin, 2007a, 2009b). However, the use of nearest neighbor matching can result in poor quality matches (Smith, 1997). As noted, nearest neighbor with the greedy algorithm selects the *best* available match. That does not necessarily mean that the absolute difference between the propensity scores is small; it just means that it is the smallest out of the available matches.

Nearest Neighbor with caliper. Given that nearest neighbor matching may not minimize the absolute difference between treatment and comparison group members on the distance measure (propensity scores or Mahalanobis distances), researchers often specify a caliper when conducting nearest neighbor matching (Austin, 2011a; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008). A caliper limits the maximum distance allowed for creating matches on the metric of the distance measure. The formula for nearest neighbor matching with calipers is provided in Equation 7 (Pan & Bai, 2015, p. 7),

$$d(i,j) = min_j\{\left|e(X_i) - e\left(X_j\right)\right| < b\} \quad (7)$$

where $d(i, j)$ is the distance (e.g. propensity score or Mahalanobis distance, depending on the selected distance measure) between treatment group member *i* and comparison group member *j* and $min_j\{|e(X_i) - e(X_j)|\}$ specifies, for treatment group member *i*, select the comparison group member that results in the minimum absolute difference in the distance measure *but only if* the absolute difference is less than *b*, where *b* is the research-specified distance.

The matching process employs the same algorithm as nearest neighbor matching; however, treatment group members are only matched to comparison pool members if the propensity scores are within the specified caliper distance (typically in standard deviation units of the logit of the propensity score). Calipers tend to result in matches that are more similar on the covariates (e.g., better quality matches) than nearest neighbor matching. However, the use of calipers could result in decreased sample size, as unmatched treatment group members are excluded from the matched data sets if there are no matches within the specified distance (Austin, 2013; Jacovidis et al., in press; Stuart, 2010).

Calipers can be applied to almost any matching method; however, it is most common to use calipers with nearest neighbor matching (Austin, 2009b, 2011a; Stuart, 2010). Further, the current study only applied calipers to nearest neighbor matching on the propensity score distance measure. Although researchers can employ the caliper of their choice, results from simulation studies have suggested that calipers of 0.2 standard deviations of the logit of the propensity score, or 0.02 or 0.03 standard deviations of the propensity score are preferred (Austin, 2009b, 2010a). The current study examined the performance of nearest neighbor matching (using propensity scores as the distance measure) with three different calipers: 1) the recommended caliper of 0.2 standard

deviations of the logit of the propensity score, 2) a more liberal caliper of 0.3 standard deviations of the logit of the propensity score and 3) a more stringent caliper of 0.1 standard deviations of the logit of the propensity score.

Optimal matching. Another approach to creating a matched comparison group is to use the optimal matching method, which employs an optimal algorithm (Gu & Rosenbaum, 1993; Guo & Fraser, 2015; Ho et al., 2007, 2011; Stuart, 2010). Optimal matching considers the overall set of matches when choosing individual matches, with the goal of minimizing the global distance measure (Rosenbaum, 2002). The optimal matching method starts with a treatment group member and selects the comparison pool member with the closest absolute difference between propensity scores. Once a match is created, the optimal algorithm proceeds to the next match. However, optimal matching evaluates the total distance between matched groups at each step and may alter earlier matching decisions, if the change will yield the smallest average absolute distance across all matched pairs (Gu & Rosenbaum, 1993; Guo & Fraser, 2015; Ho et al., 2007, 2011; Pan & Bai, 2015; Stuart, 2010). That is, matched pairs made earlier in the process may be modified at later stages if the modification will minimize the overall distance between the matched groups. Thus, the key distinction between the greedy algorithm used in nearest neighbor matching and the optimal algorithm used in optimal matching is whether or not the matches are re-evaluated and modified throughout the matching process.

If researchers have a large reservoir of potential comparison group members, it is not uncommon for researchers to use optimal matching to create matches with a 2 to 1 ratio – that is, two comparison group members are matched to every one treatment group member (Smith 1997; Stuart, 2010). Further, some researchers have recommended

matching each treatment group member to two comparison group members (2:1), suggesting that the 2:1 match is more efficient (Haviland, Nagin, & Rosenbaum, 2007). Conversely, other researchers argue that selecting multiple comparison group members could result in unbalanced groups, as the second, third, or forth closest matches are less similar to the treatment group member than the first closest match (Stuart, 2010). This study examined the performance of one comparison to one treatment group member (optimal 1:1) and two comparison to one treatment group members (optimal 2:1).

Step 4: Create Matched Groups. The next step is to create the matched comparison group. There are a number of software options available to perform matching including SPSS, SAS, STATA, and R (R Core Team, 2016). The MatchIt (Ho et al., 2011) package in R is one of the most comprehensive matching packages available and can implement a wide variety of matching methods (Schuler, 2015). Among other matching methods, the MatchIt (Ho et al., 2011) package can be used to conduct nearest neighbor, nearest neighbor with caliper, optimal, and Mahalanobis distance matching. Thus, the current study used the MatchIt (Ho et al., 2011) package in R. For more information on the software available to perform propensity score matching, see Schuler (2015).

Step 5: Diagnose Matches. The purpose of matching is to balance the distributions of the covariates for the treatment and matched comparison groups. As such, it is paramount that researchers compare propensity scores and covariates across groups to ensure that the groups are properly balanced. Recall that propensity scores are the probability of treatment, given a set of covariates (Luellen et al., 2005) and one assumption of propensity score matching is that all relevant covariates have been

included in the matching model. Thus, after controlling for the covariates, assignment to the treatment group is essentially random and each individual has the same probability of treatment (e.g., propensity scores; Rosenbaum & Rubin, 1983). However, if after matching, the propensity scores are not balanced, then the propensity score model must be misspecified (Diamond & Sekhon, 2013). That is, the matching model does not include all relevant covariates.

Diagnosing matches directly relates to Rubin's Causal Model (Rubin, 1974). If assumptions are met, propensity score matching results in an estimate of the counterfactual when the matched comparison group is similar to the treatment group on a set of covariates (Rosenbaum & Rubin, 1983). However, if assumptions are not met (e.g., the matching model is misspecified), then the comparison group is not similar to the treatment group on the covariates, and the comparison group cannot serve as an accurate estimate of the counterfactual (Diamond & Sekhon, 2013). Thus, it is important to diagnose the quality and quantity of matches.

Quality of Matches. Once matched sets have been created, researchers should evaluate the quality of their matches. This should include an examination of both propensity score balance and individual covariates balance. There are several approaches to assessing the quality of matches numerically and visually (Caliendo & Kopeinig, 2008; Pan & Bai, 2015; Stuart, 2010). The most commonly used approaches are described below.

Numeric Diagnosis of Balance. Numeric diagnosis of balance can be examined via 1) the standardized mean difference, 2) the variance ratio, and 3) the percent bias reduction. Each of these techniques can be used with propensity scores or the individual

covariates. The techniques are described below along with recommendations for what constitutes balance with each index.

Standardized mean difference. The standardized mean difference can be used to evaluate both propensity score balance and individual covariate balance (Austin, 2009a; Stuart, 2010). Equation 8 provides the computation the standardized mean difference for propensity scores and continuous covariates (Cohen's *d*; Austin, 2009a),

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{comparison})}{\sqrt{\frac{s^2_{treatment} + s^2_{comparison}}{2}}} \quad (8)$$

where $\bar{x}$ is the respective group mean on the propensity scores or the individual covariate and s^2 is the respective group variance of the propensity scores or the individual covariate. For propensity scores, the standardized mean difference should be close to zero (Austin, 2011b). For continuous covariates, the standardized mean difference should be less than 0.25 standard deviation units (Stuart, 2010; What Works Clearinghouse, 2014); however, the closer to zero, the better.

The standardized mean difference for categorical covariates is provided in Equation 9 (Austin, 2009a),

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{comparison})}{\sqrt{\frac{\hat{p}_{treatment}(1-\hat{p}_{treatment}) + \hat{p}_{comparison}(1-\hat{p}_{comparison})}{2}}} \quad (9)$$

where $\hat{p}$ is the respective group mean of the dichotomous categorical covariate (e.g., the proportion of individuals in the group coded 1). The standardized mean difference for categorical covariates should be less than 0.10 (Austin, 2009a); however, values closer to zero indicate better balance. Additionally, frequencies or odds ratios should be examined to determine whether the comparison group has over- or underrepresentation compared to the treatment group (Austin, 2009a).

Variance ratio. Another approach to assessing balance is to compare the variances of the propensity scores between groups (Stuart, 2010). Equation 10 displays the formula for calculating the variance ratio,

$$variance\ ratio = \frac{s^2_{treatment}}{s^2_{comparison}} \quad (10)$$

where s^2 is the respective group variance on the propensity scores. Ideally, the variance ratio should be close to one (Rubin, 2001; Stuart, 2010), indicating that the variances of the propensity scores between the two groups are about equal. Although it is more common to use this technique to assess propensity score balance, it can be used to assess individual covariate balance.

Percent bias reduction. Another approach to assessing balance is to examine the percent bias reduction (Bai, 2015; Cochran & Rubin, 1973; Pan & Bai, 2015). The percent bias reduction can be calculated using Equation 11,

$$PBR = \frac{bias_{before\ matching} - bias_{\ after\ matching}}{bias_{before\ matching}} * 100 \quad (11)$$

where bias is the difference between treatment and comparison group propensity scores before and after matching, respectively. Adequate percent balance reduction is typically considered 80% and above (Bai 2013; Cochran & Rubin, 1973). Percent balance reduction can be used to assess the balance of individual covariates and propensity scores.

Visual diagnosis of balance. In addition to numeric balance, there are several options to visually assess propensity score and individual covariate balance, including jitter graphs, cumulative density plots, quantile-quantile (QQ) plots, standardized difference (effect size) plots and histograms (Ho et al., 2007; Schuler, 2015; Stuart, 2010; Stuart & Rubin, 2008). Visual inspection involves a subjective decision from the

researchers on whether the groups are balanced. The current study did not include visual diagnosis of balance; however, these graphs can be easily created with the MatchIt package in R (Ho et al., 2011). For guidance on jitter graphs, QQ plots, standardized difference (effect size) plots, and histograms, see Ho, Imai, King, and Stuart (2011) or Schuler (2015), which provide step-by-step instructions. R code for creating cumulative density plots may be found in Harris and Horst (2016).

Quantity of Matches. As noted above, some matching methods will exclude unmatched treatment group members from the matched data sets. Further, it is possible that one matching technique will create closely matched groups (i.e., high quality matches), yet only maintain a portion of the original treatment group (i.e., low quantity matches). It is worth noting that the creation of quality matches should not come at the expense of decreased treatment group sample size and subsequent loss of information (Austin, 2013; Jacovidis et al., in press; Stuart, 2010). However, assessing the quantity of matches is straightforward. Researchers should simply examine the number of treatment group members (e.g., raw numbers, percentages, or proportions) who were able to be successfully matched. The researcher, then, needs to weigh the benefit of having closely balanced groups against the cost of losing information or sample size (Austin, 2013; Jacovidis et al., in press; Stuart, 2010). More importantly, if any treatment group members were dropped from the matched data sets, the researcher should ensure that this does not affect the representation of the treatment group (Austin, 2013; Jacovidis et al., in press; Rosenbaum & Rubin, 1985; Stuart, 2010). Loss of representation may limit the generalizability of the results to a subset of the treatment group, instead of the entire treatment group. For example, if all treatment group members from one racial group

were dropped from the matched data sets, then the researcher could no longer generalize the results back to that racial group. The current study examined the tradeoffs that arise between the quantity and quality of matches created and how that tradeoff influences the resulting outcomes analyses.

Step 6: Examine Group Differences on the Outcome. Researchers are typically interested in examining whether the treatment group differs from the comparison group on some outcome of interest. The previous steps ultimately help researchers get to the point where they can examine group differences on the outcome of interest. To avoid knowledge of the outcome influencing researchers' decisions throughout the propensity score matching process, it is best practice for the outcome variable to be merged onto the data set *after* matched groups are created and the quality of matches are evaluated. If researchers are interested in examining the relationship between the covariates and the outcome variable, this should be done after groups are matched to maintain alignment with best practices (Stuart & Rubin, 2008).

As noted previously, there are two indices that are frequently used to estimate the average treatment effects in propensity score matching: the ATE and the ATT (Caliendo & Kopeinig, 2008). In education research and evaluation studies, researchers are typically interested in the ATT. Further, the ATT is the approach that is typically taken when propensity score matching is used to create matched groups. Given that the current study focused on comparing matching methods, the ATT was estimated as an index of the average treatment effect.

Once the matched group is created, the outcomes analyses to estimate ATT are straightforward. Researchers compare the treatment and matched comparison group by

conducting the appropriate inferential tests dictated by the research question of interest (Caliendo & Kopeinig, 2008; Gu & Rosenbaum, 1993; Ho et al., 2007; Stuart, 2010; Stuart & Rubin, 2008). However, it has been recommended in the propensity score literature that any covariates included in the matching model that remain unbalanced after matching should be included in the outcome analyses (Pan & Bai, 2015; Rosenbaum & Rubin, 1985). Further, including unbalanced covariates in the outcome analyses has been shown to produce accurate estimates of treatment effects regardless of the choice of propensity score matching methods (Schafer & Kang, 2008; Shadish et al., 2008). However, this technique does not appear to be a recommendation that researchers use very often in applied practice (e.g., Clark & Cundiff, 2011; Lu, Zanutto, Hornik, & Rosenbaum, 2001; Morgan, Frisco, Farkas & Hibel, 2010; Olitsky, 2013). As such, the current study included two outcome analyses to estimate the ATT: regression with no covariates and regression with unbalanced covariates.

Research Comparing Matching Methods

Given that matching methods employ different distance measures (i.e., propensity scores or Mahalanobis distances), matching algorithms (i.e., greedy or optimal), and rules for comparison group member selection, each technique could potentially select different comparison group members from the overall comparison pool to create the matched comparison groups. Moreover, matched comparison group composition could vary considerably depending on the matching algorithm used. This will not only affect the quality of matches, but the selection of matching method may also affect the results of any outcome analyses.

There are a few notable studies that have examined the performance of matching methods under various conditions. Most of these studies assessed performance in terms of how well the matching methods were able to balance the groups on the distance measure and the covariates or selection bias reduction. Only a few studies have extended the evaluation of the matching methods to the impact it has on the outcome analyses (Austin, 2013; Jacovidis et al., in press; Stone & Tang, 2013). However, researchers have noted that there are few studies that have been conducted to systematically examine which propensity score matching methods perform well under which data conditions (Austin, 2013; Bai, 2015).

Again, it is worth noting that a discussion of matching techniques that result in a weighted comparison group is beyond the scope of the current study. Thus, the comparison of matching methods focused solely on those that result in a matched comparison group (e.g., nearest neighbor matching *with* and *without* calipers and optimal matching). Research on distance measures and matching algorithms has been discussed previously; however, they are summarized here for convenience.

The differences among the matching methods are largely a result of the distance measure and matching algorithm employed. Recall that Mahalanobis distances equally weights all covariates, while propensity sores weight covariates by how well they predict group membership. Simulation studies have shown that when there are a small number of covariates (e.g., 2 to 8), the two distance measures result in comparable balance (Gu & Rosenbaum, 1993; Zhao, 2004); however, when there are a large number of covariates (e.g., 20), propensity scores result in better balanced group than Mahalanobis distances (Gu & Rosenbaum, 1993).

Also, recall that the key distinction between the greedy and optimal algorithms is whether or not the matches are re-evaluated and modified throughout the matching process. Research has shown that greedy and optimal matching approaches generally result in selection of the same comparison group members from the overall comparison group. Consequently, greedy and optimal matching approaches perform comparably in creating groups with balanced covariates (Austin, 2009b, 2013; Bai, 2013; Gu & Rosenbaum, 1993). However, when treatment group members compete for comparison group members, the optimal algorithm outperforms the greedy algorithm (Gu & Rosenbaum, 1993). Further, optimal matching performs better at reducing the distance between matching *pairs* (e.g., a direct comparison-to-treatment group member match). Thus, if the researcher is interested in well-matched *pairs*, instead of just well-matched *groups*, then optimal matching may be preferable (Gu & Rosenbaum, 1993; Schuler, 2015; Stuart, 2010).

In one study, researchers examined how well the greedy and optimal algorithms recover simulated treatment group effects ($d = 0.2$). The algorithms performed similarly in recovery of the treatment effect; however, optimal performed slightly better than the greedy algorithm with the smallest treatment group sample size ($n = 30$). Further, power was low, but comparable between the two algorithms. Low power is unsurprising given the small sample sizes ($n = 30$ and 60) and small effect size ($d = 0.2$; Stone & Tang, 2013).

There is contrary evidence on the number of comparison group members to be matched to each treatment group member (e.g., Haviland et al., 2007; Stuart, 2010) when the optimal algorithm is used. It is important to note that the quality of matches obtained

when matching multiple comparison group members to each treatment group member will likely depend on common support. That is, if there is sufficient overlap in the propensity scores between the two groups, then selecting multiple matches should still result in balanced groups; however, if there is *not* sufficient overlap in the propensity scores between the two groups, then selecting multiple matches should result in unbalanced groups, as subsequent matches are likely to be less similar than the first match that this made (Stuart, 2010).

Researchers have consistently shown that covariates and propensity scores are more balanced and percent bias reduction is greater when nearest neighbor matching with calipers is employed, compared to nearest neighbor matching without calipers (Austin, 2009b, 2013; Bai, 2015; Jacovidis et al., in press; Rosenbaum & Rubin, 1985). Further, the more stringent the caliper, the better balance between the matched groups (Austin, 2009b, 2010a; Dehejia & Wahba, 2002; Jacovidis, in press). Moreover, researchers have noted that the order in which treatment group members are matched does not affect the performance of nearest neighbor matching with and without calipers (Austin, 2013).

As noted previously, calipers can be applied when Mahalanobis distances are used as the distance measure. Mahalanobis distances matching with calipers resulted in better quality matches than were obtained without calipers (Bai, 2013; Rosenbaum & Rubin, 1985). However, it is worth noting that in one of the studies (Rosenbaum & Rubin, 1985), propensity scores were included as a variable in the calculation of the Mahalanobis distances. This is a key distinction as propensity scores weight the covariates and Mahalanobis distances balance covariates equally. Thus, if propensity

scores are used in the calculation of Mahalanobis distances, then covariates are no longer *equally* weighted.

Conversely, nearest neighbor matching with calipers also results in a loss of treatment group members (e.g., Austin, 2009b, 2013; Bai, 2015; Jacovidis et al., in press). Moreover, as the caliper becomes more stringent, the loss in treatment group members is greater (e.g., Austin, 2009b, 2013; Dehejia & Wahba, 2002; Jacovidis et al, in press). For example, in one simulation study where common support was manipulated, the nearest neighbor matching with a 0.25 caliper resulted in only 35% to 55% of the treatment group members being retained. That is, 45% to 65% of the treatment group members could not be matched (Bai, 2015). Additionally, in an applied study, 81% to 84% of the treatment group was retained when calipers of 0.1 to 0.3 were applied. Although, the majority of the treatment group members were retained, there was substantial loss of minority representation in the treatment group (Jacovidis et al., in press).

It is up to the researcher to balance the quality and quantity of matches when creating a matched comparison group. This is a difficult task. Obviously, researchers want to balance groups on the covariates, while also maintaining the treatment group sample size. However, this may not always be possible. If researchers are concerned with equity and representativeness (e.g., generalizability), they may wish to choose a matching technique that does not compromise quantity (e.g., nearest neighbor, optimal). However, as noted above, matching techniques that select the closest available match may still result in unbalanced groups. Additionally, there may be expectations from funding agencies that require close balance (quality). Thus, researchers may choose a matching technique that does not compromise quality (e.g., caliper matching). Further,

when groups are not balanced, the comparison group may not be a viable estimate of the counterfactual. However, applying a caliper may result in a decrease in the treatment group sample size, as treatment group members who do not have an adequate match are excluded from the matched data set. When treatment group members are excluded, it is important for researchers to examine the representativeness of the samples. For example, if the matching procedure results in the loss of a minority group representation, then researchers are no longer generalizing the findings back to the same population (Austin, 2013; Jacovidis et al., in press; Rosenbaum & Rubin, 1985). However, because the outcome variable of interest is not used in the matching procedure, any number of matching methods can be evaluated. Researchers can then select the matching technique that results in the best balance (Ho et al., 2011).

As noted, few researchers have extended the examination of matching method performance to include outcome analyses (Austin, 2013; Jacovidis et al., in press; Stone & Tang, 2013). In a simulation study comparing 12 matching methods (different variations of nearest neighbor matching with and without calipers and optimal matching), caliper matching resulted in more accurate estimates of the simulated treatment effect (d = -0.02) than the nearest neighbor matching without calipers and optimal matching (Austin, 2013). In an applied study, it was demonstrated that different decisions could be made about whether there was a statistically significant difference between groups, depending on which matching method was used; however the true difference between the treatment and comparison groups was unknown (Jacovidis et al., in press). Thus, it was difficult to know which matching technique performed best (Jacovidis et al., in press).

There are a number of limitations regarding the matching methods literature that are necessary to address. First, many of the studies examining propensity score matching methods have focused on comparing matched comparison group methods used to obtain ATT estimates with weighted comparison group methods used to obtain ATE (e.g., Austin, 2007b; Austin, Grootendorst, Normand, & Anderson, 2007; Austin & Schuster, 2016; Harder et al., 2010). Although this is a much needed line of research, it is not particularly helpful for researchers who are interested in choosing among matching methods used to obtain ATT estimates. Second, the studies focusing on comparing matching methods used to obtain ATT estimates have not been systematic. Given the disorganization in the research, it is difficult to make a cohesive case for the use of specific matching methods under specific conditions. Third, most studies examining matching methods compare propensity score and covariate balance and bias reduction, but do not include an examination of outcome analyses. Fourth, in many studies, applied data are used. Although this is not problematic when evaluating balance, it is difficult to examine treatment effects, as true group differences are unknown.

Finally, in simulation studies, the data are not simulated realistically. For example, covariates are simulated to be all continuous or all binary (e.g., Austin, 2011a, 2013), when in reality, most researchers use a combination of the two. Perhaps certain matching methods perform better with certain kinds of covariates (e.g., maybe it is easier to match binary covariates than it is to match continuous covariates). Further, the covariates in simulation studies are often simulated to be independent (e.g., Austin, 2011a, 2013); however, in social sciences, constructs are rarely independent. Perhaps

matching methods perform better when covariates are independent (e.g., collinearity in regression). Both of these issues are empirical questions that researchers could explore.

Some researchers have noted that selecting a matching method is less important than selecting the covariates used in the propensity scores matching model (Steiner et al., 2010). However, this has led to the current mindset in propensity score matching: as long as the researcher has selected appropriate covariates, the matching method does not matter. Although matching method may be *less* important than covariate selection, given the results of the studies comparing matching methods, it does not appear that matching method is of no concern.

The Current Study

As described above, there are a number of decisions that researchers make at each step in the propensity score matching process. This study focused mainly on the decisions related to the selection of the matching method. Matching methods employ different distance measures (i.e., propensity scores or Mahalanobis distances), matching algorithms (i.e., greedy or optimal), and rules for comparison group member selection. Thus, each technique could result in matched comparisons groups that vary considerably depending on the matching algorithm used. Selection of matching method not only affects the quality of matches, but may also affect the results of any outcome analyses. Further, there are key limitations to the current matching method literature that make it difficult to recommend the use of specific matching methods under specific conditions. Clearly, one study is not going to be able to address all of the current limitations, but a collective, more coherent program of research is needed to provide guidance to practitioners on which matching methods perform the best under which conditions. This

study was one in a line of research on matching method performance. Thus, the purpose of this study was to examine and compare common matching techniques used to estimate ATT. Specifically, the current study addressed four research questions.

1. How do the most common matching methods differ, in terms of quantity (i.e., number of matches) and quality (i.e., covariate balance) of matches?
2. Once matched comparison groups are formed, how do the results of group comparisons (e.g., significance tests) compare across the different matching methods?
3. How well do the various matching methods recover the true treatment effect (e.g., difference between the group means)?
4. What conditions (e.g., true difference between the means, matching method, comparison-to-treatment ratio, sample size, and outcome analysis) are optimal in obtaining accurate estimates of parameters?

CHAPTER 3

Method

In the current study, data were simulated to empirically investigate the performance of common matching methods under known and systematically manipulated conditions. Because population parameters, such as differences between group means, are not known in applied studies, simulation is needed to compare the accuracy of matching methods. Applied studies can show that matching methods yield different estimates; however, they cannot show which matching method produces the most accurate estimates. Although simulation studies can never completely capture the complexities of real data situations, the utility of simulation results are dependent on the representativeness of the conditions that are being modeled. If the conditions are not similar to those found in real data, the utility of the study is limited. The current study focused on comparing matching methods under manipulated conditions that were representative of program evaluation and effectiveness studies.

Data Generation

Data were simulated to reflect values found in higher education, using a recent study by Jacovidis and her colleagues (in press). In the Jacovidis et al., (in press) study, the data were gathered from 3,287 undergraduate first-year students from a public university in the mid-Atlantic US, which included 3,201 comparison group members and 86 treatment group members. The study by Jacovidis and her colleagues (in press) focused on group differences between the treatment group and a matched comparison group on an information literacy test that was administered to all first-year students. The treatment group consisted of a subset of the population that has historically

underperformed on the information literacy test and had received targeted interventions in recent years. Jacovidis and her colleagues (in press) noted that a discussion of the nature of the treatment was intentionally omitted from their article, as the study focused on comparing matching methods, rather than on the impact of treatment program.

The study by Jacovidis and her colleagues (in press) included 12 covariates: 4 continuous covariates (SAT math, SAT verbal, conscientiousness, and work avoidance) and 8 categorical covariates (gender and race/ethnicity). The measure of conscientiousness, work avoidance, and information literacy was completed in a secure, proctored environment, which helped to provide a standardized testing experience for all students completing the measure. Demographic information and SAT scores were retrieved from student records. The covariates were selected using a similar method to that used in applied practice. That is, of the variables that the researchers had available, they chose the ones that best aligned with theory and previous research (Jacovidis et al., in press). Following the procedures recommended by Stuart and Rubin (2008), the relationships between the covariates and group membership or the outcome of interest were not examined prior to matching. However, after matched groups were created, Jacovidis and her colleagues (in press) examined those relationships to ensure that the selected covariates were related to both group selection and the outcome of interest.

In the current study, 6 covariates (4 continuous covariates and 2 categorical covariates) were simulated based on the data presented by Jacovidis and her colleagues (in press). The four continuous covariates represented SAT math (*X1*), SAT verbal (*X2*), work avoidance (*X3*), and conscientiousness (*X4*). The two categorical covariates represented gender (*X5*) and race/ethnicity (*X6*).

The correlations among the covariates, and between the covariates and the outcome, were calculated for the real data. For the categorical covariates, the correlations were dependent on the proportion within each group. For example, the expected correlation between X6 and X1 should increase if the split on X6 was change from 5%-95% to 15%-85%. Thus, for any correlation between a categorical covariate and a continuous covariate, the correlation was replaced with the biserial correlation. Similarly, the correlation between X5 and X6 was replaced with the tetrachoric correlation. These correlations are shown in Table 1. The resulting regression coefficients (predicting either the propensity scores or the outcome) represent the coefficients from a probit regression. These covariances and regression coefficients were then used to simulate the data for the current study.

Table 1

Generating Variances and Covariances

	X1	X2	X3	X4	X5	X6
X1	1.0000					
X2	0.4300	1.0000				
X3	-0.1220	-0.1160	1.0000			
X4	0.1520	0.0940	-0.3720	1.0000		
X5	-0.3297	-0.0984	0.2121	-0.3054	1.0000	
X6	-0.4058	-0.2664	0.0143	-0.0389	-0.0260	1.0000

All data were simulated using R version 3.3.2 (R Core Team, 2016). First, six continuous covariates were simulated.[1] All covariates were drawn from a multivariate normal distribution with means of 0 and variance-covariance displayed in Table 1 using the RMVNORM function in the MTVNORM package (Genz et al., 2016).

To obtain group assignment, the underlying likelihood of treatment group membership in probits was simulated as a function of the six continuous covariates and a

random error term drawn from a standard normal distribution. Equation 12 specifies the equation used to obtain probits, with coefficients obtained using the correlations in Table 1 and the correlation between the covariates and group membership in the data from Jacovidis et al. (in press),[2]

$$P(x_i) = -0.015(X1_i) - 0.301(X2_i) + 0.088(X3_i) + 0.084(X4_i) \\ -0.117(X5_i) + 0.308(X6_i) + e_i \quad (12)$$

where $P(x_i)$ is the underlying likelihood of treatment group membership (e.g., probit) based on the covariates for person *i*, $X1_i$-$X6_i$ represent person *i*'s scores on the covariates, and $e_i \sim N(0,1)$. Then, simulees were assigned to treatment and comparison groups using a cut point based on the percentiles corresponding to the proportion of the sample assigned to the treatment group, which varied by condition. In later analyses, X5 and X6 were not used in their continuous form. Specifically, the two covariates representing gender and race (*X5* and *X6* in Table 1, respectively) were dichotomized such that *X5* (gender) was split 60%-40% and *X6* (race/ethnicity) was split 15%-85%. The cut point was determined based on the *z*-score corresponding to 60% (for *X5*) or 15% (for *X6*) in a cumulative normal distribution.[3] This resulted in four continuous covariates (*X1*-*X4*) and two dichotomous covariates (*X5*-*X6*).

Next, outcome scores were simulated as a function of group membership, six covariates,[4] and a random term representing unexplained variance and error. The coefficients in Equation 13 were based on the relationship between the covariates and information literacy scores in the data from Jacovidis et al. (in press),

$$Y_i = d(Group_i) + 0.158(X1_i) + 0.418(X2_i) + 0.049(X3_i) + \\ 0.029(X4_i) + 0.087(X5_i) - 0.035(X6_i) + e_i \quad (13)$$

where Y_i is the simulated outcome score based on the covariates for person *i*, *d* represents the simulated effect size (specified at one of four levels, described further below), $Group_i$ is the group membership for person *i*, $X1_i$-$X6_i$ represent person *i*'s scores on the four continuous and two dichotomous covariates, and e_i ~N(0, 0.7401388). Finally, *Y* was standardized. This was done so that the within-group standard deviation was one.[5] With the pooled within-group standard deviation of Y set to one in the population, the difference between the means was on the Cohen's *d* metric. Appendix A includes the syntax used to simulate and analyze the data.

Conditions

In the current study, five factors were manipulated: effect size, matching method, comparison-to-treatment ratio, treatment group sample size, and type of outcome analysis. A description of each of the manipulations and rationale for the selected conditions are provided below.

Effect size. The true effect size was systematically manipulated at four levels: 0.0, 0.2, 0.5, and 0.8. These values align with the effect size benchmarks suggested by Cohen (1988) for small, medium, and large effects. Although Cohen's benchmarks have become the standard in interpreting the magnitude of effect sizes, some researchers (e.g., Hill, Bloom, Black, & Lipsey, 2008) have suggested that the magnitude of the effect size should be interpreted based on the research or evaluation context, as Cohen originally urged. Thus, what may be viewed as a small effect size for one context, can be viewed as a large effect size in another context.

Hill and her colleagues (2008) provided three empirical benchmarks that consider the research and evaluation context specific to achievement. The first empirical

benchmark relied on the expectations for growth over time. The researchers examined seven nationally normed reading tests and six nationally normed math tests across elementary and secondary grades. Standardized effect sizes (i.e., Cohen's *d*) between grades ranged from 0.00 to 1.03, with the effect sizes decreasing from first to twelfth grade. The second empirical benchmark involved examining demographic group or school performance differences. Thus, the researchers examined reading and math differences by gender, race/ethnicity, socioeconomic status, and school performance. Standardized effect sizes between groups ranged from 0.04 to 1.04. The third empirical benchmark involved comparing the observed effect sizes to effect size results from past research for similar interventions and target populations. Hill and her colleagues (2008) presented a summary of student achievement effect sizes for random assignment studies of educational interventions by elementary, middle, and high school. The mean effect sizes ranged from 0.07 to 0.51 (Hill et al., 2008). Contextually, these effect sizes are particularly relevant for this study given the focus on student achievement. Moreover, the range of effect sizes for the three empirical benchmarks are not substantially different than the range of benchmarks suggested by Cohen (1988).

The study by Jacovidis and her colleagues (in press) provided further context for expected effect sizes for the current study. In the original study, the observed effect sizes ranged from 0.25 to 0.76, across the total sample and the various matched groups. Thus, it seems reasonable that the current study should examine effect sizes close to these values. It is also worth noting that these effect sizes align with Cohen's benchmarks (1988) and the student achievement empirical benchmarks of Hill and her colleagues (2008).

One goal of the current study was to generalize beyond student achievement and information literacy interventions to a larger context of program evaluation studies. Thus, it is important to ensure that the manipulated effect sizes are typical for a variety of program evaluation contexts. In an extensive review of 302 meta-analyses across a range of psychological, educational, and behavioral interventions, researchers found a mean effect size of 0.50 (*SD* = 0.29; Lipsey, 2002). Accordingly, the mean effect size reported by Lipsey (2002) aligned with Cohen's benchmark for a medium effect, one standard deviation above the mean aligned with Cohen's benchmark for a large effect, and one standard deviation below the mean effect size aligned with Cohen's benchmark for a small effect.

Regardless of whether Cohen's benchmarks *should* be interpreted as small, medium, and large, the range of the benchmarks represent the magnitude of effect sizes observed in student achievement, the specific context of information literacy in first-year collect students, and the broader context of program evaluation. Thus, effect sizes for this study were manipulated to be 0.0, 0.2, 0.5, and 0.8. The effect size was defined as the mean difference (between treatment and comparison) in the outcome, divided by the pooled within-group standard deviation.

Matching method. Eight matching methods were used to create matched comparison groups for the treatment group: random sampling, nearest neighbor (using the default order of matching treatment group members with the highest propensity scores first and those with the lowest propensity scores matched last; Ho et al., 2011), nearest neighbor with calipers (0.3, 0.2, and 0.1 times the standard deviation of the propensity

scores), optimal (1:1 and 2:1 ratios), and Mahalanobis distance matching without calipers. All matching was conducted *without* replacement.

Estimated propensity scores computed via logistic regression served as the distance measures for all of the propensity score techniques. Propensity scores represent the probability of participation, given the set of covariates. It is important to note that random sampling and Mahalanobis distance matching are not *propensity* score techniques. The random sampling technique was included because it is commonly used in practice and allowed for a comparison of the propensity score techniques to a more traditional technique. Mahalanobis distance matching was included because it is advantageous over propensity score matching in certain situations (e.g., King & Nielsen, 2016; Zhao, 2004). The MatchIt (Ho et al., 2011) R package was used to conduct nearest neighbor, nearest neighbor with caliper, optimal, and Mahalanobis distance matching. Additional R code was written for random sampling, as this technique is not offered via the MatchIt package.

Comparison-to-treatment ratio. The ratio of comparison group members to treatment group members before matching was manipulated at four levels: 3:1, 4:1, 5:1, and 6:1. The selected ratios were not meant to be exhaustive; they were meant to serve as a starting point. Although there is some research on comparison-to-treatment ratios, there is little practical guidance in the literature on how much larger the comparison pool needs to be than the treatment group. That is, there is no consensus on the *minimum* ratio of comparison-to-treatment group members. Researchers recommend that the larger the comparison pool, the better (Bai, 2015; Rubin, 1979). This makes sense intuitively—the

larger the comparison pool, the more likely an adequate match can be found for the treatment group members, assuming adequate common support.

It also seems reasonable that there would be a point of diminishing returns. Thus, it may not be worthwhile to increase the comparison pool past a certain point. Moreover, increasing the comparison pool may be cost prohibitive if the researcher has to collect covariate data rather than or in addition to using extant data, especially if the researcher is using a proprietary measure. For example, suppose a researcher is evaluating a retention intervention for students who have low institutional commitment. The researcher should match students on variables related to institutional commitment. It is unlikely that this information is already being collected by the institution and will likely need to be collected by the researcher. Further, say the researcher uses a proprietary measure that costs $5 per administration to obtain the covariates of interest. If the treatment group was composed of 100 students, then the cost of administering the measure to the potential comparison group would range from $1500 (3:1 ratio) to $3000 (6:1 ratio) depending on the comparison-to-treatment group ratio. Thus, the researcher would spend *twice* as much obtaining a 6:1 ratio than obtaining a 3:1 ratio. Moreover, this is based on the recommendation that larger comparison pools are better than smaller comparison pools, even though the improvements in percent bias reduction from 2:1 to 9:1 is modest (Rubin, 1979).

In sum, the comparison-to-treatment ratio was manipulated at four levels (3:1, 4:1, 5:1, and 6:1). The inclusion of optimal 2:1 matching required a minimum ratio of 3:1 to avoid matching every comparison group member to a treatment group member. That is, the 2:1 ratio was not included because it would result in the selection of the full

comparison pool when the optimal 2:1 matching method was used. The ratio was increased incrementally for the remaining ratios. Again, these ratios are meant to serve as a starting point and additional ratios may be necessary in future studies.

Treatment group sample size. Sample size of the treatment group was manipulated at two levels: 30 and 100. Given the comparison-to-treatment ratio was also manipulated, this resulted in the generation of 90 to 600 comparison group members. The total sample size varied based on the sample size of the treatment group and the comparison-to-treatment ratio. The comparison pool sample size for each treatment group sample size and comparison-to-treatment ratio is presented in Table 2.

Table 2

Comparison Pool Sample Size by Treatment Group Sample Size and Comparison-to-Treatment Ratio

Treatment Sample Size	Comparison-to-Treatment Ratio			
	3:1	4:1	5:1	6:1
30	90	120	150	180
100	300	400	500	600

Sample size was examined because propensity score matching was developed as a large sample size technique; however, it has been applied in small sample situations (e.g., small-scale program evaluations; Stone & Tang 2013). For example, the National Science Foundation's *Handbook for Mixed Method Evaluations* provides an illustrative example of a program evaluation for education researchers on using mixed method approaches in their evaluation design. The example describes an undergraduate faculty enhancement program focusing on preservice mathematics. The two-year intervention involves workshops throughout the academic year, summer sessions, demonstrations of model teaching, and individual coaching; it is designed to serve 25 faculty members

(National Science Foundation, 1997). Larger samples sizes are often not practical when programs offer more direct and intensive services.

In a propensity score matching study, Stone and Tang (2013) manipulated treatment group samples sizes at 30 and 60 stating that these values were "chosen to represent smaller treatment group sizes that are consistent with typical educational program evaluations," (p. 4). In a meta-analysis of randomized and quasi-experiments evaluating education programs, almost 30% of the reviewed studies (published and unpublished or "gray literature") involved sample sizes below 100 (Cheung & Slavin, 2016). Further, there are mixed perspectives on whether propensity score matching should be used with smaller sample sizes (e.g. Bai, 2015; Dehejia & Wahba, 2002; Rubin, 1979, 1997; Stone & Tang, 2013). Thus, additional research is needed to determine the appropriateness of propensity score matching with small treatment group sample sizes. The two sample sizes investigated in this study were selected to represent a small sample size (30) that might be seen in a small-scale program evaluation study with one cohort and a larger sample size (100) that might be more characteristic of a program evaluation study with multiple cohorts.

Outcome Analyses. The outcome analyses were manipulated at two levels: regression with group membership predicting the outcome variable and regression with group membership and any unbalanced covariates predicting the outcome variable. Recommendations in the propensity score literature are that when conducting the outcome analyses, any covariates included in the matching model that remain unbalanced after matching should be included as predictors in the model (Rosenbaum & Rubin, 1985). However, this does not appear to be a recommendation that researchers use in

practice (e.g., Clark & Cundiff, 2011; Lu, Zanutto, Hornik, & Rosenbaum, 2001; Morgan, Frisco, Farkas & Hibel, 2010; Olitsky, 2013). Thus, the two approaches to outcome analyses were examined to determine which approach produces more accurate estimates of the group differences. Equation 14 displays the regression equation for group membership predicting the outcome variable,

$$\widehat{Y}_i = \alpha + \beta_1(Group_i) \tag{14}$$

where $\hat{Y}_i$ is the predicted outcome score for person i, α is the predicted outcome score for the comparison group, β_1 is the unstandardized regression coefficient associated with group, and $Group_i$ represents group membership for person i. Equation 15 displays the regression equation for group membership and any unbalanced covariates predicting the outcome variable,

$$\widehat{Y}_i = \alpha + \beta_1(Group_i) + \sum \beta_k(X_{ki}) \tag{15}$$

where β_1 is the unstandardized regression coefficient associated with group after controlling for the other variables in the model, β_k is the unstandardized regression coefficient associated with the unbalanced covariate included in the model after controlling for the other variables included in the model, and X_{ki} represents the score on the unbalanced covariate for person i. Each unbalanced covariate has its own β_k and X_{ki} term. All other terms are defined above. Theoretically, if the matching technique results in unbalanced covariates after matching, then the estimated treatment effects obtained using Equation 15 should be more accurate than the estimated treatment effects obtained in Equation 14; however, if the matching method balanced the covariates well, then the two equations should result in comparable estimates.

Summary. The five conditions were fully crossed to explore potential interactions among the conditions. Specifically, data were generated for each of the effect sizes by comparison-to-treatment ratios by sample sizes combinations (32 conditions). The simulation process was replicated 1,000 times for each condition, resulting in 32,000 unique data sets. For each simulated data set, the eight matching methods were employed to create matched comparison groups, then the two outcome analysis approaches were applied to examine the group differences between the treatment group and matched comparison groups. Appendix B displays the simulated conditions.

Evaluation Criteria

Performance of the matching methods was evaluated in a number of ways. First, matches were diagnosed in terms of quantity and quality (propensity score and covariate balance) of matches. Then, outcome analyses were conducted. Both significance tests and effect sizes were of interest in outcome analyses. Each of the criteria used to diagnoses matches and evaluate outcome analyses are described in further detail below.

Diagnosing Matches. The purpose of matching is to balance the distributions of the covariates for the treatment and matched comparison groups. As such, it is paramount that researchers examine the quality of matches to ensure that groups are properly balanced on covariates and propensity scores. Further, if treatment group members were excluded because an adequate comparison match was not available, researchers need to ensure that the matched treatment group is representative of the original treatment group sample. Two criteria were used to diagnosis matches.

Quality of Matches. Quality of matches was determined by examining propensity score balance and individual covariate balance. Propensity score balance was evaluated

via the standardized mean difference, variance ratio, and percent bias reduction in propensity scores. High quality matches are evidenced by mean differences on the propensity scores near zero (Austin, 2011b), propensity score variance ratios near one (Stuart, 2010), and percent bias reduction values at 80% or above (Bai, 2013; Cochran & Rubin, 1973).

Covariate balance was examined by comparing the treatment and matched comparison group on *each* covariate after matching. For continuous covariates, an effect size (Cohen's *d*) was examined; groups should be less than 0.25 standard deviation units apart (Stuart, 2010) to be considered balanced. For categorical covariates, the standardized difference (similar for Cohen's *d* for categorical covariates) was examined; groups should be less than 0.10 standard deviation units apart (Austin, 2009a) to be considered balanced. Additionally, frequencies on the categorical covariates were examined to determine whether the comparison group had over- or underrepresentation compared to the treatment group. The current study did not include visual diagnosis of balance.

Quantity of Matches. Quantity of matches can be assessed by examining the number of treatment group members who were successfully matched. As noted previously, a matching algorithm may result in adequate covariate balance (i.e., quality of matches); however, it may come at the cost of sample size (i.e., quantity of matches). Thus, it is the responsibility of the researcher to balance these competing goals. This study included an examination of the tradeoff between the quality and quantity of the matched and how it impacts the results of the outcomes analyses. In the current study, the

proportion of treatment group members who were successfully matched was examined to facilitate comparisons across replications and conditions.

Outcome Analyses. Outcome analyses were the primary focus of this study. Regression analyses were used to examine group differences on the outcomes between the treatment group and their matched comparison group. Type I error, power, and the estimated effect sizes were used to evaluate the accuracy of the resulting group comparisons. Outcome analyses were conducting using SAS, version 9.4.

Type I Error. Type I error is the probability of rejecting the null hypothesis when the null is true. In simulations, Type I error can be empirically determined when data are simulated under the null distribution. Type I error was defined as the proportion of replications where the groups were flagged as significantly different when there was no true difference (d = 0.0). The nominal alpha was set to 0.05. As such, it was expected that a Type I error would be observed about 5% of the time.

Power. Power is the probability of detecting an effect, given that an effect exists. Again, in simulations, power can be empirically determined when data are simulated under an alternative distribution (i.e., there is a true effect). Power was defined as the proportion of replications where the groups were flagged as significantly different when there was a true difference (d = 0.2, 0.5, and 0.8). Table 3 displays what the expected power would be under random assignment, which provides a benchmark for the power that could be achieved with propensity score matching; alpha was set to 0.05 for all power analyses. Power was calculated using an online power calculator (Soper, 2017). It is important to note that the significance tests are underpowered in the current study; however, significance tests were still conducted as a precursor for examining estimated

effect size, as is typically done in practice. The primary interest was on the accuracy of the estimated effect sizes.

Table 3

Statistical Power for Each Effect Size and Sample Size Combination

Sample Size Per Group	0.2	0.5	0.8
30	0.1151	0.4764	0.8602
100	0.2900	0.9402	0.9999

Estimated Effect Size. The unstandardized effect size was defined as the difference between the group means. In the population, the unstandardized effect size was equivalent to Cohen's *d* (a standardized effect size) because the pooled within-group standard deviation was scaled to equal one. The unstandardized effect size was used to avoid confounding errors in estimating the mean with errors in estimating the standard deviation. Two indices were reported for the difference between the means: bias and root mean squared error (RMSE).

Bias. Bias is the difference between the estimated parameter and the generating true parameter value, averaged across replications. Bias should be close to 0, indicating that on average, the estimated parameter is approximately the same as the true parameter value. To calculate bias for each parameter, the true population value is subtracted from the average estimate value across replications. Equation 16 presents this computation,

$$Bias_{\theta} = \frac{\sum_{r=1}^{R}(\hat{\theta}_r - \theta)}{R} \quad (16)$$

where $\hat{\theta}_r$ is the parameter estimate from the *r*th replication, θ is the true parameter value, and *R* is the total number of replications. In the current study, the parameter (θ) is the difference between the group means.

RMSE. RMSE combines both bias and sampling variability of parameter estimates across replications. It is calculated by taking the difference between the estimated parameter and the generated true parameter value. These values are squared and averaged across replications. The squaring is done so negative values do not cancel out positive values. This value is the mean squared error (MSE). The square root is taken to obtain the RMSE. RMSE values should be close to 0. The computational formula is presented in Equation 17,

$$RMSE_\theta = \sqrt{\frac{\sum_{r=1}^{R}(\widehat{\theta}_r - \theta)^2}{R}} = \sqrt{Bias_\theta^2 + SE_\theta^2} \quad (17)$$

where SE_θ^2 is the empirical standard error of the parameter (i.e., the standard deviation of the parameter estimates across replications). The other elements in the equation are defined above. Further, the variance explained in mean difference and squared mean difference was examined to determine which factors or interactions made a meaningful difference in parameter recovery using ANOVAs. Statistical significance was not examined, as the significance tests were overpowered. An effect size, specifically η^2, was used to determine which conditions were meaningful. Finally, taken together, these results were used to provide preliminary recommendations regarding which matching method to use under which conditions. Table 4 presents the alignment of the evaluation criteria described above with the research questions for the current study.

Table 4

Alignment of the Evaluation Criteria with Research Questions

Evaluation Criteria	Research Question			
	1	2	3	4
Quality of Matches	X			
Standardized mean difference	X			
Variance Ratio	X			
Percent Bias Reduction	X			
Quantity of Matches	X			
Percentage of successful matches	X			
Type I Error		X		
Power		X		
Bias			X	
RMSE			X	
Partitioning Variance				X

CHAPTER 4

Results

The purpose of the current study was to examine and compare common matching techniques used to estimate ATT. First, matching methods were compared in terms of the quantity and quality of matches. Then, outcome analyses were conducted to determine whether conclusions regarding group differences and estimated effect sizes depended on the matching technique used. Differences across effect size, treatment group sample size, comparison-to-treatment ratio, and analysis technique were also examined. The results are summarized in the following sections.

Research Question 1: Quality and Quantity of Matches

The first research question was aimed at exploring how the matches created from the various matching techniques differed in terms of quality and quantity of matches. Quality of matches was determined by examining propensity score balance and individual covariate balance. Quantity of matches was assessed by examining the percentage of treatment group members who were successfully matched. Given that the covariates were simulated independently of the treatment effects, quality and quantity of the matches were consistent across effect sizes.

Quality of matches. As shown in Table 5, propensity score balance was evaluated using three metrics: average standardized mean difference, variance ratio, and percent bias reduction. The standardized mean difference of the treatment and comparison propensity scores was calculated by dividing the average mean difference in the propensity scores across replications by the square root of the average pooled variance across replications. The standardized mean difference of the propensity scores

should be close to 0 (Austin, 2011b). Although the standardized mean difference was small across all propensity score matching methods, it was the smallest for the nearest neighbor matching with calipers. Moreover, as the caliper became more stringent, the standardized mean difference decreased, indicating that smaller calipers result in better quality matches than larger calipers. The standardized mean differences for nearest neighbor and optimal 1:1 matching were comparable, suggesting that treatment group members did not compete for comparison group matches. Additionally, the standardized mean difference for optimal 2:1 matching was consistently larger than the other matching methods. This is unsurprising, given that subsequent matches are not as similar to the treatment group member as the first match (Stuart, 2010), thus introducing additional imbalance in the propensity scores. Also worth noting, there is little variability around the mean difference across replications. These patterns were consistent across simulation conditions. Further, the mean difference generally decreased as treatment group sample size increased and as comparison-to-treatment group ratio increased.

Table 5

Propensity Score Balance Before and After Matching Across Conditions

Method	Std. Mean Difference	*SD* of Std. Mean Difference	Variance Ratio	Percent Bias Reduction
Treatment Group Sample Size = 30				
Comparison-to-Treatment = 3:1				
Before Matching	0.1496	0.0623	1.9189	--
NN	0.0382	0.0328	1.5692	74.4%
NN3	0.0007	0.0014	1.0030	99.5%
NN2	0.0004	0.0009	1.0013	99.8%
NN1	0.0001	0.0005	0.9999	99.9%
Op1	0.0363	0.0339	1.5441	75.8%
Op2	0.0935	0.0550	2.0098	37.5%

(continued)

Method	Std. Mean Difference	*SD* of Std. Mean Difference	Variance Ratio	Percent Bias Reduction
Comparison-to-Treatment = 4:1				
Before Matching	0.1370	0.0569	2.1487	--
NN	0.0254	0.0243	1.4601	81.5%
NN3	0.0005	0.0011	1.0018	99.6%
NN2	0.0003	0.0007	1.0005	99.8%
NN1	0.0001	0.0004	1.0000	99.9%
Op1	0.0233	0.0253	1.4254	83.0%
Op2	0.0624	0.0440	1.9527	54.5%
Comparison-to-Treatment = 5:1				
Before Matching	0.1264	0.0528	2.3942	--
NN	0.0186	0.0201	1.4049	85.3%
NN3	0.0004	0.0009	1.0011	99.7%
NN2	0.0002	0.0007	1.0007	99.8%
NN1	0.0001	0.0003	1.0000	99.9%
Op1	0.0169	0.0207	1.3749	86.6%
Op2	0.0450	0.0375	1.8759	64.4%
Comparison-to-Treatment = 6:1				
Before Matching	0.1155	0.0484	2.5121	--
NN	0.0135	0.0154	1.3379	88.3%
NN3	0.0003	0.0008	1.0013	99.7%
NN2	0.0002	0.0006	0.9998	99.8%
NN1	0.0000	0.0003	0.9999	100.0%
Op1	0.0118	0.0159	1.3025	89.8%
Op2	0.0325	0.0301	1.7311	71.9%
Treatment Group Sample Size = 100				
Comparison-to-Treatment = 3:1				
Before Matching	0.1097	0.0308	1.7136	--
NN	0.0164	0.0121	1.3325	85.1%
NN3	0.0007	0.0005	1.0051	99.3%
NN2	0.0004	0.0003	1.0027	99.6%
NN1	0.0002	0.0002	1.0006	99.9%
Op1	0.0152	0.0127	1.3115	86.2%
Op2	0.0586	0.0251	1.8247	46.6%

(continued)

Method	Std. Mean Difference	*SD* of Std. Mean Difference	Variance Ratio	Percent Bias Reduction
Comparison-to-Treatment = 4:1				
Before Matching	0.1034	0.0287	1.9195	--
NN	0.0102	0.0088	1.2501	90.1%
NN3	0.0005	0.0004	1.0042	99.5%
NN2	0.0003	0.0003	1.0018	99.7%
NN1	0.0001	0.0001	1.0007	99.9%
Op1	0.0090	0.0093	1.2247	91.3%
Op2	0.0367	0.0205	1.7179	64.5%
Comparison-to-Treatment = 5:1				
Before Matching	0.0967	0.0257	2.0715	--
NN	0.0068	0.0062	1.1959	92.9%
NN3	0.0004	0.0004	1.0026	99.6%
NN2	0.0002	0.0002	1.0015	99.8%
NN1	0.0001	0.0001	1.0004	99.9%
Op1	0.0058	0.0064	1.1718	94.0%
Op2	0.0243	0.0154	1.5774	74.9%
Comparison-to-Treatment = 6:1				
Before Matching	0.0910	0.0252	2.2235	--
NN	0.0051	0.0051	1.1686	94.4%
NN3	0.0003	0.0003	1.0020	99.7%
NN2	0.0002	0.0002	1.0012	99.8%
NN1	0.0001	0.0001	1.0003	99.9%
Op1	0.0043	0.0052	1.1453	95.3%
Op2	0.0173	0.0131	1.4820	81.0%

Note. NN = nearest neighbor matching, NN3 = nearest neighbor matching with a caliper of 0.3, NN2 = nearest neighbor matching with a caliper of 0.2, NN1 = nearest neighbor matching with a caliper of 0.1, Op1 = optimal 1:1 matching, and Op2 = optimal 2:1 matching. Propensity scores were not calculated for random sampling and Mahalanobis distance matching. As such, these matching methods were not included above. The standardized mean differences were calculated by subtracting the mean propensity score for the comparison group from the mean propensity score for the treatment group. The variance ratios were calculated by taking the average propensity score variance for the treatment group across replications and dividing it by the average propensity score variance for the comparison group across replications.

The variance ratio of the treatment and comparison propensity scores was calculated by taking the average propensity score variance for the treatment group across replications and dividing it by the average propensity score variance for the comparison group across replications. The variance ratio should be close to 1 (Stuart, 2010),

indicating that the variance of the propensity scores is about the same across the two groups. There was more variability in the variance ratios across the matching methods than there was for the standardized mean differences. Variance ratios were most balanced (e.g., closest to 1) for the nearest neighbor matching method with calipers. Moreover as the caliper decreased (e.g., became more stringent), the variance ratio was closer to 1. The variance ratios for nearest neighbor and optimal 1:1 matching were comparable. This, again, suggests that treatment group members did not compete for comparison group matches. Additionally, the variance ratio for optimal 2:1 matching was consistently larger than the other matching methods. These patterns were consistent across simulation conditions. Further, for the optimal and nearest neighbor matching without calipers, the variance ratio was closer to 1 as treatment group sample size increased and as comparison-to-treatment group ratio increased.

Percent bias reduction should be 80% or above (Bai, 2013; Cochran & Rubin, 1973). Although the percent bias reduction was above 80% for most propensity score matching methods, it was largest for nearest neighbor matching with calipers. Moreover, as the caliper decreased (e.g., became more stringent), the percent bias reduction increased, suggesting better quality matches than with the larger calipers. Percent bias reduction for nearest neighbor and optimal 1:1 matching were comparable, with optimal 1:1 consistently resulting in slightly larger percent bias reduction than nearest neighbor matching. Again, this suggests that treatment group members did not compete for comparison group matches. Additionally, the percent bias reduction for optimal 2:1 matching was consistently smaller than the other matching methods. These patterns held across simulation conditions. Further, the percent bias reduction generally increased as

treatment group sample size increased and as comparison-to-treatment group ratio increased, except when using the nearest neighbor matching with calipers, where the percent bias reduction was nearly always close to 100%.

Covariate balance was examined by comparing the treatment and matched comparison group on *each* continuous covariate after matching. The average standardized mean differences (e.g., Cohen's *d*) between treatment and comparison groups for the continuous covariates were examined; the groups should be less than 0.25 standard deviation units apart (Stuart, 2010) to be considered balanced. Tables 6 and 7 present the covariate balance across the different matching methods. The absolute value of the standardized differences for the continuous covariates before matching ranged from 0.06 to 0.65 standard deviation units; the differences for the continuous covariates in the random sample were about the same as before matching. The standardized mean difference between groups on the continuous covariates increased as treatment group sample size and comparison-to-treatment ratio increased for random sampling; this was also true regarding group differences on the covariates before matching.

Table 6

Continuous Covariate Balance Before and After Matching Across Conditions

Method	Treatment N = 30				Treatment N = 100			
	X1	X2	X3	X4	X1	X2	X3	X4
	Comparison-to-Treatment = 3:1							
Before Matching	-0.35	-0.60	0.10	0.08	-0.35	-0.59	0.11	0.07
Ran	-0.35	-0.61	0.11	0.06	-0.35	-0.58	0.11	0.07
NN	-0.06	-0.09	0.02	0.01	-0.04	-0.05	0.01	0.00
NN3	0.01	0.00	0.01	-0.01	0.00	0.00	0.00	0.00
NN2	0.01	0.01	0.01	-0.01	0.00	0.00	0.00	0.01
NN1	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00

(continued)

Method	Treatment N = 30				Treatment N = 100			
	X1	X2	X3	X4	X1	X2	X3	X4
Op1	-0.06	-0.08	0.01	0.00	-0.03	-0.04	0.01	0.00
Op2	-0.17	-0.28	0.04	0.03	-0.15	-0.22	0.03	0.02
Mah	-0.15	-0.28	0.05	0.03	-0.10	-0.20	0.05	0.02
Comparison-to-Treatment = 4:1								
Before Matching	-0.36	-0.61	0.13	0.07	-0.36	-0.61	0.12	0.06
Ran	-0.37	-0.61	0.13	0.06	-0.36	-0.61	0.12	0.06
NN	-0.04	-0.06	0.02	0.01	-0.02	-0.03	0.01	0.01
NN3	-0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.00
NN2	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00
NN1	-0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
Op1	-0.04	-0.06	0.01	0.00	-0.01	-0.03	0.00	0.01
Op2	-0.12	-0.17	0.03	0.01	-0.08	-0.12	0.02	0.00
Mah	-0.12	-0.24	0.06	0.03	-0.08	-0.17	0.04	0.02
Comparison-to-Treatment = 5:1								
Before Matching	-0.38	-0.64	0.11	0.08	-0.37	-0.63	0.12	0.07
Ran	-0.39	-0.63	0.10	0.07	-0.37	-0.63	0.12	0.07
NN	-0.03	-0.04	0.00	0.00	-0.01	-0.02	0.00	0.00
NN3	-0.01	0.00	-0.01	0.00	0.00	0.00	0.00	-0.01
NN2	-0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00
NN1	-0.02	0.01	-0.01	0.02	-0.01	0.00	0.00	0.00
Op1	-0.02	-0.04	0.00	0.01	-0.01	-0.02	0.01	0.00
Op2	-0.08	-0.11	0.01	0.01	-0.05	-0.07	0.01	0.00
Mah	-0.10	-0.22	0.04	0.03	-0.07	-0.16	0.04	0.02
Comparison-to-Treatment = 6:1								
Before Matching	-0.37	-0.65	0.12	0.07	-0.38	-0.64	0.12	0.07
Ran	-0.37	-0.64	0.13	0.06	-0.38	-0.65	0.12	0.07
NN	-0.02	-0.04	0.01	0.01	-0.01	-0.01	0.00	0.00
NN3	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NN2	0.00	0.00	0.01	-0.01	0.00	0.00	0.00	0.00
NN1	-0.01	0.00	0.01	-0.01	0.00	0.00	0.00	-0.01
Op1	-0.02	-0.03	0.01	0.00	-0.01	-0.01	0.00	0.00
Op2	-0.06	-0.09	0.02	0.00	-0.04	-0.05	0.01	0.00
Mah	-0.09	-0.21	0.05	0.02	-0.07	-0.15	0.04	0.01

Note. Ran = random sampling, NN = nearest neighbor matching, NN3 = nearest neighbor matching with a caliper of 0.3, NN2 = nearest neighbor matching with a caliper of 0.2, NN1 = nearest neighbor matching with a caliper of 0.1, Op1 = optimal 1:1 matching, Op2 = optimal 2:1 matching, and Mah = Mahalanobis distance matching.

Most of the matching techniques resulted in average standardized mean differences between the groups on the continuous covariates less than 0.25 standard deviation units. However, nearest neighbor matching with the calipers produced the most equivalent matched comparison group, with standard deviation unit differences for the continuous covariates of 0.02 or less. Continuous covariate balance for nearest neighbor and optimal 1:1 matching were comparable and performed only slightly worse than nearest neighbor matching with calipers. Optimal 2:1 and Mahalanobis distance matching were comparable, and performed worse than the other matching methods at balancing the continuous covariates. Again, these patterns were consistent across simulation conditions. Further, the standardized mean difference between treatment and comparison groups on the continuous covariates generally decreased as treatment group sample size and comparison-to-treatment group ratio increased, with the exception of nearest neighbor matching with calipers, where the standardized mean difference was nearly zero in all conditions.

Categorical covariate balance was assessed by comparing the proportion of treatment group members to the proportion of comparison group members after matching and by examining standardized mean difference (similar to Cohen's *d* for categorical covariates). The standardized mean difference between the treatment and comparison groups for categorical covariates should be less than 0.1 (Austin, 2009a). Table 7 presents the covariate balance for the categorical covariates across the different matching methods. The absolute value of the standardized mean difference for the categorical covariates ranged from 0.11 to 0.45 standard deviation units before matching; the

differences for the categorical covariates in the random sample were about the same as the full sample before matching.

Table 7

Categorical Covariate Balance Before and After Matching Across Conditions

Method	X5			X6		
	Treatment Proportion	Comparison Proportion	Std. Mean Difference	Treatment Proportion	Comparison Proportion	Std. Mean Difference
			N=30			
			Comparison-to-Treatment = 3:1			
Before Matching	0.56	0.62	-0.12	0.27	0.11	0.41
Ran	0.56	0.62	-0.12	0.27	0.11	0.42
NN	0.56	0.57	-0.02	0.27	0.21	0.13
NN3	0.59	0.59	-0.01	0.16	0.16	0.01
NN2	0.59	0.59	-0.01	0.14	0.14	0.01
NN1	0.60	0.59	0.00	0.13	0.13	0.01
Op1	0.56	0.57	-0.01	0.27	0.22	0.12
Op2	0.56	0.59	-0.05	0.27	0.15	0.29
Mah	0.56	0.57	-0.02	0.27	0.22	0.10
			Comparison-to-Treatment = 4:1			
Before Matching	0.55	0.61	-0.12	0.28	0.12	0.43
Ran	0.55	0.61	-0.12	0.28	0.12	0.44
NN	0.55	0.56	-0.01	0.28	0.25	0.09
NN3	0.58	0.58	-0.01	0.18	0.17	0.01
NN2	0.59	0.59	-0.01	0.16	0.16	0.02
NN1	0.59	0.60	-0.01	0.14	0.14	0.01
Op1	0.55	0.56	-0.01	0.28	0.25	0.08
Op2	0.55	0.57	-0.03	0.28	0.19	0.22
Mah	0.55	0.57	-0.03	0.28	0.27	0.04
			Comparison-to-Treatment = 5:1			
Before Matching	0.55	0.61	-0.12	0.30	0.12	0.44
Ran	0.55	0.61	-0.13	0.30	0.12	0.46
NN	0.55	0.55	-0.01	0.30	0.27	0.06

(continued)

Method	X5			X6		
	Treatment Proportion	Comparison Proportion	Std. Mean Difference	Treatment Proportion	Comparison Proportion	Std. Mean Difference
NN3	0.57	0.58	0.00	0.19	0.18	0.02
NN2	0.58	0.58	0.00	0.17	0.17	0.01
NN1	0.58	0.59	-0.02	0.15	0.14	0.02
Op1	0.55	0.55	-0.01	0.30	0.27	0.06
Op2	0.55	0.56	-0.02	0.30	0.23	0.16
Mah	0.55	0.56	-0.02	0.30	0.29	0.02
Comparison-to-Treatment = 6:1						
Before Matching	0.55	0.61	-0.13	0.31	0.12	0.45
Ran	0.55	0.61	-0.13	0.31	0.12	0.47
NN	0.55	0.55	0.00	0.31	0.29	0.04
NN3	0.57	0.56	0.01	0.20	0.20	0.01
NN2	0.57	0.58	-0.01	0.18	0.18	0.03
NN1	0.58	0.58	-0.01	0.16	0.15	0.03
Op1	0.55	0.55	-0.01	0.31	0.29	0.04
Op2	0.55	0.55	-0.02	0.31	0.25	0.11
Mah	0.55	0.56	-0.02	0.31	0.30	0.01
N=100						
Comparison-to-Treatment = 3:1						
Before Matching	0.56	0.61	-0.11	0.26	0.11	0.40
Ran	0.56	0.61	-0.11	0.26	0.11	0.40
NN	0.56	0.56	-0.01	0.26	0.23	0.08
NN3	0.57	0.57	0.00	0.19	0.18	0.02
NN2	0.58	0.58	0.00	0.17	0.17	0.01
NN1	0.58	0.59	-0.01	0.15	0.15	0.01
Op1	0.56	0.56	-0.01	0.26	0.23	0.07
Op2	0.56	0.58	-0.04	0.26	0.16	0.26
Mah	0.56	0.57	-0.02	0.26	0.25	0.03
Comparison-to-Treatment = 4:1						
Before Matching	0.55	0.61	-0.12	0.28	0.12	0.42
Ran	0.55	0.61	-0.12	0.28	0.12	0.43
NN	0.55	0.56	-0.01	0.28	0.27	0.04

(continued)

Method	X5			X6		
	Treatment Proportion	Comparison Proportion	Std. Mean Difference	Treatment Proportion	Comparison Proportion	Std. Mean Difference
NN3	0.57	0.57	0.00	0.21	0.21	0.01
NN2	0.57	0.57	0.00	0.19	0.19	0.01
NN1	0.58	0.58	0.00	0.17	0.17	0.01
Op1	0.55	0.56	-0.01	0.28	0.27	0.04
Op2	0.55	0.56	-0.02	0.28	0.21	0.18
Mah	0.55	0.56	-0.02	0.28	0.28	0.01
Comparison-to-Treatment = 5:1						
Before Matching	0.55	0.61	-0.13	0.30	0.12	0.44
Ran	0.55	0.61	-0.12	0.30	0.12	0.45
NN	0.55	0.55	0.00	0.30	0.28	0.03
NN3	0.56	0.56	0.00	0.23	0.22	0.01
NN2	0.57	0.57	0.00	0.21	0.21	0.02
NN1	0.57	0.57	0.00	0.18	0.18	0.01
Op1	0.55	0.55	0.00	0.30	0.29	0.03
Op2	0.55	0.56	-0.01	0.30	0.24	0.12
Mah	0.55	0.55	-0.01	0.30	0.30	0.00
Comparison-to-Treatment = 6:1						
Before Matching	0.55	0.61	-0.12	0.31	0.12	0.45
Ran	0.55	0.61	-0.12	0.31	0.12	0.46
NN	0.55	0.55	-0.01	0.31	0.30	0.02
NN3	0.56	0.56	0.00	0.24	0.24	0.01
NN2	0.56	0.57	0.00	0.23	0.22	0.01
NN1	0.57	0.57	0.00	0.20	0.19	0.02
Op1	0.55	0.55	0.00	0.31	0.30	0.02
Op2	0.55	0.55	-0.01	0.31	0.27	0.08
Mah	0.55	0.55	0.00	0.31	0.31	0.00

Note. Ran = random sampling, NN = nearest neighbor matching, NN3 = nearest neighbor matching with a caliper of 0.3, NN2 = nearest neighbor matching with a caliper of 0.2, NN1 = nearest neighbor matching with a caliper of 0.1, Op1 = optimal 1:1 matching, Op2 = optimal 2:1 matching, and Mah = Mahalanobis distance matching.

Although propensity score and Mahalanobis distance matching techniques created more balanced groups than before matching, nearest neighbor matching with the calipers produced the most equivalent matched comparison group, with absolute standard

deviation unit differences for the categorical covariates of 0.03 or less. Categorical covariate balance for nearest neighbor, optimal 1:1, and Mahalanobis distance matching were comparable, and performed only slightly worse than nearest neighbor matching with calipers. Moreover, Mahalanobis distance matching resulted in slightly better balanced categorical covariates than nearest neighbor and optimal 1:1 when treatment group sample size was 100. For X5, optimal 2:1 matching performed comparably to nearest neighbor, optimal 1:1, and Mahalanobis distance matching for the 4:1 to 6:1 comparison-to-treatment ratios. However, optimal 2:1 matching performed worse than the other matching methods at balancing X6.

All matching methods balanced X5 well; however, it was more difficult to balance X6. Further, it is important to note that in some replications, the entire representation of one group on X6 was excluded from analysis due to lack of an adequate match. This is particularly problematic for generalizability. That is, when the representation of one group is lost, then the results no longer generalize back to the original treatment group. Thus, the generalizability of the results is limited by the representation of the matched treatment group.

Covariates were considered unbalanced if the absolute value of the standardized mean difference was greater than 0.25 for continuous covariates (Stuart, 2010) or 0.10 for categorical covariates (Austin, 2009a). The percentage of replications in which each covariate was unbalanced was examined by condition and is presented in Table 8. Random sampling had the highest percentage of replications with unbalanced covariates across all covariates, with X4 and X5 being unbalanced less frequently than the other covariates. This pattern held across treatment group sample sizes and comparison-to-

treatment group ratios. Across the other matching methods, the percentages of unbalanced covariates was larger when treatment group sample size was small.

Table 8

Proportion of Replications with Unbalanced Covariates by Covariate and Conditions

Method	X1	X2	X3	X4	X5	X6
Treatment Group Sample Size = 30						
Comparison-to-Treatment = 3:1						
Ran	67%	92%	35%	31%	71%	88%
NN	8%	10%	6%	8%	45%	57%
NN3	21%	17%	24%	24%	69%	70%
NN2	29%	24%	30%	30%	79%	71%
NN1	40%	33%	46%	44%	77%	65%
Op1	8%	10%	6%	7%	44%	53%
Op2	28%	55%	5%	4%	46%	83%
Mah	26%	56%	10%	10%	25%	33%
Comparison-to-Treatment = 4:1						
Ran	68%	92%	40%	35%	74%	90%
NN	10%	7%	8%	9%	45%	47%
NN3	21%	14%	25%	24%	61%	72%
NN2	26%	22%	30%	28%	74%	73%
NN1	36%	30%	43%	40%	80%	70%
Op1	9%	6%	8%	10%	46%	45%
Op2	11%	27%	2%	2%	40%	76%
Mah	16%	49%	11%	9%	20%	14%
Comparison-to-Treatment = 5:1						
Ran	70%	93%	38%	36%	73%	93%
NN	8%	5%	9%	10%	50%	43%
NN3	22%	12%	24%	23%	58%	70%
NN2	27%	19%	29%	30%	71%	72%
NN1	36%	24%	38%	38%	82%	69%
Op1	8%	5%	9%	9%	47%	41%
Op2	6%	12%	1%	1%	36%	67%
Mah	13%	43%	7%	6%	17%	6%

(continued)

Method	X1	X2	X3	X4	X5	X6
Comparison-to-Treatment = 6:1						
Ran	67%	94%	41%	37%	73%	93%
NN	9%	4%	11%	11%	50%	37%
NN3	21%	13%	23%	23%	57%	68%
NN2	25%	17%	27%	28%	68%	75%
NN1	34%	23%	39%	35%	81%	73%
Op1	8%	5%	10%	10%	48%	38%
Op2	3%	6%	2%	1%	37%	56%
Mah	11%	37%	7%	5%	10%	1%
Treatment Group Sample Size = 100						
Comparison-to-Treatment = 3:1						
Ran	74%	99%	17%	11%	62%	99%
NN	0%	0%	0%	0%	22%	35%
NN3	1%	0%	1%	1%	26%	23%
NN2	1%	0%	1%	1%	30%	30%
NN1	2%	0%	3%	3%	38%	33%
Op1	0%	0%	0%	0%	22%	31%
Op2	8%	38%	0%	0%	15%	94%
Mah	2%	28%	0%	0%	7%	8%
Comparison-to-Treatment = 4:1						
Ran	80%	100%	19%	11%	64%	99%
NN	0%	0%	0%	0%	25%	18%
NN3	1%	0%	1%	1%	27%	24%
NN2	1%	0%	1%	2%	32%	23%
NN1	2%	0%	3%	4%	38%	38%
Op1	0%	0%	0%	0%	26%	16%
Op2	1%	4%	0%	0%	8%	80%
Mah	0%	14%	0%	0%	4%	1%
Comparison-to-Treatment = 5:1						
Ran	81%	99%	17%	12%	64%	100%
NN	0%	0%	1%	0%	28%	16%
NN3	0%	0%	1%	1%	27%	25%
NN2	1%	0%	2%	2%	29%	28%
NN1	3%	0%	3%	2%	41%	36%
Op1	0%	0%	0%	0%	28%	15%
Op2	0%	0%	0%	0%	5%	58%
Mah	0%	8%	0%	0%	1%	0%

(continued)

Method	X1	X2	X3	X4	X5	X6
			Comparison-to-Treatment = 6:1			
Ran	82%	100%	19%	12%	64%	100%
NN	0%	0%	1%	1%	29%	14%
NN3	1%	0%	1%	1%	28%	22%
NN2	2%	0%	2%	2%	31%	27%
NN1	2%	0%	3%	4%	42%	33%
Op1	0%	0%	0%	1%	26%	16%
Op2	0%	0%	0%	0%	7%	34%
Mah	0%	5%	0%	0%	0%	0%

Note. Ran = random sampling, NN = nearest neighbor matching, NN3 = nearest neighbor matching with a caliper of 0.3, NN2 = nearest neighbor matching with a caliper of 0.2, NN1 = nearest neighbor matching with a caliper of 0.1, Op1 = optimal 1:1 matching, Op2 = optimal 2:1 matching, and Mah = Mahalanobis distance matching.

Nearest neighbor matching also had a high percentage of replications with unbalanced covariates. Moreover, as the caliper became more stringent, the percentage of replications with unbalanced covariates increased. This pattern held across comparison-to-treatment group ratios for a treatment group sample size of 30. When the treatment group sample size was 100, the percentage of replications with unbalanced continuous covariate was much smaller; however, the percentages of replications with unbalanced categorical covariates were still large.

The percentage of replications with unbalanced covariates varied across conditions (e.g. treatment group sample size and comparison-to-treatment group ratio) for nearest neighbor, optimal 1:1 and 2:1, and Mahalanobis distance matching. Thus, these methods were compared within treatment group sample sizes.

When the treatment group sample size was 30, nearest neighbor and optimal 1:1 matching were comparable for both continuous and categorical covariates. Mahalanobis distance matching resulted in a higher percentage of unbalanced continuous covariates than nearest neighbor and optimal 1:1 matching; however, Mahalanobis distance

matching resulted in a smaller percentage of unbalanced categorical covariates (X5 and X6) than nearest neighbor and optimal 1:1 matching. The performance of optimal 2:1 matching varied by covariate. Specifically, the percentage of replications when X3 and X4 were unbalanced was consistently smaller than the other covariates and smaller than the other matching methods. The percentage of replications when X1 and X2 were unbalanced was larger when the comparison-to-treatment group ratio was smaller; however, the percentages were comparable to nearest neighbor and optimal 1:1 matching when the comparison-to-treatment group ratio was at least 5:1. Finally, optimal matching resulted in a higher percentage of unbalance for X6 than X5 for comparison-to-treatment group ratios of 3:1 to 5:1.

When the treatment group sample size was 100, the percentage of replications with unbalanced continuous covariates (X1, X3, and X4) was comparable for nearest neighbor, optimal 1:1, optimal 2:1, and Mahalanobis distance matching; however, the percentage of replications where X2 was unbalanced was higher for Mahalanobis distance matching than the other matching methods. Optimal 2:1 was comparable to these other methods when the comparison-to-treatment group ratio was at least 4:1. However, the percentage of replications with unbalanced categorical covariates was much smaller for Mahalanobis distance matching than nearest neighbor, optimal 1:1, and optimal 2:1 matching. Additionally, optimal 2:1 resulted in a smaller percentage of replications with unbalanced X5 than nearest neighbor and optimal 1:1 matching. Conversely, optimal 2:1 resulted in a larger percentage of replications with unbalanced X6 than nearest neighbor and optimal 1:1 matching.

Quantity of matches. Quantity of matches was assessed by examining the *percentage* of treatment group members who were successfully matched. Table 9 lists the average percentage of treatment group members who were retained after matching by condition. The percentage of the comparison group that was retained was a function of the sample size of the treatment group and matching method used. The full treatment sample was retained for all of the matching techniques except for the nearest neighbor with calipers.[6] For the nearest neighbor with caliper methods, as the caliper size decreased (e.g., became more stringent) so did the proportion of the treatment group that was successfully matched. This pattern held across conditions (e.g., treatment group sample size and comparison-to-treatment group ratio). Additionally, a larger percentage of treatment group members were successfully matched as the comparison-to-treatment group ratio increased. Further, a smaller percentage of treatment group members were retained when the treatment group sample size was smaller (e.g., 30). This is particularly problematic as the treatment group was already fairly small so the loss of treatment group members may result in too few matches to conduct the outcome analyses of interest, as well as loss of power.

Table 9

Quantity of Matches After Matching Across Conditions

Method	Treatment *N* = 30		Treatment *N* = 100	
	M	*SD*	*M*	*SD*
Comparison-to-Treatment = 3:1				
Ran	100%	0%	100%	0%
NN	100%	0%	100%	0%
NN3	66%	7%	83%	3%
NN2	58%	7%	79%	3%
NN1	43%	8%	69%	4%

(continued)

Method	Treatment $N = 30$		Treatment $N = 100$	
	M	*SD*	*M*	*SD*
Op1	100%	0%	100%	0%
Op2	100%	0%	100%	0%
Mah	100%	0%	100%	0%
Comparison-to-Treatment = 4:1				
Ran	100%	0%	100%	0%
NN	100%	0%	100%	0%
NN3	70%	7%	86%	3%
NN2	62%	7%	82%	3%
NN1	48%	8%	73%	3%
Op1	100%	0%	100%	0%
Op2	100%	0%	100%	0%
Mah	100%	0%	100%	0%
Comparison-to-Treatment = 5:1				
Ran	100%	0%	100%	0%
NN	100%	0%	100%	0%
NN3	72%	7%	87%	3%
NN2	66%	7%	84%	3%
NN1	52%	8%	76%	3%
Op1	100%	0%	100%	0%
Op2	100%	0%	100%	0%
Mah	100%	0%	100%	0%
Comparison-to-Treatment = 6:1				
Ran	100%	0%	100%	0%
NN	100%	0%	100%	0%
NN3	74%	7%	88%	3%
NN2	68%	7%	85%	3%
NN1	55%	8%	77%	3%
Op1	100%	0%	100%	0%
Op2	100%	0%	100%	0%
Mah	100%	0%	100%	0%

Note. Ran = random sampling, NN = nearest neighbor matching, NN3 = nearest neighbor matching with a caliper of 0.3, NN2 = nearest neighbor matching with a caliper of 0.2, NN1 = nearest neighbor matching with a caliper of 0.1, Op1 = optimal 1:1 matching, Op2 = optimal 2:1 matching, and Mah = Mahalanobis distance matching.

Research Question 2: Type I Error and Power

The second research question concerned how the results of group comparisons (e.g., significance tests) compared across the matching techniques and conditions. Type I

error was examined when the true effect between the groups after matching was simulated to be zero (i.e., $d = 0$). Power was examined when the true effect between the groups after matching was simulated to be greater than zero (i.e., $d = 0.2$, 0.5, and 0.8).

Type I Error. Type I error was defined as the proportion of replications where the groups were significantly different when there was no true difference ($d = 0.0$). The nominal alpha was set to 0.05, thus it was expected that a Type I error would be observed about 5% of the time. Figures 2 and 3 display the Type I error across conditions for the treatment group sample sizes of 30 and 100, respectively. Type I error was within the nominal rate for most of the matching methods across all conditions. Regardless of treatment group sample size and comparison-to-treatment ratio, the inclusion of the unbalanced covariates resulted in a Type I error rate around 5%. When unbalanced covariates were not included in the analyses, random sampling resulted in a Type I error rate of about 25% for a treatment group sample size of 30 and about 60% to 70% for a treatment group sample size of 100. Optimal 2:1 matching also resulted in in a Type I error rate that was slightly above 5% when the comparison-to-treatment group ratio was 3:1 (treatment $N = 30$ and 100) and 4:1 (treatment $N = 100$). Additionally, Mahalanobis distance matching resulted in a Type I error rate slightly over 5% for most comparison-to-treatment ratio when treatment group sample size was 100.

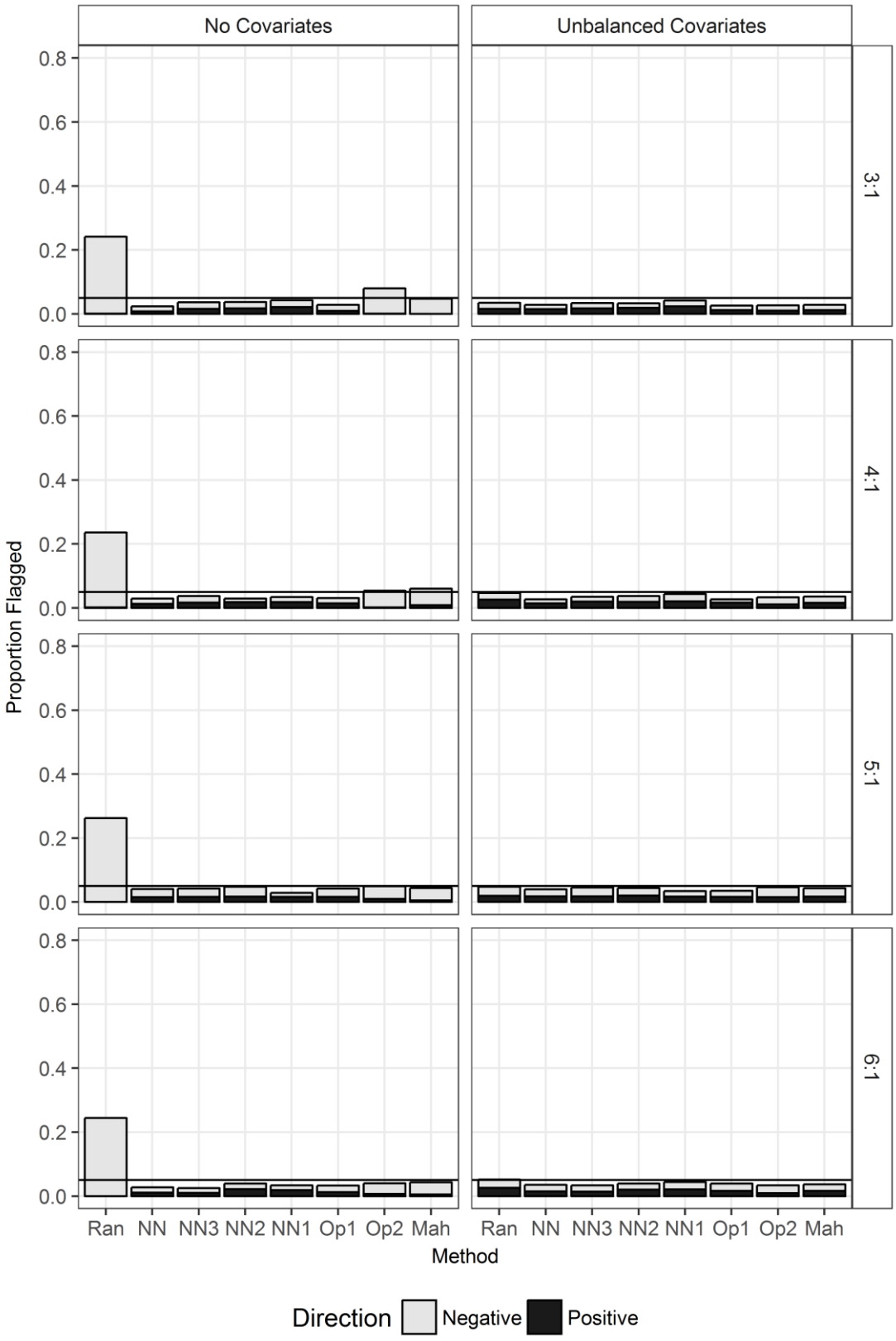

Figure 2. Type I error across conditions, treatment $N = 30$. Negative direction indicates the estimated treatment effect favored the comparison group.

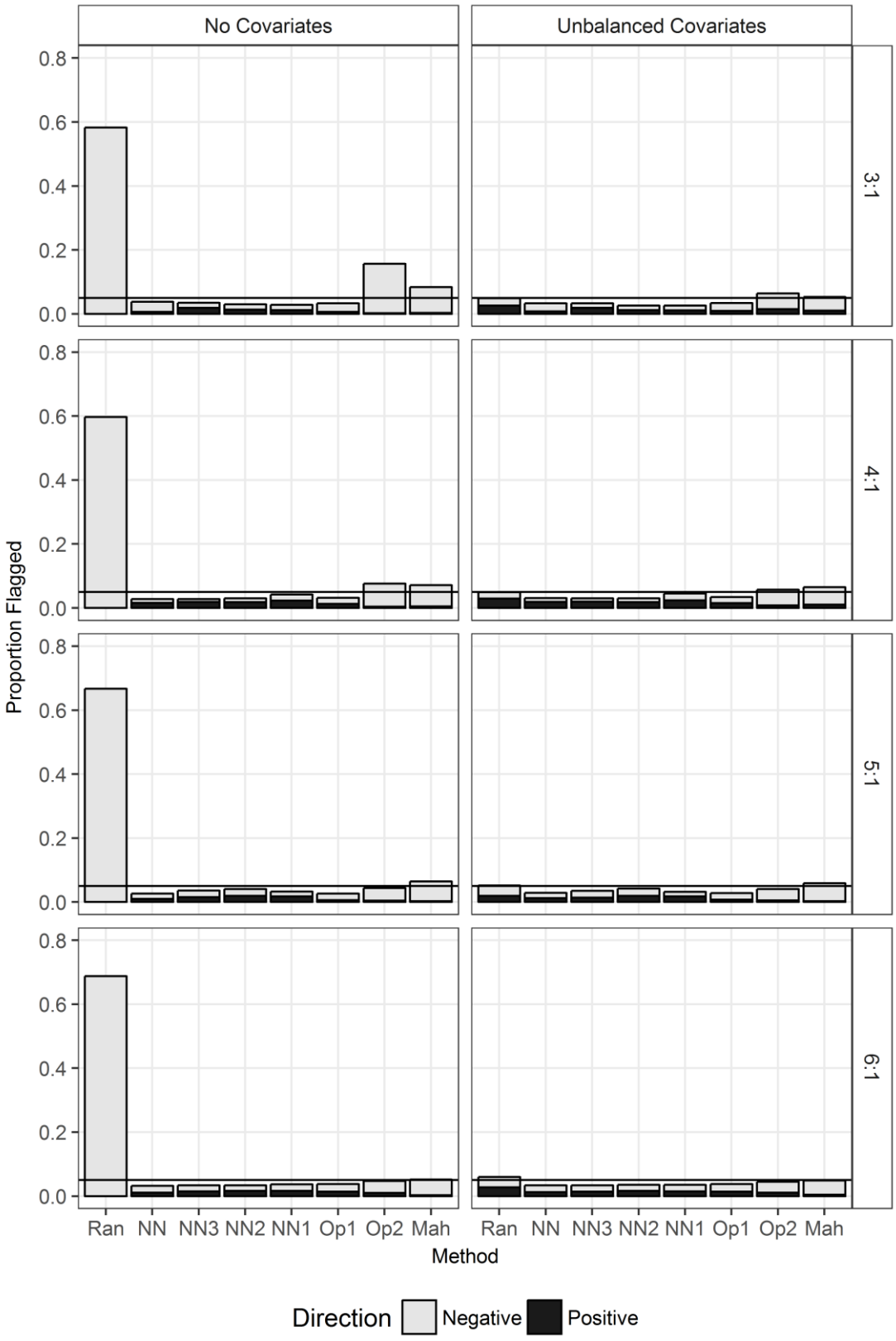

Figure 3. Type I error across conditions, treatment $N = 100$. Negative direction indicates the estimated treatment effect favored the comparison group.

Power. Power was defined as the proportion of replications where the groups were significantly different when there was a true difference (d = 0.2, 0.5, and 0.8). It is worth noting that, because the significance tests was two-tailed, a literal definition of power includes significance in either direction. However, the results below include a differentiation between power in the correct direction (the mean on the outcome for the treatment group was significantly *higher* than the mean on the outcome for the comparison group) and power in the incorrect direction (the mean on the outcome for the treatment group was significantly *lower* than the mean on the outcome for the comparison group). Figures 4 and 5 display power in the correct direction across conditions for the treatment group sample sizes of 30 and 100, respectively.

Unsurprisingly, power was lower for the smaller effect sizes. Additionally, power was lower when treatment group sample size was 30 than when treatment group sample size was 100. When treatment group samples size was 30, power was lower for nearest neighbor matching with calipers than for the other matching methods (except random sampling); power was lower for more stringent calipers than for more liberal calipers. This is unsurprising given that there was a loss of sample size when calipers were applied. Although still low, power was higher for nearest neighbor and optimal 1:1 and 2:1 matching. Power was higher for Mahalanobis distance matching when unbalanced covariates were included in the analyses than when no covariates were included in the analyses; analysis did not impact power for the other matching methods. Moreover, comparison-to-treatment group ratio did not affect power across the matching methods.

When the sample size was 100 and the effect size was 0.5 or 0.8, power across the different matching methods was close to 1, expect for random sampling. When effect

size was 0.2, power was around 0.25 for all matching methods, except for random sampling and Mahalanobis distance matching; power was a little lower for these methods. Generally, this pattern held regardless of comparison-to-treatment ratio. Power was lower for random sampling when no covariates were included in the analyses than when unbalanced covariates were included in the analyses; including unbalanced covariates in the analysis affected power minimally for the other matching methods.

As noted previously, power can also be in the incorrect direction. In the current study, power in the incorrect direction meant that the mean of the outcome for the treatment group was statistically significantly *lower* than the mean of the outcome for the comparison group. Random sampling had a higher proportion of power in the incorrect direction than the other matching methods, except when treatment group sample size was 100 and unbalanced covariates were included in the group comparisons on the outcome. When sample size was 30 and effect size was 0.2, most matching methods had some power in the incorrect direction. However, the power in the wrong direction was small (e.g., less than 1%). Additionally, when the effect size was 0.5, a small number of matching methods has some power in the incorrect direction. The pattern was inconsistent across matching methods. When sample size was 100, there were fewer instances of power in the wrong direction; however, power in the wrong direction for random sampling increased with the larger treatment group sample size. Appendix C presents graphs for power in the incorrect direction across conditions.

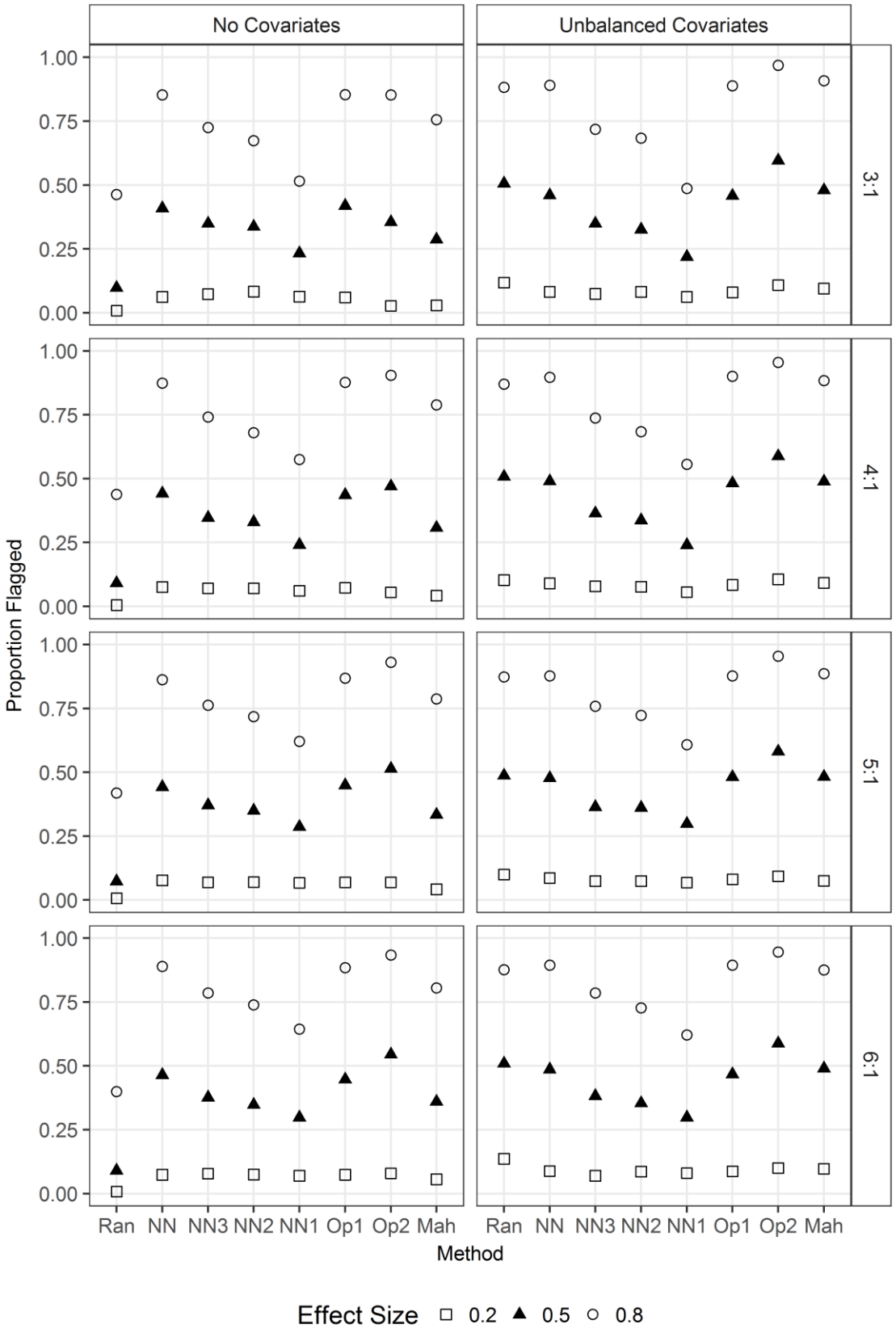

Figure 4. Power in the correct direction across conditions, treatment $N = 30$.

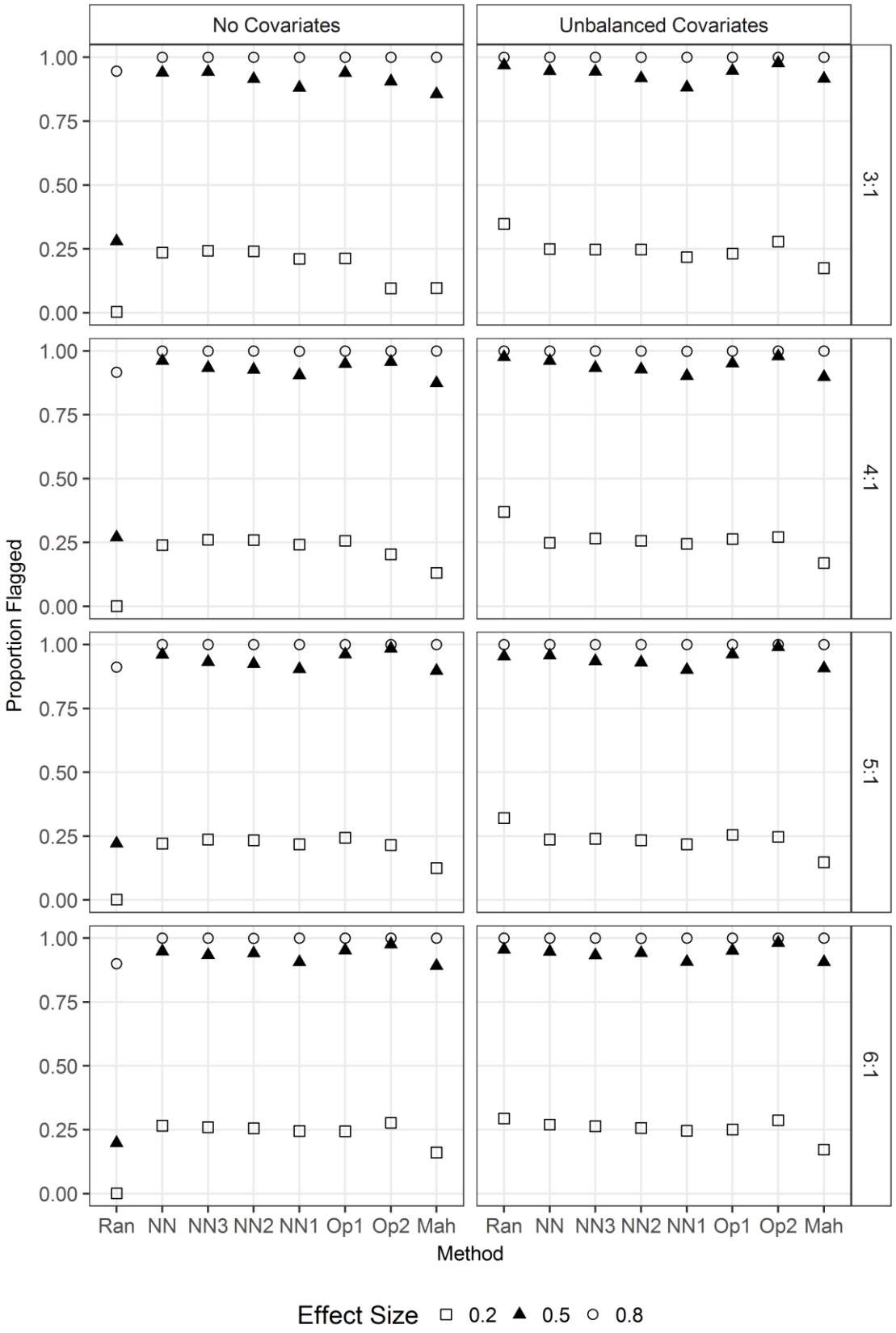

Figure 4. Power in the correct direction across conditions, treatment $N = 100$.

Research Question 3: Treatment Effect Recovery

The third research question was aimed at exploring how well the matching methods recovered the true treatment effect (e.g., differences between the group means). Recovery of the true treatment effect was determined by examining bias and RMSE of the effect size estimates. Bias and RMSE did not differ across effect sizes, thus, the results summarized below apply across effect sizes.

Bias. Bias is the difference between the estimated parameter and the generating true parameter value, averaged across replications. Thus, bias values closer to 0 are desirable, indicating that on average, the estimated parameter is approximately the same as the true parameter value. Given that the parameter of interest was the estimated treatment effect, bias is on a Cohen's *d* metric. As shown in Figure 5, bias was consistently negative. For an effect size of 0, negative bias indicated that the comparison group scored higher on the outcome than the treatment group. For other effect sizes (i.e., 0.2, 0.5, and 0.8), negative bias indicated that the treatment effect was estimated to be lower than the true treatment effect, which was simulated to be positive (favoring the treatment group). Prior to matching, the comparison group had higher values on the covariates. Thus, when matching did not completely balance the covariates, the comparison group's higher values on the covariates led to negatively biased estimates of the treatment effect.

Overall, bias was negligible for most of the matching methods. Random sampling was the most biased when covariates were not included in the outcome analysis; however when unbalanced covariates were included in the outcome analysis, bias for random sampling was close to 0. Additionally, bias for random sampling was comparable across

treatment group sample sizes and comparison-to-treatment ratios. Nearest neighbor matching without calipers was slightly negatively biased when covariates were not included in the outcome analyses. The inclusion of the unbalanced covariates resulted in bias closer to 0. Bias for nearest neighbor matching was smaller when treatment group sample size and comparison-to-treatment group ratio was larger. Additionally, the bias for nearest neighbor matching with calipers was close to 0 regardless of treatment group sample size, comparison-to-treatment ratio, and whether unbalanced were included in the outcome analyses. Optimal 1:1 matching was slightly biased, with greater bias when the treatment group sample size was 30 and when unbalanced covariates were not included in the outcome analyses. Optimal 2:1 and Mahalanobis distance matching had the largest bias (other than random sampling) and were comparable. Bias for these two matching techniques decreased as treatment group sample size and comparison-to-treatment ratio increased. Additionally, optimal 2:1 and Mahalanobis distance matching was less biased when unbalanced covariates were included in the outcome analyses.

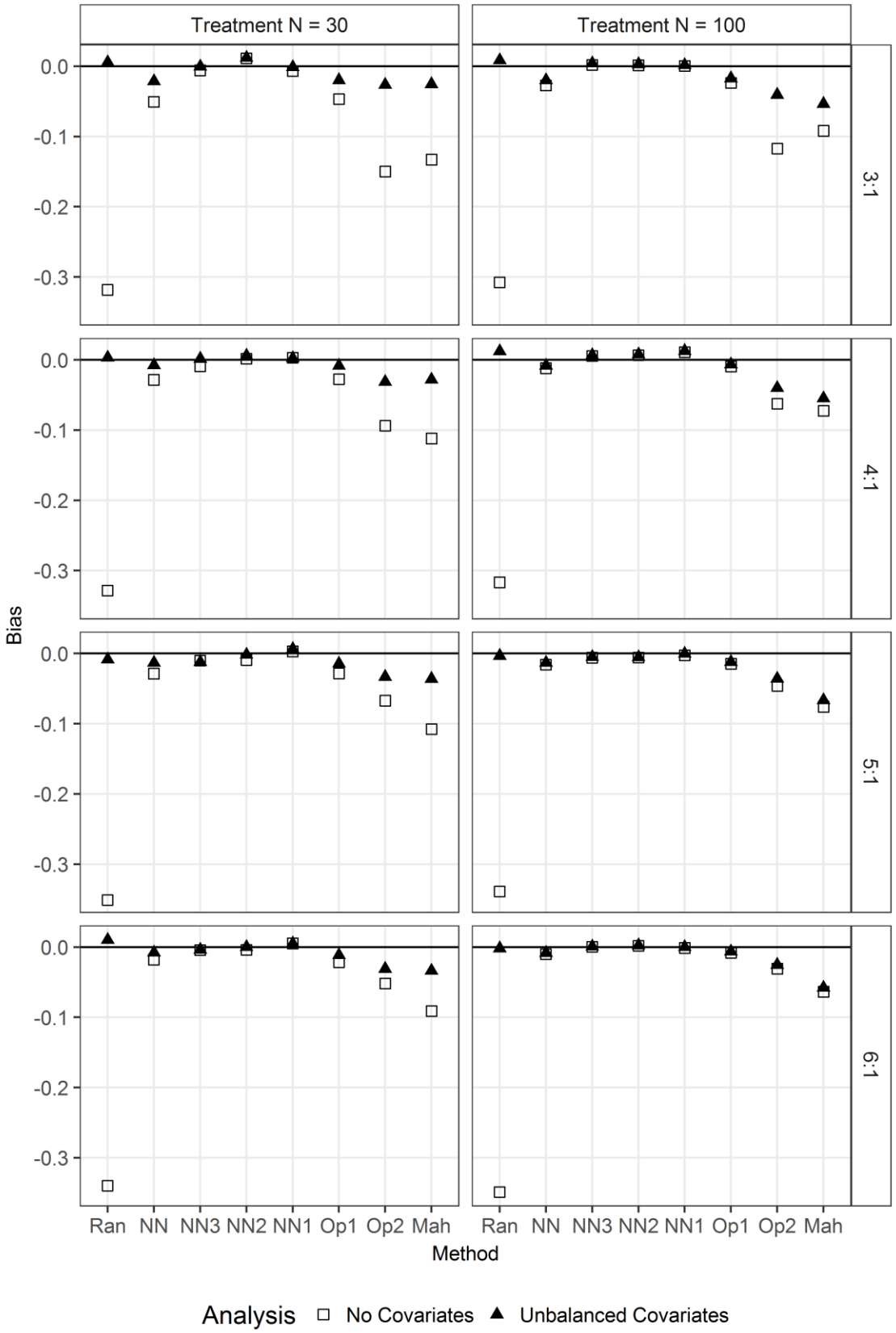

Figure 5. Treatment effect bias across conditions.

RMSE. RMSE is an index the combines bias and the average variability between the true and estimated parameters, across replications. Thus, RMSE values closer to 0 are desirable. Overall, there was more variability in RMSE across matching methods when the treatment group sample size was 30 than when the treatment group sample size was 100. When treatment group sample size was 100, the matching methods were comparable, except random sampling with no covariates in the outcome analyses. However, when treatment group sample size was 30, nearest neighbor matching with calipers resulted in the largest RMSE values. RMSE was smaller when treatment group sample size was larger (N = 100). Moreover, comparison-to-treatment ratio did not impact RMSE. Generally, whether unbalanced covariates were included in the outcome analysis did not affect RMSE, except for when random sampling was used. For random sampling, including unbalanced covariates in the outcome analyses resulted in smaller RMSE values than when unbalanced covariates were not included in the outcome analyses, largely due to the decrease in bias.

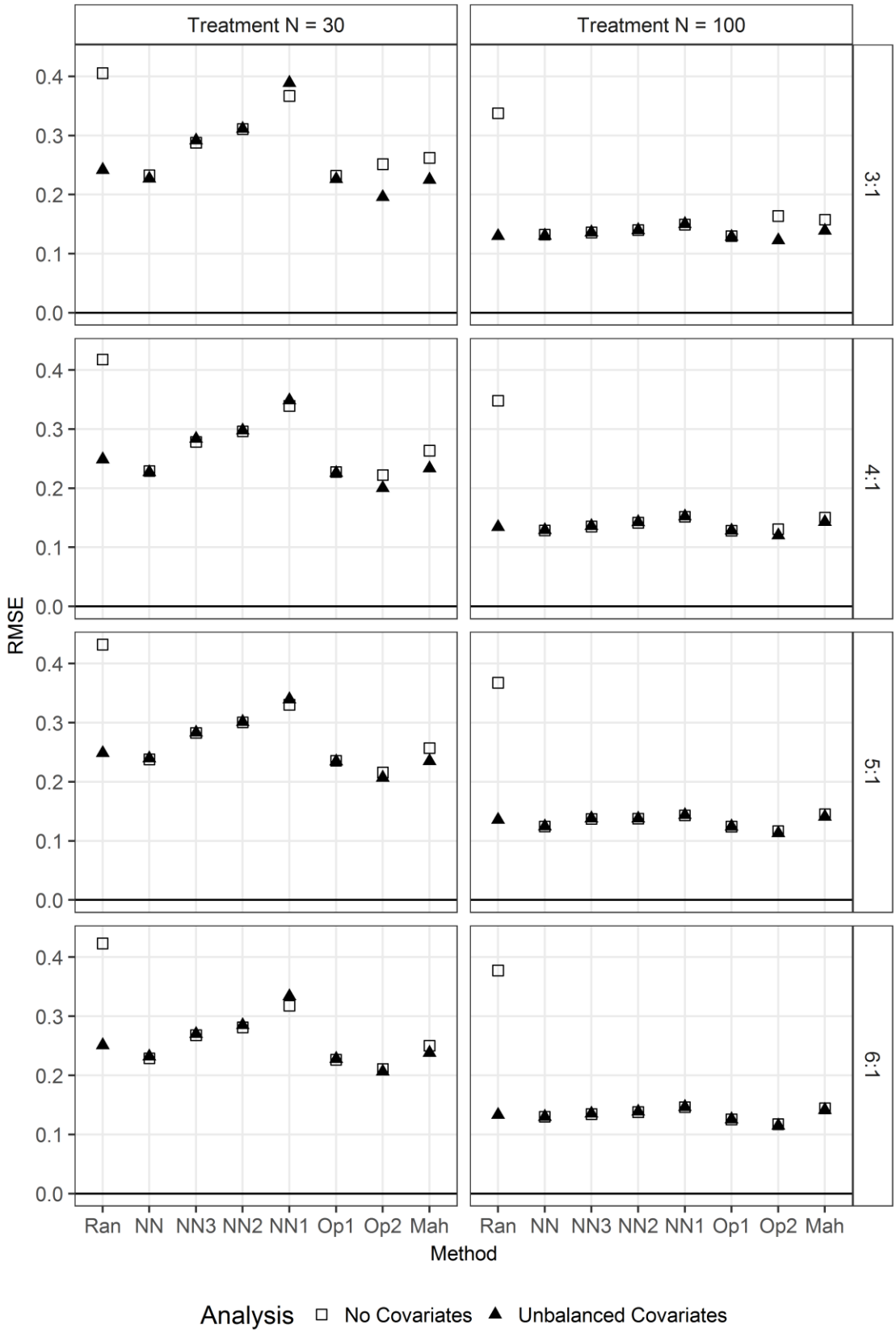

Figure 6. Treatment effect RMSE across conditions.

Research Question 4: Explaining Variability

The fourth research question concerned what conditions were optimal in obtaining accurate estimates of the effect size parameter. That is, how much of the variability in the difference between the estimated and true parameters could be explained by the manipulated conditions (e.g., effect size, matching method, comparison-to-treatment ratio, sample size, and outcome analysis)? This was assessed by examining the variability in the mean difference (similar to bias, but for one replication) and squared difference (similar to MSE, but for one replication) across replications. Most of the findings described below were hinted at in the descriptions of Figures 5 and 6; however, this analysis gives more precise values to how much variance was explained by each condition.

Table 10 presents the variance explained in the difference and square difference. Overall, most of the conditions did not impact true effect size recovery. Only three conditions explained more than 1% of the variance in the difference between the estimated and true effect sizes. The interaction of method by analysis ($\eta^2 = 5.68\%$) and the main effect for method ($\eta^2 = 5.63\%$) explained the most variance in the estimated and true parameter differences. Intuitively, this makes sense: the inclusion of unbalanced covariates in the outcome analysis (i.e., analysis) was more beneficial for some matching methods than others (e.g., including unbalanced covariates made more of a difference for random sampling than for nearest neighbor matching). The main effect for method indicates that the difference between the estimated and true effect sizes varied across methods, which is also to be expected. Further, the main effect for analysis ($\eta^2 = 1.64\%$) suggests that even though analysis interacted with matching method, averaging over the

methods the difference between the estimated and true effect sizes was closer to 0 when unbalanced covariates were included in the outcome analyses.

Only four conditions explained more than 1% of the variance in the squared difference between the estimated and true effect sizes. The main effect for treatment group sample size ($\eta^2 = 7.61\%$) and the main effect for method ($\eta^2 = 4.76\%$) explained the most variance in the squared difference between the estimated and true parameters across replications. The main effect for treatment group sample size indicates that the squared difference between the estimated and true effect sizes varied across the treatment group sample sizes. Again, this makes sense, as more variability would be expected when the sample size was smaller. The main effect for method suggests that the squared difference between the estimated and true effect sizes varied across methods, which is also to be expected. The interaction of method by analysis ($\eta^2 = 3.70\%$) indicates that the inclusion of unbalanced covariates in the outcome analysis (i.e., analysis) was more beneficial for some matching methods than others. Finally, the interaction of method by treatment group sample size ($\eta^2 = 1.27\%$) suggests that the squared difference between the estimated and true effect sizes across replication for the methods depended on the treatment group sample size.

Table 10

Variance Explained in the Estimated and True Effect Size Difference and Squared Difference Across Conditions

Condition	Difference	Squared Difference
d	0.02%	0.00%
Method	5.63%	4.76%
Treatment *N*	0.01%	7.61%
Ratio	0.03%	0.02%
Analysis	1.64%	0.61%
d * Method	0.00%	0.00%
d * Treatment *N*	0.00%	0.00%
d * Ratio	0.00%	0.00%
d * Analysis	0.00%	0.00%
Method * Treatment *N*	0.01%	1.27%
Method * Ratio	0.11%	0.12%
Method * Analysis	5.68%	3.70%
Treatment *N* * Ratio	0.00%	0.02%
Treatment *N* * Analysis	0.04%	0.00%
Ratio * Analysis	0.02%	0.00%
d * Method * Treatment *N*	0.00%	0.00%
d * Method * Ratio	0.00%	0.00%
d * Method * Analysis	0.00%	0.00%
d * Treatment *N* * Ratio	0.00%	0.00%
d * Treatment *N* * Analysis	0.00%	0.00%
d * Ratio * Analysis	0.00%	0.00%
Method * Treatment *N* * Ratio	0.01%	0.06%
Method * Treatment *N* * Analysis	0.05%	0.02%
Method * Ratio * Analysis	0.08%	0.04%
Treatment *N* * Ratio * Analysis	0.00%	0.00%
d * Method * Treatment *N* * Ratio	0.00%	0.00%
d * Method * Treatment *N* * Analysis	0.00%	0.00%
d * Method * Ratio * Analysis	0.00%	0.00%
d * Treatment *N* * Ratio * Analysis	0.00%	0.00%
Method * Treatment *N* * Ratio * Analysis	0.00%	0.01%
d * Method * Treatment *N* * Ratio * Analysis	0.00%	0.00%

Note. *d* = standardized effect size (Cohen's *d*), method = matching method, Treatment *N* = treatment group sample size, ratio = comparison-to-treatment ratio, and analysis = type of analysis (regression with no covariates or regression with unbalanced covariates).

CHAPTER 5

Discussion

The purpose of this study was to examine common matching techniques to determine how they differ in terms of the *quantity* and *quality* of matches and whether the results of subsequent group comparisons (e.g., significance test results, effect sizes) vary across the different matching techniques and manipulated conditions (i.e., effect size, treatment group sample size, comparison-to-treatment group ratio, and inclusion of unbalanced covariates in the outcome analyses).

Summary of Results

The first research question addressed how the matching methods differed in terms of the quality and quantity of matches. Although most of the matching techniques created matched comparison groups that were more equivalent to the treatment group than before matching, nearest neighbor matching with calipers resulted in the best quality matches (e.g., propensity score and individual covariate balance) compared to the other matching methods. Balance on the propensity scores and individual covariates was more favorable as the caliper became more stringent. Additionally, nearest neighbor and optimal 1:1 matching resulted in similar balance on the propensity scores and individual covariates. This suggests that treatment group members did not compete for comparison matches during the matching process. If there were insufficient overlap in propensity scores between the treatment and comparison group, the matches made earlier in the nearest neighbor matching process might be much better than the matches made later in the nearest neighbor matching process. Optimal matching would help balance this and thus might result in better matches overall. The relative comparability of matching for

nearest neighbor and optimal matching suggests that this was not the case. Moreover, optimal 2:1 and Mahalanobis distance matching resulted in comparable propensity score and individual covariate balance and performed worse than the other matching methods. Across matching methods, propensity score and individual covariate balance was slightly improved when the treatment group sample size was larger and when the comparison-to-treatment ratio was larger.

Importantly, X6 (the covariate representing race/ethnicity) was the least balanced across the matching methods. This may be due to the proportion of individuals in that group. It appears that matching methods balance better when group membership for categorical covariates are more equal (e.g., closer to a 50/50 split) than when group membership for the categorical covariates are unequal (e.g., closer to a 15/85 split). Further, in some replications, the entire representation of one group on X6 (in this case, African American representation) was excluded from analysis due to lack of an adequate match. This is particularly problematic for generalizability. That is, when the representation of one group is lost, then the results no longer generalize back to the original treatment group. Thus, the generalizability to the results are limited to the representation of the matched treatment group.

The full treatment group was retained for all of the matching techniques, except nearest neighbor with calipers. As the caliper became more stringent, the proportion of the treatment group that was successfully matched decreased. As the treatment group sample size and comparison-to-treatment group ratio decreased, the percentage of treatment group members who were successfully matched also decreased. This is particularly problematic as the treatment group was already fairly small. The loss of

treatment group members results in a loss of power, and could result in too few matches to conduct the outcome analyses of interest (e.g., in some replications, only seven treatment group members were successfully matched). Further, when treatment group members are excluded from the matched data, not only do researchers risk a loss in representativeness in the treatment group, but the treatment group members who are excluded likely had a higher propensity for treatment, thus the treatment group members that remain are the ones who had a lower propensity for being in the treatment group in the first place.

The second research question explored how the results of group comparisons varied across the matching methods, with a focus on Type I error and power. Type I error was highest for random sampling; however, the Type I error was close to 5% when unbalanced covariates were included in the outcome analyses. Additionally, optimal 2:1 and Mahalanobis distance matching resulted in slightly inflated Type I error rates. Type I error was within the nominal rate (e.g., around 5%) for the other matching methods. Treatment group sample size and comparison-to-treatment ratio made little difference in Type I error rates for all matching methods, except random sampling.

Unsurprisingly, power was lower when effect size and treatment group sample size was smaller. When the sample size was 100 and the effect size was 0.5 or 0.8, power across the different matching methods was close to 1, except for random sampling. Overall, nearest neighbor matching with calipers resulted in lower power than the other matching methods. Additionally, power decreased as the calipers became more stringent. This is unsurprising given that there was a loss of sample size when calipers were applied. Comparison-to-treatment group ratio did not affect power across the matching

methods. Further, when the effect size was 0.2, many of the matching methods had power in the incorrect direction; this was more problematic for random sampling than the other matching methods.

The third research question was concerned with the recovery of the true effect size, with a focus on bias and RMSE. Overall, bias was close to 0 for most of the matching methods, with optimal 2:1 and Mahalanobis distance matching being the most biased. Additionally, random sampling was biased when unbalanced covariates were not included in the outcome analyses. Inclusion of unbalanced covariates did not affect bias for the other matching methods. Further, treatment group sample size and comparison-to-treatment group ratio had little impact on bias across the matching methods.

RMSE was larger when the treatment group sample size was small. This was to be expected, as smaller sample sizes tend to result in more variability due to sampling error. There was more variability in RMSE across matching methods when the treatment group sample size was small. Matching methods were comparable when the treatment group sample size was large. Nearest neighbor matching with calipers resulted in the largest RMSE values when treatment group sample size was small. Moreover, RMSE was not affected by comparison-to-treatment ratio. Further, RMSE was not impacted by whether unbalanced covariates were included in the outcome analysis, except for random sampling.

The fourth research question addressed whether the variability in the differences between the estimated and true parameters across replications could be explained by the manipulated conditions (e.g., effect size, matching method, comparison-to-treatment ratio, sample size, and analysis technique). Method and analysis explained the most

variability in the difference and squared difference between the estimated and true parameters. Treatment group sample size also explained a notable percentage of the variability in the squared difference between the estimated and true parameters. These results make sense intuitively and relate to the implications for practice presented below.

Study Limitations and Future Research

As noted, this study was one in a line of research that is needed to provide guidance for practitioners on the selection of matching methods. The findings from this study demonstrated that matching method impacted Type I error, power, and estimated effect size; however; as with all studies, this study has a few notable limitations. First, this study included a small number of conditions. Also some conditions were adequately represented (e.g., comparison-to-treatment group ratio, effect sizes, and whether unbalanced covariates were included in the outcome analyses); other conditions were limited. For example, this study only examined two treatment group sample sizes. Additional treatment group sample sizes should be examined to ensure that these findings generalize. It would be useful to identify the minimum treatment group sample size necessary to obtain accurate effect size estimates.

Similarly, this study only included a limited number of matching methods, all of which are used to estimate ATT. It may be beneficial to examine how methods used to estimate ATE impacts Type I error, power, and estimated effect size. Additionally, nearest neighbor and optimal matching performed comparably, suggesting that treatment group members did not compete for comparison matches. This will not always be the case. Thus, it would be beneficial to determine how the competition for matches impacts the Type I error, power, and estimated effect size for these two techniques. Moreover,

when nearest neighbor matching with calipers was used, treatment group members were lost, but more importantly, representation of certain groups was decreased or lost, thus limiting generalizability. It would be useful to examine whether the same impacts on Type I error, power, and estimated effect size are observed when treatment group members are lost, but representation is *not* jeopardized.

A second limitation of this study was that the covariance matrix of the covariates was held constant across effect sizes. This might be realistic if the effect sizes corresponded to different dosages of the treatment; in this context, the size of the treatment effect should not be related to the correlations between the covariates and the outcome. However, if the varying effect sizes corresponded to the effect sizes in different populations, then the correlations might not be the same across effect sizes. The relationships among the covariates could be manipulated in future studies. Additionally, the relationships between the covariates and the outcome were held constant. It might also be useful to manipulate the relationships between the covariates and group selection, the outcome, or both. For example, if the covariates are highly correlated with group membership, but have low correlations with the outcome, then this may lead to more biased effect size estimates. This is an empirical question that could be answered in a future study.

A third limitation of this study is that it included a small number of covariates. Research has shown that the difference between Mahalanobis distance and propensity score matching is more pronounced when a large number (e.g., at least 8) of covariates are included in the matching model. Thus, future studies should include a larger number of covariates. Moreover, this study only included two categorical covariates—one with

about equal group proportions (60% and 40% split) and one with unequal group proportions (85% and 15% split). Future studies should include a larger number of categorical covariates with more variety in the group proportions.

A fourth limitation of this study, and propensity score matching studies in general, is that fit of the logistic regression model to predict group membership was not examined. Further, some of the covariates included in the matching model were not significant predictors of group membership. Although it is not common to examine fit in propensity score matching studies, this information is available to researchers who are interested in examining model-data fit. Future studies could include an examination of fit, as well as an examination of the utility of covariates for matching. Moreover, model-data fit and predictive utility of the covariates could be manipulated in future studies.

Implications for Practice

Although nearest neighbor with caliper resulted in the best propensity score and individual covariate balance (quality), the loss of treatment group members was concerning (quantity). Ultimately, it is up to the researcher to balance the quality and quantity of matches when creating a matched comparison group. If researchers are concerned with equity and representativeness (e.g., generalizability, internal validity), a matching technique that does not compromise quantity may be the most appropriate option. However, if researchers are concerned with obtaining groups that are equivalent on background and experience variables, then the use of a caliper may be most appropriate (e.g., external validity). Additionally, there may be expectations from funding agencies that require close balance. When treatment group members are excluded due to lack of an adequate match, it is important for the researcher to examine

the representativeness of the samples to ensure that generalizability is not limited. Researchers may opt for a combination of matching methods to ensure representativeness in the matched groups. For example, in some cases, it may be beneficial to use exact matching on certain variables (e.g., X6 in the current study), and then use propensity score matching with calipers to balance on the remaining covariates. This approach would help to ensure that X6 is represented and balanced in the final matched data set.

Matching method had an impact on Type I error, power, and estimated effect size but only in certain situations. This is unsurprising, given that different matching techniques may create comparison groups that are composed of different subsets of individuals from the entire comparison pool. Overall, random sampling, optimal 2:1 matching, and Mahalanobis distance matching performed worse than the other matching methods. Findings from this study suggest that they should not be used. Additionally, nearest neighbor matching with calipers did not perform as well as nearest neighbor and optimal 1:1 matching. This is likely due to the loss of the treatment group sample size and potentially the loss of representation among the treatment group members who were successfully matched. Matching method did not affect the outcome analyses (i.e., Type I error and power) when there was no effect ($d = 0$) or when there was at least a moderate effect ($d \geq 0.5$) *and* a large treatment group sample size ($N = 100$).

Treatment group sample size made some difference in the quality and quantity of matches and the significance tests and estimated effect sizes. Additional research is needed to determine the minimum sample size necessary to obtain accurate effect size estimates. Although comparison-to-treatment ratio resulted in some improvements across the conditions, the difference was minimal for this study. This provides some evidence

that a 3:1 ratio may be sufficient. Moreover, effect size made very little difference in this study, aside from its effect on power. If the researcher suspects that the effect size may be small, then the researcher should consider whether power should be addressed in other ways (e.g., increasing sample size or alpha). Further, whether unbalanced covariates were included in the outcome analyses made some difference in the quality and quantity of matches and the resulting effect size estimates. When unbalanced covariates are included in the analyses, the techniques becomes an ANCOVA. Thus, the researcher should only use this technique when the assumptions of ANCOVA are met.

Conclusions

In sum, the choice of matching technique is not without consequence. It dictates both the quality and quantity of the matches obtained and the resulting outcome analyses and estimated effect sizes. Although nearest neighbor matching with calipers tends to result in better matches, it can also result in the loss of treatment group members. When treatment group members are excluded from the matched groups due to lack of adequate match, the researcher should ensure that this does not impact generalizability of the results. If representation is compromised, the researcher may want to select a different matching method, such as nearest neighbor or optimal 1:1 matching. Otherwise, the matching methods appear to be comparable. Given that outcome variables are not used in the matching procedure, researchers can examine propensity score and covariates balance for the different matching methods and select the method that results in the best balance between the quality and quantity of matches (Ho et al., 2011). Although this is a difficult decision, it is up to the researcher to decide how to best balance the quality and

quantity of matches, while recognizing that this decision can impact the accuracy of the outcome analyses.

Endnotes

[1] Six continuous covariates were simulated; however two of the six covariates were dichotomized later in the data generation process. This resulted in four continuous covariates and two categorical covariates.

[2] The coefficients for Equation 12 were calculated as $\beta = S_{xx}^{-1} S_{px}$, where S_{xx} is the covariance matrix for the covariates in Table 1 and S_{px} is the vector of correlations between each covariate and probit, calculated by changing the observed point-biserial correlations to biserial correlations and the phi correlations to tetrachoric correlations. Thus, coefficients were equivalent to the standardized coefficients from a probit regression (where the error term is standard normal).

[3] This dichotomization was done after simulating the underlying likelihood of treatment group membership so that the coefficients could be left in terms of the continuous X5 and X6. Alternatively, the coefficients could have been transformed for the specific proportions used here, and the transformed coefficients could have been substituted in Equation 12 and applied to the dichotomous X5 and X6.

[4] Dichotomized covariates were used to simulate outcome scores.

[5] With an effect size of 0, the ~$N(0, 0.74)$ distribution yielded a total variance of 1 in Y. Because the mean value of Y was slightly lower in the treatment group, due to differences on the covariates, the within-group variance was slightly less than 1 before the final standardization.

[6] It is important to note that only the caliper methods can result in the loss of treatment group members; the other methods in the current study result in a match regardless of how close the matches are.

Appendix A

Syntax for Data Simulation and Analysis

R Code for Generating Data Sets

```
#########################################################################
#
# DISSERTATION: Propensity Score Matching Simulation Study
# Jessica Jacovidis
# February 10, 2017
#
#########################################################################
#########################################################################

# Setting working directory
setwd("E:/PSYC 900 - Dissertation/Syntax")

# Check working directory
# getwd()

source("simFunctions_DIS.R") #activate functions

#########################################################################

# Installing and loading necessary packages

#install.packages("mvtnorm")
require(mvtnorm)

#install.packages("psych")
require(psych)

#install.packages("OpenMx")
require(OpenMx)

#install.packages("MatchIt")
require(MatchIt)

#install.packages("optmatch")
require(optmatch)

#for write.fwf
#install.packages("gdata")
require(gdata)

#########################################################################
#########################################################################

# Preliminary Stuff before data simulation

# DO NOT MODIFY

#########################################################################
```

```
# before simulating anything, get the MREST-specific correlations
# get biserial rs from point biserial, and tetrachorics from phis
# math, verbal, cons, WA, MREST;
# observed correlations are condCorr, latent are corrCond

MfemP=.628
MblackP=.055
MfbCorr=-.01
MftCorr=-.03
MbtCorr=.128
MtreatP=86/3200
MfbP=(MfbCorr)*sqrt(MfemP*(1-MfemP)*MblackP*(1-MblackP))+MfemP*MblackP
#percent female and black
MftP=(MftCorr)*sqrt(MfemP*(1-MfemP)*MtreatP*(1-MtreatP))+MfemP*MtreatP
#percent female and treatment
MbtP=(MbtCorr)*sqrt(MtreatP*(1-MtreatP)*MblackP*(1-
MblackP))+MtreatP*MblackP #percent black and treatment
MgenderCorr=c(-.258,-.077,.166,-.239,.015)
MblackCorr=c(-.198,-.130,.007,-.019,-.113)
MtreatCorr=c(-.088,-.146,.028,.017)
z=qnorm(MfemP)
ordinate=1/sqrt(2*pi)*exp(-z^2/2)
corrGender=sqrt(MfemP*(1-MfemP))*MgenderCorr/ordinate
z=qnorm(MblackP)
ordinate=1/sqrt(2*pi)*exp(-z^2/2)
corrBlack=sqrt(MblackP*(1-MblackP))*MblackCorr/ordinate
temp=tetrachoric(c(MfemP,MblackP,MfbP))
corrFB=temp$rho
temp=tetrachoric(c(MfemP,MtreatP,MftP))
corrFT=temp$rho
temp=tetrachoric(c(MtreatP,MblackP,MbtP))
corrBT=temp$rho

#corrGender
#corrBlack
#c(corrFB,corrFT,corrBT)

#corrX1: categorical variables are latent
#use this only for the multivariate normal draws
#corrX2: categorical variables are observed categories, so depends on
choice of simulated percent

corrX1=matrix(c(1,.430,-.122,.152,corrGender[1],corrBlack[1],
              .430,1, -.116,.094,corrGender[2],corrBlack[2],
              -.122,-.116,1,-.372,corrGender[3],corrBlack[3],
               .152,.094,-.372,1,corrGender[4],corrBlack[4],
               corrGender[1:4],1,corrFB,
               corrBlack[1:4],corrFB,1),6,6)
covX1=corrX1
#now correlation with outcome;
corrXY1=c(.323,.482,-.015,.054,corrGender[5],corrBlack[5])
Ycoef1=solve(corrX1) %*% corrXY1  #solve means inverse
Ycoef1  #this is the model for simulating MREST scores, with gender and
Black as continuous latent variables

#now correlation with propensity, as a normal variable, not a logistic
variable
```

```
#start with observed correlation with treatment
obscorrXP=c(-.088,-.146,.028,.017)
temp=MtreatP
z=qnorm(MtreatP)
ordinate=1/sqrt(2*pi)*exp(-z^2/2)
temp=sqrt(MtreatP*(1-MtreatP))*obscorrXP/ordinate
corrXP=c(temp,corrFT,corrBT)
Pcoef=solve(corrX1) %*% corrXP #solve means inverse
Pcoef  #this is the model for simulating propensity scores, with gender
and Black as continuous latent variables
covX1=corrX1   #because standardized variables
#theoretical cov between normalp and Y estimated when cat variables
latent
temp=Pcoef %*% t(Ycoef1)
temp2=temp*covX1 #If some covariates are used only for predicting
propensity or only for Y, need to pull out the appropriate elements
from X
covnormPY=sum(temp2)

#########################################################################
#########################################################################

# Additional Preliminary Stuff before data simulation

# NEED TO MODIFY:

# treatP to reflect the proportion of the sample that is treatment
# VARIES BY CONDITION

#########################################################################

#now work out population values for specific conditions
treatP=.142857
femP=.60
blackP=.15
lbound <- c(-Inf, -Inf ) # Integrate from -Infinity to 0 on first
variable
ubound <- c(qnorm(femP), qnorm(blackP)) # From 0 to +Infinity on
second, and from 1 to 2.5 on third
fbP=omxMnor(matrix(c(1,corrFB,corrFB,1),nrow=2,ncol=2), c(0,0), lbound,
ubound)
#expected value of the observed r between dichotomous female and black
fbCorr=(fbP-femP*blackP)/sqrt(femP*(1-femP)*blackP*(1-blackP))
z=qnorm(femP)
ordinate=1/sqrt(2*pi)*exp(-z^2/2)
genderCorr=corrGender*ordinate/sqrt(femP*(1-femP))
z=qnorm(blackP)
ordinate=1/sqrt(2*pi)*exp(-z^2/2)
blackCorr=corrBlack*ordinate/sqrt(blackP*(1-blackP))


#now observed correlation matrix in the population
#if the continuous covariates have error, divide by reliability here
#otherwise, just replace tetrachorics with phis for a given condition
corrX2=matrix(c(1,.430,-.122,.152,genderCorr[1],blackCorr[1],
              .430,1, -.116,.094,genderCorr[2],blackCorr[2],
              -.122,-.116,1,-.372,genderCorr[3],blackCorr[3],
```

```
               .152,.094,-.372,1,genderCorr[4],blackCorr[4],
               genderCorr[1:4],1,fbCorr,
               blackCorr[1:4],fbCorr,1),6,6)
corrX2 #this wont vary with the treatment proportion
#now correlation with outcome; #could adjust for unreliability in X or
Y
corrXY2=c(.323,.482,-.015,.054,genderCorr[5],blackCorr[5])


Ycoef2=solve(corrX2) %*% corrXY2
Ycoef2 #observed coefficients for simulating MREST
covX2=corrX2  #this will only work if categorical variables have been
std too
#explained variance in Y--expected value of the observed variables--
need this later for simulation
temp=Ycoef2 %*% t(Ycoef2)*covX2 #NOT matrix multiplication
varExpY=sum(temp)

#variance in normalP
temp=Pcoef%*%t(Pcoef)*covX1 #NOT matrix multiplication  #adjust if not
all covariates used in propensity
normPvar=sum(temp)+1  #add in error variance
Yvar=1 #because working with correlation matrix -- have to calculate if
cov matrix
corrnormPY=covnormPY/sqrt(Yvar*normPvar) #biserial correlation
#now theoretical point-biserial corr between observed group and Y
z=qnorm(treatP)
ordinate=1/sqrt(2*pi)*exp(-z^2/2)
corrGY=corrnormPY*ordinate/sqrt(treatP*(1-treatP))
##std group difference on Y due ONLY to the covariates (does not
include d)
#YdiffCov=2*corrGY/sqrt(1-corrGY^2)
YdiffCov=corrGY/sqrt((1-corrGY^2)*treatP*(1-treatP))
#YdiffCov


########################################################################
########################################################################

############  Now simulate data. Loop for replications will start here

########################################################################
########################################################################

# NEED TO MODIFY:

# d to reflect effect size
# VARIES BY CONDITION

# Nexaminee to reflect the total sample size
# VARIES BY CONDITION

# THERE IS A NOTE BELOW THAT WE NEED TO CHANGE d IN THE SIMFUN_DIS FILE
# BUT I DON'T THINK WE DO.  IT SEEMS THAT d AND Nexaminee IS PASSED TO
# THE SIMFUN_DIS FILE, SO WE ONLY NEED TO SET IT HERE.

# NEED TO ADD TO THE NAMING CONVENTION BELOW (TO SAVE OUT FILE)
```

```
# HOW DO WE DO THAT?

#########################################################################

#remember to change d (second argument to simfun)

d=0.8
Nexaminee=700

set.seed(80313)
for (rep in 1:1000) {
   #rm(X)  #because I kept regenerating X and wanted to clear it out--
move to end of loop

   simdat<-simfun(Nexaminee,d)

   random<-RandomSamp(simdat)
   random=subset(random,select=c(ID,random))
   colnames(random)=c("ID","random")

   NN<-NNmatch(simdat)
   NN=subset(NN,select=c(ID,distance, weights))
   colnames(NN)=c("ID","NNdist","NNwgt")

   # Compute the SD of the Propensity Scores to create calipers
   ps.sd = sd(NN$NNdist)
   ps.sd

   NN3<-NN3match(simdat)
   NN3=subset(NN3,select=c(ID,distance, weights))
   colnames(NN3)=c("ID","NN3dist","NN3wgt")

   NN2<-NN2match(simdat)
   NN2=subset(NN2,select=c(ID,distance, weights))
   colnames(NN2)=c("ID","NN2dist","NN2wgt")

   NN1<-NN1match(simdat)
   NN1=subset(NN1,select=c(ID,distance, weights))
   colnames(NN1)=c("ID","NN1dist","NN1wgt")

   Opt1<-Opt1Match(simdat)
   Opt1=subset(Opt1,select=c(ID,distance, weights))
   colnames(Opt1)=c("ID","Opt1dist","Opt1wgt")

   Opt2<-Opt2Match(simdat)
   Opt2=subset(Opt2,select=c(ID,distance, weights))
   colnames(Opt2)=c("ID","Opt2dist","Opt2wgt")

   Mahal<-MahalMatch(simdat)
   Mahal<-subset(Mahal,select=c(ID, weights))
   colnames(Mahal)=c("ID", "Mahalwgt")

   #Merge together files
   alldat<-AllMerge()

#save(alldat, file=paste0("resultsD", as.integer(d), rep, ".Rdata"))
#or
```

```
#write.fwf(data.frame(rep,d,alldat), file=paste0("resultsD", d*10, "r",
rep, ".dat"),
# append=TRUE, colnames=TRUE)
write.table(data.frame(rep,d,alldat), file=paste0("resultsD", d*10,
"N", Nexaminee, "r", rep, ".dat"),
 quote=FALSE, sep="\t", row.names=FALSE, col.names=TRUE,na = ".")
}

##############################################################################
##############################################################################
```

R Code for Simulation Function

```
##############################################################################
##############################################################################

# SIMULATION FUNCTION STARTS HERE

##############################################################################
##############################################################################

simfun <- function(Nexaminee,d) {
X=rmvnorm(Nexaminee, rep(0,6), covX1, method="chol")
#X=data.frame(X) #so I can just use the name X1, etc.
#str(X) #checking
mycut=qnorm(1-femP)
female01=ifelse(X[,5]>mycut,1,0)
mycut=qnorm(1-blackP)
black01=ifelse(X[,6]>mycut,1,0)
oldX=X #backup
female=(female01-mean(female01))/sd(female01)
black=(black01-mean(black01))/sd(black01)
X[,5]=female
X[,6]=black
#table(X[,6])
Perr=rnorm(Nexaminee) # "error" in propensity score
normalP= as.vector(oldX %*% Pcoef + Perr) #these are the coefficients
for the latent categorical variables
#Pcoef
#standardize
#variance in logitP
temp=Pcoef%*%t(Pcoef)*covX1 #NOT matrix multiplication  #adjust if not
all covariates used in propensity
normPvar=sum(temp)+1  #add in error variance approximating logistic
error
normalP=normalP/sqrt(normPvar)
#var(normalP)
#mean(normalP)
mycut=quantile(normalP,1-treatP)
group=ifelse(normalP>mycut,1,0)
#mean(group)
#by(X,group,colMeans)
#Pb=glm(formula=group ~ X1+X2+X3+X4+X5+X6, data=data.frame(oldX),
family=binomial)
#Pb$coefficients
```

```
#temp=Pb$coefficients[2:7]
#temp/Pcoef
e=rnorm(Nexaminee)*sqrt(1-varExpY)
#this makes the within-group variance 1
Y=as.vector((d*group +(X %*% Ycoef2 +e)/sqrt(Yvar))*sqrt(1+treatP*(1-
treatP)*YdiffCov^2))
mydata=data.frame(1:Nexaminee,Y,X,female01,black01,group)
colnames(mydata)[1]="ID"
mydata
}   #end simulate data function


#############################################################################
#############################################################################

# MATCHING FUNCTIONS START HERE

#############################################################################
#############################################################################

# Random Matching

RandomSamp <- function(mydata) {
random2<-subset(mydata,group==1)
temp=dim(random2)
numTreat=temp[1]
random1<-subset(mydata,group==0)
random1<-random1[sample(1:nrow(random1), numTreat, replace=FALSE),]
tryrandom<-rbind(random1,random2)
tryrandom$random<-1
tryrandom
} #end random match function

#############################################################################

# Nearest Neighbor PSM

NNmatch <- function(mydata) {
try1_NN = matchit(group~X1+X2+X3+X4+X5+X6,method="nearest",
data=mydata, ratio=1)
try1_NN
summary(try1_NN)
NN<-match.data(try1_NN)
#tapply (NN$distance,NN$ATHLETE, var)
#plot(try1_NN, type="jitter")
#write.csv(NN, file="try1_NN.csv")
} #end NN match function

#############################################################################

# Nearest Neighbor PSM with .3 caliper

NN3match <- function(mydata) {
try1_NNCAL3=matchit(group~X1+X2+X3+X4+X5+X6,method="nearest",
data=mydata, ratio=1, caliper = 0.30*ps.sd)
try1_NNCAL3
summary(try1_NNCAL3)
```

```
NN3<-match.data(try1_NNCAL3)
#tapply (NNCL3$distance,NNCL3$ATHLETE, var)
#plot(try1_NNCAL3, type="jitter")
#write.csv(NNCL3, file="try1_NNCAL3.csv")
} #end NN .3 caliper match function

#########################################################################

# Nearest Neighbor PSM with .2 caliper

NN2match <- function(mydata) {
try1_NNCAL2=matchit(group~X1+X2+X3+X4+X5+X6,method="nearest",
data=mydata, ratio=1, caliper = 0.20*ps.sd)
try1_NNCAL2
summary(try1_NNCAL2)
NN2<-match.data(try1_NNCAL2)
#tapply (NNCL2$distance,NNCL2$ATHLETE, var)
#plot(try1_NNCAL2, type="jitter")
#write.csv(NNCL2, file="try1_NNCAL2.csv")
} #end NN .2 caliper match function

#########################################################################

# Nearest Neighbor PSM with .1 caliper

NN1match <- function(mydata) {
try1_NNCAL1=matchit(group~X1+X2+X3+X4+X5+X6,method="nearest",
data=mydata, ratio=1, caliper = 0.10*ps.sd)
try1_NNCAL1
summary(try1_NNCAL1)
NN1<-match.data(try1_NNCAL1)
#tapply (NNCL1$distance,NNCL1$ATHLETE, var)
#plot(try1_NNCAL1, type="jitter")
#write.csv(NNCL1, file="try1_NNCAL1.csv")
} #end NN .1 caliper match function

#########################################################################

# Optimal Matching (1:1)

Opt1Match <- function(mydata) {
try1_OPTIMAL = matchit(group~X1+X2+X3+X4+X5+X6,method="optimal",
data=mydata, ratio=1)
try1_OPTIMAL #1:1 ratio
summary(try1_OPTIMAL)
Opt1<-match.data(try1_OPTIMAL)
} #end Optimal 1:1 match function

#########################################################################

# Optimal Matching (2:1)

Opt2Match <- function(mydata) {
try1_OPTIMAL2 = matchit(group~X1+X2+X3+X4+X5+X6,method="optimal",
data=mydata, ratio=2)
try1_OPTIMAL2 #2:1 ratio
summary(try1_OPTIMAL2)
```

```
Opt2<-match.data(try1_OPTIMAL2)
} #end Optimal 2:1 match function


#########################################################################

# Mahalanobis Distance Matching
# This is the syntax suggested by Kosuke Imai on the MatchIt listserve

MahalMatch <- function(mydata) {
try1_Mahal = matchit(group~X1+X2+X3+X4+X5+X6, data=mydata,
method="nearest", distance="mahalanobis",
mahvars=c("X1","X2","X3","X4","X5","X6"),ratio=1, caliper=1000)
try1_Mahal
summary(try1_Mahal)
Mahal<-match.data(try1_Mahal)
} #end Mahalanobis Distance match function

#########################################################################
#########################################################################

# MERGING FUNCTION STARTS HERE

#########################################################################
#########################################################################

AllMerge<-function(){
   alldat=merge(simdat,random,by="ID",all=TRUE)
   alldat=merge(alldat,NN,by="ID",all=TRUE)
   alldat=merge(alldat,NN3,by="ID",all=TRUE)
   alldat=merge(alldat,NN2,by="ID",all=TRUE)
   alldat=merge(alldat,NN1,by="ID",all=TRUE)
   alldat=merge(alldat,Opt1,by="ID",all=TRUE)
   alldat=merge(alldat,Opt2,by="ID",all=TRUE)
   alldat=merge(alldat,Mahal,by="ID",all=TRUE)
alldat
}

#########################################################################
#########################################################################
```

SAS Syntax for Data Analysis

```
options nocenter;
options nonotes;
%let path=E:\PSYC 900 - Dissertation\Data;
libname lib1 "E:\PSYC 900 - Dissertation";

*Macro to read in the data;
*Need to change the value of d;
*Will save out a complete file for each d;
%macro readin(values);
%let D=8;
%let count=%sysfunc(countw(&values));
%do i = 1 %to &count;
```

```
%let value=%qscan(&values,&i,%str(,));
%put &value;
%do rep=1 %to 1000;
data d1;
infile "&path\resultsD&D.N&value.r&rep..dat" missover firstobs=2
dlm='09'x ;
input rep d ID Y X1 X2 X3 X4 X5 X6 female01 black01 group random
NNdist NNwgt NN3dist NN3wgt NN2dist NN2wgt NN1dist
NN1wgt Opt1dist Opt1wgt Opt2dist Opt2wgt Mahalwgt;
totalN=&value;
run;
proc datasets nolist; append base=lib1.D&d data=d1; run;
proc datasets nolist;  delete d1; run;
%end;
%end;
%end;
%mend;

%readin(%str(120,150,180,210,400,500,600,700));

*Concatenating SAS files for each effect size;
*make each matching method into a record;
*Creating a new variable for treatment group sample size and
comparison-to-treatment ratio;
data d0(drop=i random NNwgt NN3wgt NN2wgt NN1wgt Opt1wgt Opt2wgt
Mahalwgt
  randist NNdist NN3dist NN2dist NN1dist Opt1dist Opt2dist Maldist);
set lib1.D0 lib1.D2 lib1.D5 lib1.D8;
length method $3;
randist=.; *no distance measures for random, but want variable;
Maldist=.;
treatN=.;
ratio=.;
if totalN=120 then treatN=30;
if totalN=120 then ratio=3;
if totalN=150 then treatN=30;
if totalN=150 then ratio=4;
if totalN=180 then treatN=30;
if totalN=180 then ratio=5;
if totalN=210 then treatN=30;
if totalN=210 then ratio=6;
if totalN=400 then treatN=100;
if totalN=400 then ratio=3;
if totalN=500 then treatN=100;
if totalN=500 then ratio=4;
if totalN=600 then treatN=100;
if totalN=600 then ratio=5;
if totalN=700 then treatN=100;
if totalN=700 then ratio=6;
array mwgt[*] random NNwgt NN3wgt NN2wgt NN1wgt Opt1wgt Opt2wgt
Mahalwgt;
array dist[*] randist NNdist NN3dist NN2dist NN1dist Opt1dist Opt2dist
Maldist;
array mname[8] $ _temporary_ ("Ran", "NN0", "NN3", "NN2", "NN1", "Op1",
"Op2", "Mal");
do i = 1 to 8;
 method=mname[i];
```

```
 if mwgt[i]=1 then select=1; else select=0;
 distance=dist[i];
 output;
end;
run;

data lib1.raw; set d0; run;
proc datasets library=lib1;
 modify raw;
 index create method ;
  index create d ;
   index create treatN ;
    index create ratio ;
       index create rep ;
        index create select;
 run;

proc sort; by method d treatN ratio rep group;
run;

data Pred1; set d0;
 proc means noprint;
 by method d treatN ratio rep group;
 var X1;
 output out=PreX1Res mean=PreX1mean var=PreX1var min=PreX1min
max=PreX1max;
 run;
 data PreX1a(drop=_type_ _freq_ PreX1mean PreX1var PreX1min PreX1max);
set PreX1Res;
 if group=0;
 cX1meanPre=PreX1mean;
 cX1varPre=PreX1var;
 cX1minPre=PreX1min;
 cX1maxPre=PreX1max;
 data PreX1b(drop=_type_ _freq_ PreX1mean PreX1var PreX1min PreX1max);
set PreX1Res;
 if group=1;
 tX1meanPre=PreX1mean;
 tX1varPre=PreX1var;
 tX1minPre=PreX1min;
 tX1maxPre=PreX1max;
run;

data Pred2; set d0;
 proc means noprint;
 by method d treatN ratio rep group;
 var X2;
 output out=PreX2Res mean=PreX2mean var=PreX2var min=PreX2min
max=PreX2max;
 run;
 data PreX2a(drop=_type_ _freq_ PreX2mean PreX2var PreX2min PreX2max);
set PreX2Res;
 if group=0;
 cX2meanPre=PreX2mean;
 cX2varPre=PreX2var;
 cX2minPre=PreX2min;
 cX2maxPre=PreX2max;
```

```
 data PreX2b(drop=_type_ _freq_ PreX2mean PreX2var PreX2min PreX2max);
set PreX2Res;
 if group=1;
 tX2meanPre=PreX2mean;
 tX2varPre=PreX2var;
 tX2minPre=PreX2min;
 tX2maxPre=PreX2max;
run;

data Pred3; set d0;
 proc means noprint;
 by method d treatN ratio rep group;
 var X3;
 output out=PreX3Res mean=PreX3mean var=PreX3var min=PreX3min
max=PreX3max;
 run;
 data PreX3a(drop=_type_ _freq_ PreX3mean PreX3var PreX3min PreX3max);
set PreX3Res;
 if group=0;
 cX3meanPre=PreX3mean;
 cX3varPre=PreX3var;
 cX3minPre=PreX3min;
 cX3maxPre=PreX3max;
 data PreX3b(drop=_type_ _freq_ PreX3mean PreX3var PreX3min PreX3max);
set PreX3Res;
 if group=1;
 tX3meanPre=PreX3mean;
 tX3varPre=PreX3var;
 tX3minPre=PreX3min;
 tX3maxPre=PreX3max;
run;

 data Pred4; set d0;
 proc means noprint;
 by method d treatN ratio rep group;
 var X4;
 output out=PreX4Res mean=PreX4mean var=PreX4var min=PreX4min
max=PreX4max;
 run;
 data PreX4a(drop=_type_ _freq_ PreX4mean PreX4var PreX4min PreX4max);
set PreX4Res;
 if group=0;
 cX4meanPre=PreX4mean;
 cX4varPre=PreX4var;
 cX4minPre=PreX4min;
 cX4maxPre=PreX4max;
 data PreX4b(drop=_type_ _freq_ PreX4mean PreX4var PreX4min PreX4max);
set PreX4Res;
 if group=1;
 tX4meanPre=PreX4mean;
 tX4varPre=PreX4var;
 tX4minPre=PreX4min;
 tX4maxPre=PreX4max;
run;

data Pred5; set d0;
proc means noprint;
```

```
 by method d treatN ratio rep group;
 var female01 black01;
 output out=PreCatRes mean=PreX5mean PreX6mean var=PreX5var PreX6var;
 run;
 data PreX5a(drop=_type_ _freq_ PreX5mean PreX5var PreX6mean PreX6var);
set PreCatRes;
 if group=0;
 cX5meanPre=PreX5mean;
 cX5varPre=PreX5var;
 cX6meanPre=PreX6mean;
 cX6varPre=PreX6var;
 data PreX5b(drop=_type_ _freq_ PreX5mean PreX5var PreX6mean PreX6var);
set PreCatRes;
 if group=1;
 tX5meanPre=PreX5mean;
 tX5varPre=PreX5var;
 tX6meanPre=PreX6mean;
 tX6varPre=PreX6var;
run;

 data Pred7; set d0;
 proc means noprint;
 by method d treatN ratio rep group;
 var Y;
 output out=PreYRes mean=PreYmean var=PreYvar N=PreN;
 run;
 data PreYa(drop=_type_ _freq_ PreYmean PreYvar PreN); set PreYRes;
 if group=0;
 cmeanPre=PreYmean;
 cvarPre=PreYvar;
 cNPre=PreN;
 data PreYb(drop=_type_ _freq_ PreYmean PreYvar PreN); set PreYRes;
 if group=1;
 tmeanPre=PreYmean;
 tvarPre=PreYvar;
 tNPre=PreN;

data d1prop; set d0;
 if select=1;
 proc means noprint;
 by method d treatN ratio rep group;
 var distance;
 output out=prop mean=Propmean var=Propvar;
 run;
data tempppropa(drop=_type_ _freq_ Propmean Propvar); set prop;
 if group=0;
 cPropmean=Propmean;
 cPropvar=Propvar;
 data tempppropb(drop=_type_ _freq_ Propmean Propvar); set prop;
 if group=1;
 tPropmean=Propmean;
 tPropvar=Propvar;
run;

data d1; set d0;
 if select=1;
 proc means noprint;
```

```
 by method d treatN ratio rep group;
 var X1;
 output out=X1results mean=X1mean var=X1var min=X1min max=X1max;
 run;
 data tempX1a(drop=_type_ _freq_ X1mean X1var X1min X1max); set
X1results;
 if group=0;
 cX1mean=X1mean;
 cX1var=X1var;
 cX1min=X1min;
 cX1max=X1max;
 data tempX1b(drop=_type_ _freq_ X1mean X1var X1min X1max); set
X1results;
 if group=1;
 tX1mean=X1mean;
 tX1var=X1var;
 tX1min=X1min;
 tX1max=X1max;
run;

data d2; set d0;
 if select=1;
 proc means noprint;
 by method d treatN ratio rep group;
 var X2;
 output out=X2results mean=X2mean var=X2var min=X2min max=X2max;
 run;
 data tempX2a(drop=_type_ _freq_ X2mean X2var X2min X2max); set
X2results;
 if group=0;
 cX2mean=X2mean;
 cX2var=X2var;
 cX2min=X2min;
 cX2max=X2max;
 data tempX2b(drop=_type_ _freq_ X2mean X2var X2min X2max); set
X2results;
 if group=1;
 tX2mean=X2mean;
 tX2var=X2var;
 tX2min=X2min;
 tX2max=X2max;
run;

data d3; set d0;
 if select=1;
 proc means noprint;
 by method d treatN ratio rep group;
 var X3;
 output out=X3results mean=X3mean var=X3var min=X3min max=X3max;
 run;
 data tempX3a(drop=_type_ _freq_ X3mean X3var X3min X3max); set
X3results;
 if group=0;
 cX3mean=X3mean;
 cX3var=X3var;
 cX3min=X3min;
 cX3max=X3max;
```

```
 data tempX3b(drop=_type_ _freq_ X3mean X3var X3min X3max); set
X3results;
 if group=1;
 tX3mean=X3mean;
 tX3var=X3var;
 tX3min=X3min;
 tX3max=X3max;
run;

 data d4; set d0;
 if select=1;
 proc means noprint;
 by method d treatN ratio rep group;
 var X4;
 output out=X4results mean=X4mean var=X4var min=X4min max=X4max;
 run;
 data tempX4a(drop=_type_ _freq_ X4mean X4var X4min X4max); set
X4results;
 if group=0;
 cX4mean=X4mean;
 cX4var=X4var;
 cX4min=X4min;
 cX4max=X4max;
 data tempX4b(drop=_type_ _freq_ X4mean X4var X4min X4max); set
X4results;
 if group=1;
 tX4mean=X4mean;
 tX4var=X4var;
 tX4min=X4min;
 tX4max=X4max;
run;

data d5; set d0;
 if select=1;
proc means noprint;
 by method d treatN ratio rep group;
 var female01 black01;
 output out=CatRes mean=X5mean X6mean var=X5var X6var;
 run;
 data tempX5a(drop=_type_ _freq_ X5mean X5var X6mean X6var); set
CatRes;
 if group=0;
 cX5mean=X5mean;
 cX5var=X5var;
 cX6mean=X6mean;
 cX6var=X6var;
 data tempX5b(drop=_type_ _freq_ X5mean X5var X6mean X6var); set
CatRes;
 if group=1;
 tX5mean=X5mean;
 tX5var=X5var;
 tX6mean=X6mean;
 tX6var=X6var;
run;

data d7; set d0;
 if select=1;
```

```
 proc means noprint;
 by method d treatN ratio rep group;
 var Y;
 output out=Yresults mean=Ymean var=wvar N=nstud;
 run;
 data tempYa(drop=_type_ _freq_ Ymean wvar nstud); set Yresults;
 if group=0;
 cmean=Ymean;
 cvar=wvar;
 cN=nstud;
 data tempYb(drop=_type_ _freq_ Ymean wvar nstud); set Yresults;
 if group=1;
 tmean=Ymean;
 tvar=wvar;
 tN=nstud;

data d8; merge PreX1a PreX1b PreX2a PreX2b PreX3a PreX3b PreX4a PreX4b
PreX5a PreX5b PreYa PreYb tempPropa tempPropb
tempX1a tempX1b tempX2a tempX2b tempX3a tempX3b tempX4a tempX4b tempX5a
tempX5b tempYa tempYb;
by method d treatN ratio rep;
 PreX1diff=tX1meanPre-cX1meanPre;
 PreX2diff=tX2meanPre-cX2meanPre;
 PreX3diff=tX3meanPre-cX3meanPre;
 PreX4diff=tX4meanPre-cX4meanPre;
 PreX5diff=tX5meanPre-cX5meanPre;
 PreX6diff=tX6meanPre-cX6meanPre;
 PreYdiff=tmeanPre-cmeanPre;
 Propdiff=tPropmean-cPropmean;
 VarRatio=tPropvar/cPropvar;
 X1diff=tX1mean-cX1mean;
 X2diff=tX2mean-cX2mean;
 X3diff=tX3mean-cX3mean;
 X4diff=tX4mean-cX4mean;
 X5diff=tX5mean-cX5mean;
 X6diff=tX6mean-cX6mean;
 X1pooledVar=((tN-1)*tX1var+(cN-1)*cX1var)/(tN+cn-2);
 X2pooledVar=((tN-1)*tX2var+(cN-1)*cX2var)/(tN+cn-2);
 X3pooledVar=((tN-1)*tX3var+(cN-1)*cX3var)/(tN+cn-2);
 X4pooledVar=((tN-1)*tX4var+(cN-1)*cX4var)/(tN+cn-2);
 Ydiff=tmean-cmean;
 pooledsd=sqrt( ((tN-1)*tvar+(cN-1)*cvar)/(tN+cn-2) );
 df=tN+cN-2;
 ttest=Ydiff/(pooledsd*sqrt(1/tN+1/cN));
 p=(1-probt(abs(ttest),(tN+cN-2)))*2;
 flag=0;
 if p<.05 then do;
      if Ydiff<0 then flag=-1;
      else flag=1;
 end;
 run;
proc means; class method d treatN ratio; var Ydiff preYdiff; run;
proc freq; tables method*d*treatN*ratio*flag/list; run;

*Quantity of Matches;
data d8; set d8;
tMatch=tN/tNPre;
```

```
cMatch=cN/cNPre;
proc means; class method d treatN ratio; var cNPre tNPre cN tN cMatch
tMatch; run;

*Quality of Matches;

*Propensity Score Mean Difference;
 data d8; set d8;
 proc means; class method d treatN ratio; var Propdiff; run;

 *Propensity Score Variance Ratio;
 proc means; class method d treatN ratio; var cPropvar; run;
 proc means; class method d treatN ratio; var tPropvar; run;

 *Continuous Covariates;
 proc means; class method d treatN ratio;
    var X1diff X2diff X3diff X4diff X1pooledVar X2pooledVar X3pooledVar
X4pooledVar;
run;

 proc means; class method d treatN ratio;
    var tX1meanPre tX2meanPre tX3meanPre tX4meanPre tX5meanPre
tX6meanPre cX1meanPre cX2meanPre cX3meanPre cX4meanPre cX5meanPre
cX6meanPre;
run;

 *Categorical Covariates;
 proc means data=d8; class method d treatN ratio; var tX5meanPre
tX6meanPre tX5mean tX6mean cX5meanPre cX6meanPre cX5mean cX6mean; run;

 *Save out the working file because it takes forever to create...;
data lib1.psm; set d8;
run;

data d8; set lib1.psm; run;

data lib1.psm; set d8; run;
proc datasets library=lib1;
 modify psm;
 index create method ;
  index create d ;
   index create treatN ;
    index create ratio ;
       index create rep ;
 run;

data d8; set d8;
if tX5mean=0 AND cX5mean=0 then X5SB=0;
else X5SB=((tX5mean-cX5mean)/(sqrt(((tX5mean*(1-tX5mean))+(cX5mean*(1-
cX5mean)))/2)));
if tX6mean=0 AND cX6mean=0 then X6SB=0;
else X6SB=((tX6mean-cX6mean)/(sqrt(((tX6mean*(1-tX6mean))+(cX6mean*(1-
cX6mean)))/2)));
run;

 *if unbalanced, use covariates, otherwise covariate string is empty;
```

```
%unbalance(Ran, 0, 30);
%unbalance(Ran, 0, 100);
%unbalance(NN0, 0, 30);
%unbalance(NN0, 0, 100);
%unbalance(NN1, 0, 30);
%unbalance(NN1, 0, 100);
%unbalance(NN2, 0, 30);
%unbalance(NN2, 0, 100);
%unbalance(NN3, 0, 30);
%unbalance(NN3, 0, 100);
%unbalance(Op1, 0, 30);
%unbalance(Op1, 0, 100);
%unbalance(Op2, 0, 30);
%unbalance(Op2, 0, 100);
%unbalance(Mah, 0, 30);
%unbalance(Mah, 0, 100);
%unbalance(Ran, 0.2, 30);
%unbalance(Ran, 0.2, 100);
%unbalance(NN0, 0.2, 30);
%unbalance(NN0, 0.2, 100);
%unbalance(NN1, 0.2, 30);
%unbalance(NN1, 0.2, 100);
%unbalance(NN2, 0.2, 30);
%unbalance(NN2, 0.2, 100);
%unbalance(NN3, 0.2, 30);
%unbalance(NN3, 0.2, 100);
%unbalance(Op1, 0.2, 30);
%unbalance(Op1, 0.2, 100);
%unbalance(Op2, 0.2, 30);
%unbalance(Op2, 0.2, 100);
%unbalance(Mah, 0.2, 30);
%unbalance(Mah, 0.2, 100);
%unbalance(Ran, 0.5, 30);
%unbalance(Ran, 0.5, 100);
%unbalance(NN0, 0.5, 30);
%unbalance(NN0, 0.5, 100);
%unbalance(NN1, 0.5, 30);
%unbalance(NN1, 0.5, 100);
%unbalance(NN2, 0.5, 30);
%unbalance(NN2, 0.5, 100);
%unbalance(NN3, 0.5, 30);
%unbalance(NN3, 0.5, 100);
%unbalance(Op1, 0.5, 30);
%unbalance(Op1, 0.5, 100);
%unbalance(Op2, 0.5, 30);
%unbalance(Op2, 0.5, 100);
%unbalance(Mah, 0.5, 30);
%unbalance(Mah, 0.5, 100);
%unbalance(Ran, 0.8, 30);
%unbalance(Ran, 0.8, 100);
%unbalance(NN0, 0.8, 30);
%unbalance(NN0, 0.8, 100);
%unbalance(NN1, 0.8, 30);
%unbalance(NN1, 0.8, 100);
%unbalance(NN2, 0.8, 30);
%unbalance(NN2, 0.8, 100);
%unbalance(NN3, 0.8, 30);
```

```
  %unbalance(NN3, 0.8, 100);
  %unbalance(Op1, 0.8, 30);
  %unbalance(Op1, 0.8, 100);
  %unbalance(Op2, 0.8, 30);
  %unbalance(Op2, 0.8, 100);
  %unbalance(Mah, 0.8, 30);
  %unbalance(Mah, 0.8, 100);


 *regression stuff follows;
 %macro unbalance(method, d, treatN);
 %do rep=1 %to 1000;
 %do ratio=3 %to 6;
 data temp; set d8;
 length mycov $20;
 if method = "&method";
 if d=&d;
 if ratio=&ratio;
 if treatN=&treatN;
 if rep=&rep;
mycov=" ";
/*check my cutting and pasting here;*/
if abs(X1diff/sqrt(X1pooledvar))>.25 then do;
   substr(mycov,1)="X1"; badX1=1; end;
   else badX1=0;  *want to keep a record of which covariates were
unbalanced;
if abs(X2diff/sqrt(X2pooledvar))>.25 then do;
    substr(mycov,4)="X2"; badX2=1; end;
         else badX2=0;
if abs(X3diff/sqrt(X3pooledvar))>.25 then do;
   substr(mycov,7)="X3"; badX3=1; end;
   else badX3=0;
if abs(X4diff/sqrt(X4pooledvar))>.25 then do;
   substr(mycov,10)="X4";
   badX4=1; end;
   else badX4=0;
if abs(X5SB)>.1 then do;
   substr(mycov,13)="X5";
   badX5=1; end;
   else badX5=0;
if abs(X6SB)>.1 then do;
   substr(mycov,16)="X6";
   badX6=1; end;
   else badX6=0;
run;
data bad; set temp;
 keep badX1-badX6; run;
data _null_; set temp;
 call symput('keepcov', mycov);
 run;
 %put &keepcov;
data temp2; set lib1.raw(where=(select=1 and method = "&method" and
d=&d and ratio=&ratio and treatN=&treatN and rep=&rep));
*if select=1;
run;
proc reg; model Y = group &keepcov;
 ods output  ParameterEstimates=parmest  FitStatistics=MSE;
```

```
 run;
 quit;
 options nocenter;
data parmest; set parmest;
 if variable = "group";
data MSE; set MSE;
if label1 ="Root MSE";
MSE=nvalue1**2; *adjusted within group variance;
data parmest2; merge parmest MSE bad;
method = "&method";
d=&d;
ratio=&ratio;
treatN=&treatN;
rep=&rep;
keep d method rep ratio treatN Estimate  StdErr tValue Probt MSE badX1-
badX6;
run;
proc datasets nolist; append base=lib1.adjD0b data=parmest2; run;
proc datasets nolist;  delete temp temp2 parmest parmest2 MSE; run;
%end; /*end rep loop; */
%end; /*end ratio loop; */
 %mend;


 *Read in files and create a final adj file;
 data partA; set lib1.adjD0; run;
 data partB; set lib1.adjD2; run;
 data partC; set lib1.adjD5; run;
 data partD; set lib1.adjD8; run;

 data all; set partA partB partC partD; run;

proc sort data=all; by method d treatN ratio rep;
run;

 data lib1.adj; set all;
 run;


 proc means data=all;
 class method d treatN ratio;
 var badX1 badX2 badX3 badX4 badX5 badX6;
 run;


data adj; set lib1.adj;
 flag=0;
 if probt<.05 then do;
      if Estimate<0 then flag=-1;
      else flag=1;
 end;
 run;

 proc freq; tables method*d*treatN*ratio*flag/list; run;

 data lib1.adj; set adj;
 run;
```

```
 data prop; set lib1.propensity;
 run;

  proc means data=prop;
   class d treatN ratio group;
   var propensity;
  run;


proc sort data=prop; by d treatN ratio rep group;
run;

 proc means data=prop noprint;
 by d treatN ratio rep group;
 var propensity;
 output out=prop2 mean=Propmean var=Propvar;
 run;
data tempppropa(drop=_type_ _freq_ Propmean Propvar); set prop2;
 if group=0;
 cPropmean=Propmean;
 cPropvar=Propvar;
 data tempppropb(drop=_type_ _freq_ Propmean Propvar); set prop2;
 if group=1;
 tPropmean=Propmean;
 tPropvar=Propvar;
run;

data prop3; merge tempppropa tempppropb;
 by d treatN ratio rep;
 Propdiff=tPropmean-cPropmean;
 VarRatio=tPropvar/cPropvar;
 run;

*Propensity Scores for everyone;
data temp2; set lib1.raw(where=(method = "NN0"));
run;
proc logistic data=temp2;
 by method d treatN ratio rep;
model group(Event='1')=X1 X2 X3 X4 X5 X6;
output out=lib1.propensity predprobs=I P=propensity;
run;

 *Propensity Score Mean Difference;
 proc means data=prop3; class d treatN ratio; var Propdiff; run;

 *Propensity Score Variance Ratio;
 proc means data=prop3; class d treatN ratio; var cPropvar; run;
 proc means data=prop3; class d treatN ratio; var tPropvar; run;


data d8; set d8;
 X1PrepooledVar=((tN-1)*tX1varPre+(cN-1)*cX1varPre)/(tN+cN-2);
 X2PrepooledVar=((tN-1)*tX2varPre+(cN-1)*cX2varPre)/(tN+cN-2);
 X3PrepooledVar=((tN-1)*tX3varPre+(cN-1)*cX3varPre)/(tN+cN-2);
 X4PrepooledVar=((tN-1)*tX4varPre+(cN-1)*cX4varPre)/(tN+cN-2);
run;
```

```
  proc means; class treatN ratio;
    var PreX1diff PreX2diff PreX3diff PreX4diff X1PrepooledVar
X2PrepooledVar X3PrepooledVar X4PrepooledVar;
run;


*ANOVAs;

proc glm data=all;
class method treatN analysis d ratio;
model bias=method|treatN|analysis|d|ratio;
run;
quit;

proc glm data=all;
class method treatN analysis d ratio;
model sqdiff=method|treatN|analysis|d|ratio;
run;
quit;

data all; set all;
absdiff=sqrt(sqdiff);

proc glm data=all;
class method treatN analysis d ratio;
model absdiff=method|treatN|analysis|d|ratio;
run;
quit;



data trad; set lib1.psm;
keep method d treatN ratio rep Ydiff ttest p flag analysis;
analysis=0;
run;

data trad; set trad;
rename ttest=tValue;
run;

data trad;
retain method d treatN ratio analysis rep Ydiff tValue p flag;
set trad;
run;

data adj; set lib1.adj;
analysis=1;
keep method d treatN ratio rep Estimate tValue probt flag analysis;
run;

data adj; set adj;
rename Estimate=Ydiff;
rename probt=p;
run;

data adj;
```

```
retain method d treatN ratio analysis rep Ydiff tValue p flag;
set adj;
run;

data all; set trad adj; run;

data all; set all;
bias=ydiff-d;
sqdiff=bias**2;
run;

proc freq data=all noprint; tables
method*d*treatN*ratio*analysis*flag/out=d9; run;

data d10; set d9;
 by method d treatN ratio analysis;
count=count/1000;
select(flag);
 when(-1) which='neg';
 when(1) which='pos';
 otherwise;
end;
if flag=0 then delete;
run;

libname lib2 xport "d:\PSYC 900 - Dissertation\Power.xpt";
data lib2.d9; set d10; run;

*get raw data;
data d9; set d8;
 if flag ne 0;
keep rep method d treatN ratio flag;
run;
libname lib2 xport "d:\PSYC 900 - Dissertation\Power.xpt";
data lib2.d9; set d9; run;

data d8; set all;
proc means; by analysis method d treatN ratio;
var bias sqdiff;
output out=d9 mean=; run;
data d9; set d9;
 RMSE=sqrt(sqdiff);
 run;

 libname lib2 xport "d:\PSYC 900 - Dissertation\bias.xpt";
data lib2.d9; set d9; run;
```

R Code for Graphing

```
library(Hmisc)
library(foreign)

mydata <- sasxport.get("E:/PSYC 900 - Dissertation/Power.xpt")
str(mydata)
mydata <- transform(mydata,
```

```
 analysis = factor(analysis, levels=c(0,1), labels=c("No
Covariates","Unbalanced Covariates")),
 ratio = factor(ratio, levels=c(3,4,5,6),
labels=c("3:1","4:1","5:1","6:1")),
 method = factor(method, levels=c("Ran", "NN0", "NN3", "NN2", "NN1",
"Op1", "Op2", "Mal"), labels=c("Ran", "NN", "NN3", "NN2", "NN1", "Op1",
"Op2", "Mah")) )


bs=12 # or 18 or 24 #most text will be 80% of this--manually change the
things that aren't, below
theme_set(theme_bw(base_size=bs))
theme_update(axis.title.x=element_text(size=.8*bs),
axis.title.y=element_text(size=.8*bs),
 plot.title=element_text(size=.8*bs), panel.grid.minor=element_blank(),
legend.background = element_blank(), strip.background =
element_rect(fill = 'white'))

################################
#Type I Error
################################

temp=subset(mydata,subset=(d==0 & treatn==30))
png(file ="E:/PSYC 900 - Dissertation/Graphs/TypeI30.png", units="in",
width = 6, height = 9,res=600)
ggplot(temp,aes(method,count)) + facet_grid(ratio~analysis)+
  geom_bar(stat="identity",aes(fill=which),colour="black")+
  geom_hline(aes(yintercept=0.05)) + xlab("Method")+
  ylab("Proportion Flagged")+ ylim(0,0.8) + theme(legend.position =
"bottom")+
  scale_fill_manual(values=c("gray90","gray10"), name="Direction",
breaks=c("neg", "pos"), labels=c("Negative", "Positive"))
dev.off()

temp=subset(mydata,subset=(d==0 & treatn==100))
png(file ="E:/PSYC 900 - Dissertation/Graphs/TypeI100.png", units="in",
width = 6, height = 9,res=600)
ggplot(temp,aes(method,count)) + facet_grid(ratio~analysis)+
  geom_bar(stat="identity",aes(fill=which),colour="black")+
  geom_hline(aes(yintercept=0.05)) + xlab("Method")+
  ylab("Proportion Flagged")+ ylim(0,0.8) + theme(legend.position =
"bottom")+
  scale_fill_manual(values=c("gray90","gray10"), name="Direction",
breaks=c("neg", "pos"), labels=c("Negative", "Positive"))
dev.off()

################################
#Power
################################

#Correct Direction

temp=subset(mydata,subset=(treatn==30 & which=="pos"& d>0))
png(file ="E:/PSYC 900 - Dissertation/Graphs/Power30.png", units="in",
width = 6, height = 9,res=600)
ggplot(temp,aes(method,count)) +
geom_point(aes(method,shape=as.factor(d)),size=2) +
```

```
 facet_grid(ratio~analysis) + theme(legend.position = "bottom")+
scale_shape_manual(values=c(0,17,1),name="Effect Size")+
 xlab("Method")+ ylab("Proportion Flagged") +
scale_y_continuous(limits=c(0,1))
dev.off()

temp=subset(mydata,subset=(treatn==100 & which=="pos"& d>0))
png(file ="E:/PSYC 900 - Dissertation/Graphs/Power100.png", units="in",
width = 6, height = 9,res=600)
ggplot(temp,aes(method,count)) +
geom_point(aes(method,shape=as.factor(d)),size=2) +
 facet_grid(ratio~analysis) + theme(legend.position = "bottom")+
scale_shape_manual(values=c(0,17,1),name="Effect Size")+
 xlab("Method")+ ylab("Proportion Flagged") +
scale_y_continuous(limits=c(0,1))
dev.off()

#Incorrect Direction

temp=subset(mydata,subset=(treatn==30 & which=="neg"& d>0))
png(file ="E:/PSYC 900 - Dissertation/Graphs/IncorrectPower30.png",
units="in", width = 6, height = 9,res=600)
ggplot(temp,aes(method,count)) +
geom_point(aes(method,shape=as.factor(d)),size=2) +
 facet_grid(ratio~analysis) + theme(legend.position = "bottom")+
scale_shape_manual(values=c(0,17,1),name="Effect Size")+
 xlab("Method")+ ylab("Proportion Flagged") +
scale_y_continuous(limits=c(0,1))
dev.off()

temp=subset(mydata,subset=(treatn==100 & which=="neg"& d>0))
png(file ="E:/PSYC 900 - Dissertation/Graphs/IncorrectPower100.png",
units="in", width = 6, height = 9,res=600)
ggplot(temp,aes(method,count)) +
geom_point(aes(method,shape=as.factor(d)),size=2) +
 facet_grid(ratio~analysis) + theme(legend.position = "bottom")+
scale_shape_manual(values=c(0,17,1),name="Effect Size")+
 xlab("Method")+ ylab("Proportion Flagged") +
scale_y_continuous(limits=c(0,1))
dev.off()

################################
#Bias
################################

bias <- sasxport.get("E:/PSYC 900 - Dissertation/bias.xpt")
bias <- transform(bias,
 analysis = factor(analysis, levels=c(0,1), labels=c("No
Covariates","Unbalanced Covariates")),
 ratio = factor(ratio, levels=c(3,4,5,6),
labels=c("3:1","4:1","5:1","6:1")),
 treatn = factor(treatn, levels=c(30,100), labels=c("Treatment N =
30","Treatment N = 100")),
 method = factor(method, levels=c("Ran", "NN0", "NN3", "NN2", "NN1",
"Op1", "Op2", "Mal"), labels=c("Ran", "NN", "NN3", "NN2", "NN1", "Op1",
"Op2", "Mah")) )
```

```
temp=subset(bias,subset=(d==0))
png(file ="E:/PSYC 900 - Dissertation/Graphs/Bias0.png", units="in",
width = 6, height = 9,res=600)
ggplot(temp) + geom_point(aes(method,bias,shape=analysis),size=2)  +
 scale_shape_manual(values = c(0,17),name="Analysis") +
 facet_grid(as.factor(ratio)~as.factor(treatn))+
 theme(legend.position = "bottom")+geom_hline(aes(yintercept=0.0))+
 xlab("Method")+ ylab("Bias")
dev.off()

################################
#RMSE
################################

rmse <- sasxport.get("E:/PSYC 900 - Dissertation/bias.xpt")
rmse <- transform(rmse,
 analysis = factor(analysis, levels=c(0,1), labels=c("No
Covariates","Unbalanced Covariates")),
 ratio = factor(ratio, levels=c(3,4,5,6),
labels=c("3:1","4:1","5:1","6:1")),
 treatn = factor(treatn, levels=c(30,100), labels=c("Treatment N =
30","Treatment N = 100")),
 method = factor(method, levels=c("Ran", "NN0", "NN3", "NN2", "NN1",
"Op1", "Op2", "Mal"), labels=c("Ran", "NN", "NN3", "NN2", "NN1", "Op1",
"Op2", "Mah")) )

temp=subset(rmse,subset=(d==0))
png(file ="E:/PSYC 900 - Dissertation/Graphs/RMSE0.png", units="in",
width = 6, height = 9,res=600)
ggplot(temp) + geom_point(aes(method,rmse,shape=analysis),size=2)  +
 scale_shape_manual(values = c(0,17),name="Analysis") +
 facet_grid(as.factor(ratio)~as.factor(treatn))+
 theme(legend.position = "bottom")+geom_hline(aes(yintercept=0.0))+
 xlab("Method")+ ylab("RMSE")
dev.off()
```

Appendix B

Simulated Conditions

Table B1

Simulated Conditions

Effect Size	Treatment $N = 30$			
	3 to 1	4 to 1	5 to 1	6 to 1
0.0	1	2	3	4
0.2	5	6	7	8
0.5	9	10	11	12
0.8	13	14	15	16
Effect Size	Treatment $N = 100$			
	3 to 1	4 to 1	5 to 1	6 to 1
0.0	17	18	19	20
0.2	21	22	23	24
0.5	25	26	27	28
0.8	29	30	31	32

Note. Data files were simulated 1,000 times for each effect size, treatment group sample size, and comparison-to-treatment group ratio combination, resulting in 32,000 data sets. Then, within each data set, the eight matching methods were applied, resulted in 256,000 matched groups. Finally, two sets of analyses (regression with no covariates and regression with unbalanced covariates) were conducted for each matched group, resulting in 512,000 regressions.

Appendix C

Power in the Incorrect Direction

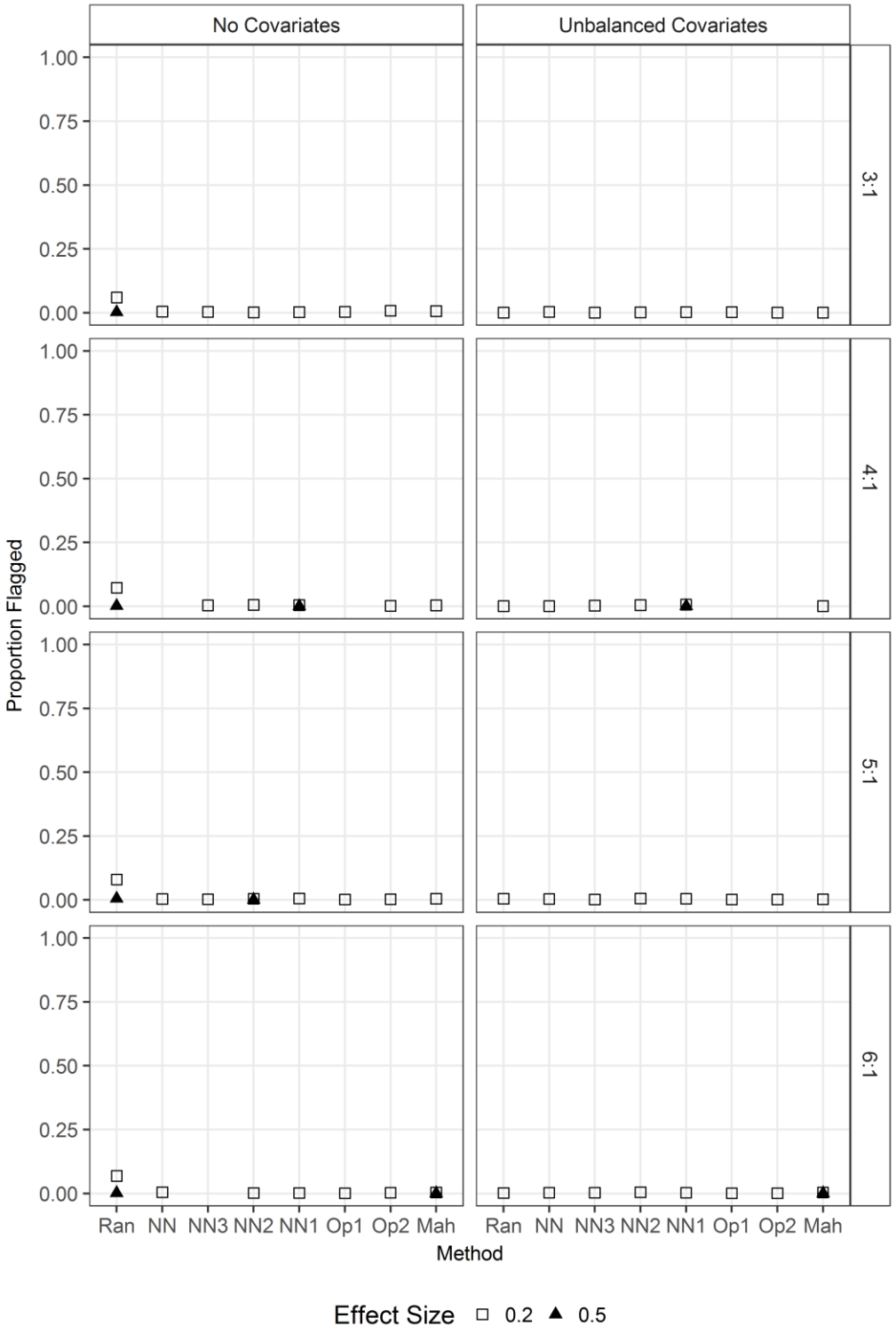

Figure C1. Power in the incorrect direction across conditions, treatment $N = 30$.

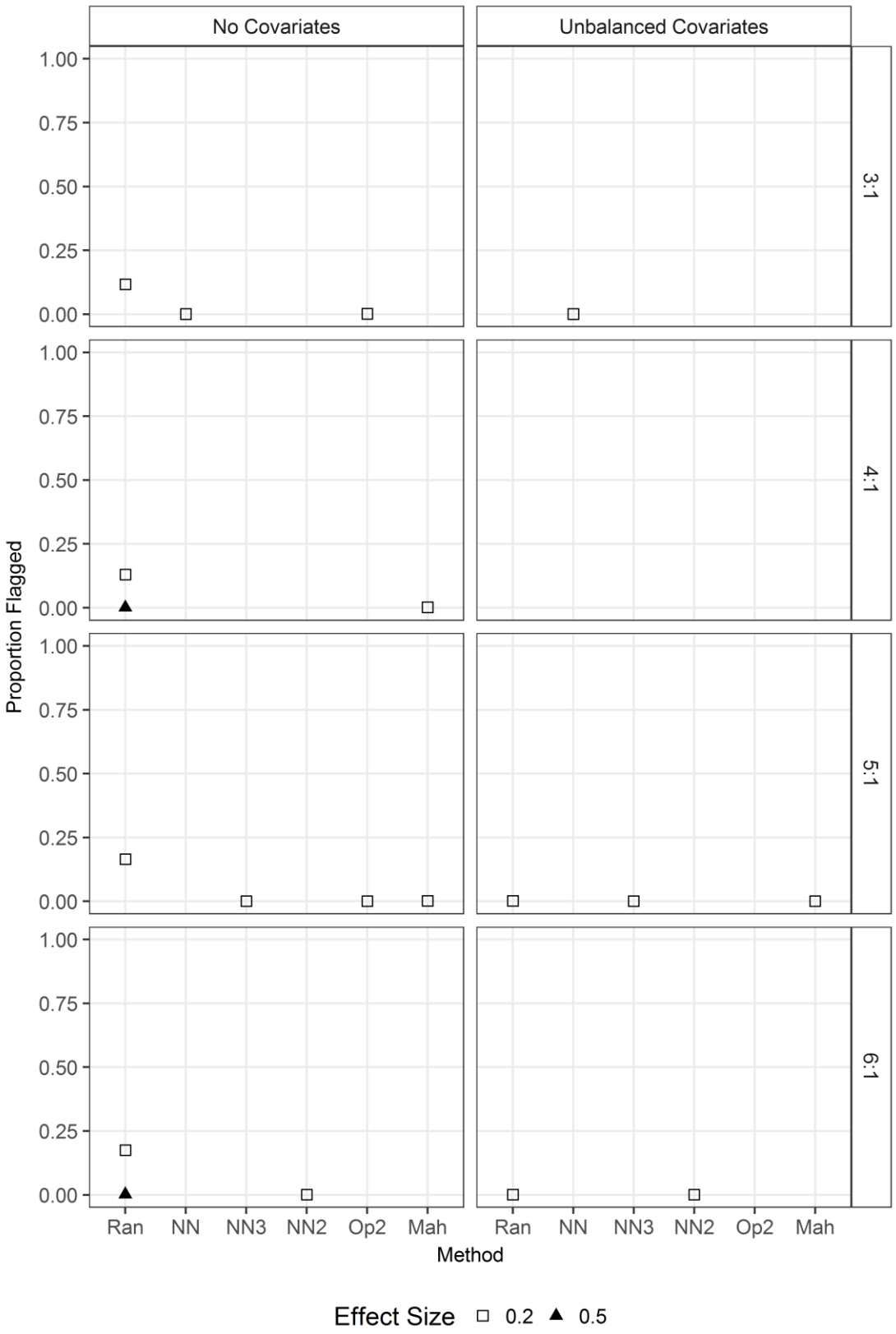

Figure C2. Power in the incorrect direction across conditions, treatment $N = 100$.

References

Austin, P. C. (2007a). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine, 27,* 2037-2049.

Austin, P. C. (2007b). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine, 26,* 3078-3094.

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

Austin, P. C. (2009b). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics, 5,* 1-21.

Austin, P. C. (2010a). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10,* 150-161.

Austin, P. C. (2010b). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology, 172,* 1092-1097.

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

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

Austin, P. C. (2013). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33,* 1057-1069.

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., Grootendorst, P., Normand, S. T., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine, 26,* 754-768.

Austin, P. C. & Schuster, T. (2016). The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. *Statistical Methods in Medical Research, 25,* 2214-2237.

Bai, H. (2011). Using propensity score analysis for making causal claims in research articles. *Educational Psychology Review, 23,* 273-278.

Bai, H. (2013). A bootstrap procedure of propensity score estimation. *The Journal of Experimental Education, 81,* 157-177.

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.

Branda, J. E. & Xieb, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review, 75,* 273-302.

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

Burgette, L. F., McCaffrey, D. F., & Griffin, B. A. (2015). Propensity score estimation with boosted regression. In W. Pan & H. Bai (Eds.), *Propensity Score Analysis: Fundamentals and Developments* (pp. 49-73). New York, NY: Guilford Publications, Inc.

Carpenter, R. G. (1977). Matching when covariables are normally distributed. *Biometrika, 64,* 299-307.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22,* 131-172.

Cheung, A. C. & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45,* 283-292.

Cochran, W. G., & Rubin, D.B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 35,* 417-446.

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

Cook, T. D. (1999, March). *Considering the major arguments against random assignment: An analysis of the intellectual cultural surrounding evaluation in American schools of education.* Paper presented at the Harvard Faculty Seminar on Experiments in Education, Cambridge, MA. Retrieved from

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.196.8497&rep=rep1&type=pdf

Clark, M. H., & Cundiff, N. L. (2011). Assessing the effectiveness of a college freshman seminar using propensity score adjustments. *Research in Higher Education, 52,* 616-639.

Davies, R. S., Williams, D. D., & Yanchar, S. (2008). The use of randomization in educational research and evaluation: A critical analysis of underlying assumptions. *Evaluation & Research in Education, 21,* 303-317.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluation the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053-1062.

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

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

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

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications.* (2nd Ed.). Los Angeles, CA: SAGE Publications.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *American Statistical Association, 99,* 609-618.

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

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

Harris, H., & Horst, S. J. (2016). A brief guide to decisions at each step of the propensity score matching process. *Practical Assessment, Research & Evaluation, 21*(4), 1-11.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12,* 247-267.

Hill, C. J., Bloom, H. S., Black, A. R., Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2,* 172-177.

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,* 199-236.

Ho, D., Imai, K., King, G., & Stuart, E. (2011). MatchIt: Nonparametric preprocessing for parametric casual inference. *Journal of Statistical Software, 42*(8), 1-28.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945-960.

Jacovidis, J. N., Foelber, K. J. & Horst, S. J. (in press). The effects of propensity score matching method on the quantity and quality of matches. *Journal of Experimental Education.* DOI: 10.1080/00220973.2016.1250209

Joffe, M. M., & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology, 150,* 327-333.

King, G. & Nielsen, R. (2016). *Why propensity scores should not be used for matching.* Retrieved from http://gking.harvard.edu/publications/why-Propensity-Scores-Should-Not-Be-Used-Formatching

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

Lipsey, M. W. (2002). Meta-analysis and program outcome evaluation. *Social Scientific Journal, 9,* 194-208.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29,* 530-558.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association, 96,* 1245-1253.

Melguizo, T., Kienzl, G. S., & Alfonso, M. (2011). Comparing the educational attainment of community college transfer students and four-year college rising juniors using propensity score matching methods. *The Journal of Higher Education, 82,* 265-291.

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

Olitsky, N. H. (2013). How do academic achievement and gender affect the earnings of STEM majors? A propensity score matching approach. *Research in Higher Education, 55,* 245-271.

Olmos, A., & Govindasamy, P. (2015). A practical guide for using propensity score weighting in R. *Practical Assessment, Research & Evaluation, 20*(13), 1-8.

Pan, W., & Bai, H. (2015). Propensity score analysis: Concepts and issues. In W. Pan & H. Bai (Eds.), *Propensity Score Analysis: Fundamentals and Developments* (pp. 3-19). New York, NY: Guilford Publications, Inc.

Pattanayak, C. W. (2015). Evaluating covariate balance. In W. Pan & H. Bai (Eds.), *Propensity Score Analysis: Fundamentals and Developments* (pp. 89-112). New York, NY: Guilford Publications, Inc.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 71,* 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association,* 79, 516-524.

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,* 33-38.

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

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. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of American Statistical Association, 81,* 961-964.

Rubin, D. R. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127,* 757-763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & 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.

Schochet, P. Z., D'Amico, R., Berk, J., Dolfin, S., & Wozny, N. (2012). *Estimated impacts for participants in the Trade Adjustment Assistance (TAA) program under the 2002 Amendments.* Princeton, NJ: Mathematica Policy Research, Inc.

Schuler, M. (2015). Overview of implementing propensity score analyses in statistical software. In W. Pan & H. Bai (Eds.), *Propensity Score Analysis: Fundamentals and Developments* (pp. 20-46). New York, NY: Guilford Publications, Inc.

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

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers?: A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103,* 1334-1344.

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

Soper, D. (2017). *Post-hoc statistical power calculator for a t-test.* Retrieved from http://www.danielsoper.com/statcalc/calculator.aspx?id=49

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics, 36,* 213-236.

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.

Steyer, R., Gabler, S., von Davier, A. A., & Nochtigall, C. (2000). Causal regression models: II. Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online, 5,* 55-87.

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.

Titus, M. A. (2007). Detecting selection bias, using propensity score matching, and estimating treatment effects: An application to the private returns to a master's degree. *Research in Higher Education, 48*, 487-521.

US Department of Education. (2015). *Skills for success grant solicitation* (CFDA Number: 84.215H). Washington, DC: US Department of Education.

US Department of Labor. (2014). *Trade adjustment assistance community college and career training (TAACCCT) grant solicitation* (Reference SGA/DFA PY13-10). Washington, DC: US Department of Labor.

What Works Clearinghouse. (2014). *What Works Clearinghouse procedures and standards (Version 2.1).* Retrieved from https://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*, 327-350.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics, 86*, 91-107.